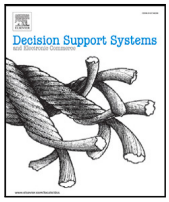
Decision Support Systems 185 (2024) 114310

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Approaches to improve preprocessing for Latent Dirichlet Allocation topic modeling

Jamie Zimmermann [a], Lance E. Champagne [b], John M. Dickens [c,*], Benjamin T. Hazen [d]

[a] Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson Air Force Base, OH, USA
[b] Associate Professor of Operations Research, Department of Operational Sciences, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH, USA
[c] Associate Professor of Supply Chain Management, Baker School of Business, The Citadel, Charleston, SC, USA
[d] Assistant Professor of Supply Chain Management, School of Business Administration: MIS, OSC and Business Analytics, University of Dayton, Dayton, OH, USA

ARTICLE INFO

Keywords:
Natural language processing
Heuristics
Principal component analysis
Latent Dirichlet allocation
Topic modeling
Stopwords

ABSTRACT

As a part of natural language processing (NLP), the intent of topic modeling is to identify topics in textual corpora with limited human input. Current topic modeling techniques, like Latent Dirichlet Allocation (LDA), are limited in the pre-processing steps and currently require human judgement, increasing analysis time and opportunities for error. The purpose of this research is to allay some of those limitations by introducing new approaches to improve coherence without adding computational complexity and provide an objective method for determining the number of topics within a corpus. First, we identify a requirement for a more robust stop words list and introduce a new dimensionality-reduction heuristic that exploits the number of words within a document to infer importance to word choice. Second, we develop an eigenvalue technique to determine the number of topics within a corpus. Third, we combine all of these techniques into the Zimm Approach, which produces higher quality results than LDA in determining the number of topics within a corpus. The Zimm Approach, when tested against various subsets of the 20newsgroup dataset, produced the correct number of topics in 7 of 9 subsets vs. 0 of 9 using highest coherence value produced by LDA.

1. Introduction

Organizations face increasing requirements to analyze and make sense of unstructured textual data [1]. Textual analysis includes various strategies and techniques to transform raw communication data into actionable intelligence [2]. For instance, data mining techniques are increasingly recognized for their capability to perform multi-dimensional analysis to assist in decision-making tasks [3]. Further, the definition of text mining is "the application of algorithms and methods from the fields of machine learning and statistics to texts with the goal of finding useful patterns" [4]. In this regard, topic modeling has grown in popularity to become a critical component of natural language processing (NLP).

Topic modeling detects hidden patterns in textual data without human intervention, wherein every word in a document may be treated as an attribute [5] and modeled as a finite mixture of topics [6]. The length of the data can range from a limited number of characters to pages of informational data and can influence the selection of technique(s) chosen for topic modeling [7]. All texts suffer from sparsity and noise, especially shorter texts [8]. Noise in textual data is information that does not provide meaning to the overall intent of the document or that distracts from the original meaning or intent of the text. Consequently, the more noise in a document the less effective topic modeling algorithms tend to be [8]. Indeed, topic modeling is prone to noise sensitivity and instability, which leads to unreliable and difficult to apply results [9]. Thus, removing as much noise as possible during pre-processing is desirable.

Our research focuses on Latent Dirichlet Allocation (LDA), a generative probabilistic approach used for modeling collections of discrete data [10]. LDA can be either supervised or unsupervised [9]. The goal of LDA is to find topics for a collection of documents [11] based on a nonhierarchical clustering of words [12]. It does not take into consideration the order of the words or the sentence structure; therefore, the word ordering is unimportant, thus creating a bag of words (BoW) [13]. A key assumption of LDA is that the BoW will maintain all relevant information [14]. It assumes all documents contain a mixture of topics [15], that is, the documents contain assorted topics, and the words within the documents are generated from those topics. Additionally, LDA assumes that the dimensionality of k (number of topics) of the

* Correspondence to: 116 Helena Park Dr, Summerville, SC 29486, USA.
E-mail addresses: Rjzim051609@gmail.com (J. Zimmermann), Lance.champagne@afit.edu (L.E. Champagne), Jdicken2@citadel.edu (J.M. Dickens), Hazenscm@gmail.com (B.T. Hazen).

https://doi.org/10.1016/j.dss.2024.114310
Received 22 November 2023; Received in revised form 21 August 2024; Accepted 22 August 2024
Available online 27 August 2024

Dirichlet distribution is known and fixed [10]. For k to be known, prior knowledge about the contents of the dataset is required [16]. LDA does not require previous training data and can handle mixed length documents, although for short messages an aggregation of the messages is needed to avoid data sparsity [17]. Collectively, we seek to relax these assumptions by introducing several novel techniques that reduce the number of subjective user inputs needed for topic modeling while providing stable, interpretable, and actionable outcomes.

LDA modeling method requires three user inputs: α, β and k [18]. Alpha, α, is the parameter that sets the prior on a per document topic distribution. A high α implies that every document is likely to contain a mixture of most topics, whereas a low α implies that the document contains fewer topics. In a low α, the topic distribution samples are near the corners, near the topics, implying that the document has only one topic. This number is between not-zero and positive infinity. Next, β sets the prior on the per topic word distribution. A high β implies that each topic may contain a mixture of most of the words, and a low β implies that a topic may contain a mixture of just a few words [18]. This number is greater than 0, not inclusive, and positive infinity. Third, the number of topics, k, is the number of topics the user wants the algorithm to extract from the corpus. The number of topic terms is the number of terms to be used in the composing of a topic, another user-specified parameter. The output of the LDA model is a list of topics and words with the associated probability that the word belongs to that topic.

There have been many advancements in the methods of topic modeling [19,20]. Despite such progress, dimensionality continues to be a challenge for text mining [21], leading to overfitting [22]. Additionally, selecting the number of topics for the methods to generate still presents a challenge and requires user input [23]. When using current techniques, the user must select the appropriate number of topics that accurately reflects the documents, directly affecting the overall results of the analysis. If the user chooses to identify too many topics, the information can become saturated and counterproductive, running the risk of "overclustering" [24]. On the other hand, if the user selects too few topics, the information may not be specific enough for the decision maker, running the risk that topics will be too broad [24]. Furthermore, the process requires the user to have prior knowledge of the dataset to select the optimal topic modeling technique for their dataset and to select the correct values for the inputs. Reducing the algorithm input decisions that are user-made reduces decision fatigue and increases reproducibility and overall algorithm performance.

The purpose of this research is to address these challenges to improve LDA topic modeling results. In the remainder of this article, we describe three new approaches aimed toward improving preprocessing for LDA topic modeling. These approaches provide both scholarly and practical contributions for anyone interested in using LDA topic modeling. Fig. 1 highlights our contributions, which are building-block techniques culminating in a less subjective topic modeling approach to enhance decision making with textual analysis methods. Overall, this is a detailed process flow chart that users can employ to enhance NLP outputs. It begins with data input and preprocessing. Next, users should develop a word cloud and refine it until useful insights can be obtained. This is our proposed Coherent Utility Process (CUP) heuristic. Then, users should perform our proposed Prominent Extraction Technique (PET) heuristic to narrow the bag of words, which feeds into a unique eigenvalue heuristic that leverages principal component analysis (PCA). Individually, each of these techniques provides value to the topic model practitioner. Collectively, these heuristics form the overall *Zimm Approach*, which we propose will provide more stable and useful results without *a priori* corpus knowledge.

Textual analysis in support of decision systems is an active area of research. Recently, scholarly works have focused on a variety of methods and potential applications of textual analysis. Hogenboom et al. [25] provides an excellent survey of NLP literature focusing on information extraction techniques. In the taxonomy proposed in [25], our proposed methods fall within "data-driven information extraction" as they rely on well-grounded statistical methods that can be applied across any domain. Our research is complementary to NLP in support of decision support systems works, but we extend the field in several important ways.

Krauss et al. [26] employed recurrent neural networks to extract features from textual data. Their research focused on a corpus with a known topic area specifically to support financial decisions. Our proposed methodology is intended to identify common topics within a potentially diverse corpus and identify keywords representative of the topics by which to categorize individual documents.

Kratzwald et al. [27] utilized deep learning techniques to uncover emotional states revealed through selection of words within documents. This research extends the rich area of sentiment analysis within NLP techniques. While an important area of NLP with broad decision support application, our proposed methods are not inherently concerned with the content of individual documents within a corpus outside commonalities that suggest clustering around one or more common topics.

Kim et al. [28] explored the visualization of convolutional neural networks used in natural language processing. Unlike these scholars whose methods are rooted in a "black box" system, we primarily deploy principal component analysis to identify statistically orthogonal topics within a text based on key word contribution to the overall variability of the corpus. PCA techniques are inherently transparent and do not require additional methods required in "black box" systems to provide explanatory power.

While many NLP techniques require curated corpora, specifically common themes among the grouped documents. Finding a reliable method for identifying the number of disparate topics within a corpus and document membership in the respective topics identified is a challenge. Myriad articles have proposed methods of identifying corpus topics, such as [14,24,29], and [16]. Despite the attention, the process remains highly subjective and often hinges on *a priori* knowledge of the corpus being investigated. We propose a methodology that advances this area of research and removes some of the inherent subjectivity in topic identification making the process more stable and repeatable by NLP practitioners.

Our first contribution develops a more customized, robust stopwords list, which results in a smaller BoW without loss of relevant information from the corpus. With our novel culling technique, CUP, we demonstrate the requirement for a robust stopwords list. Concurrently, we introduce a dimensionality-reduction technique, called PET, which employs the total number of words within a document set to produce a higher quality result from LDA topic modeling. Additionally, when we pair CUP with our bag of word dimensionality-reduction procedure, PET, we report improved LDA of output as measured by the coherence value metric.

In the second approach, we examine methods used to assist the user in determining the number of topics, k, as an input for various topic modeling techniques. Existing techniques could provide multiple numbers to the user, requiring the user to decide which is correct in absence of *a priori* knowledge on the corpus. To compensate for this deficiency, we developed an eigenvalue technique that determines the number of topics for the user as an input based on the covariance matrix of the transposed term-document matrix.

In the third contribution, we propose a new topic modeling technique, the *Zimm Approach*, that builds upon and employs the previous two contributions to address the limitations of requiring the user to input parameter values for the number of topics and the number of terms per topic into a topic model and to provide a stable output. The new technique requires the user to only input the textual data and any respective custom stopwords list the user may need. The number of topics and number of words associated with each topic is determined by the technique.

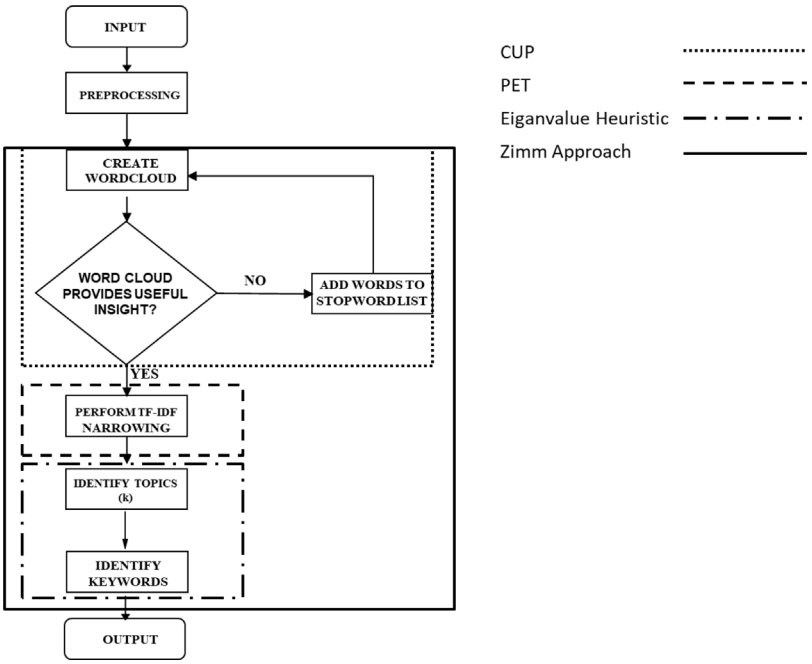


Fig. 1. Overview of contributions in topic modeling.

2. Preprocessing with coherent utility process and prominent extraction technique

The BoW is a representation of the words within a document. It is a vector representation of the document, where each element is the normalized number of occurrences of the term in the document [30]. During the computations, sequential information is not maintained [31]. The BoW is used as an input in many topic modeling techniques, such as LDA.

While the BoW is used to represent a corpus, there are limited theoretical studies on its properties [32]. The BoW suffers from high dimensionality [30] and can reach many thousands of potential predictors to assist in topic modeling [33]. Passalis and Tefas [34], Zhao et al. [35], Ljungberg [36], and Boulis and Ostendorf [37] addressed high dimensionality within the textual analysis domain; however, their techniques leave room for further improvement.

Geva and Zahavi [33] used preprocessing techniques, such as stemming and stopwords list filtering to reduce the dimensionality of the BoW. Their technique led to the need to select a specified top number of words. Despite efforts made to improve the BoW input, a methodology for bounding the Term Frequency-Inverse Document Frequency (TF-IDF) technique has not been addressed. We suggest a novel approach to narrow the BoW used in topic models based on the TF-IDF in addition to introducing a process to select words to create a unique, dataset-specific stopwords list.

In the proposed process, the textual data input remains the same, and the user or the algorithm or both still perform preprocessing to cleanse the data. Subsequently, a word cloud is created to help identify the main topic and potential subtopics of the dataset. If the word cloud does not consist of excessive noise, then the TF-IDF narrowing technique is performed and fed into the selected topic modeling. If the word cloud contains excessive noise, the user creates a unique stopwords list to assist in noise filtering, which is fed back into the creation of a new word cloud for the user to iteratively examine. We call this improved process CUP, which is an iterative process that is complete once the user is satisfied that enough noise has been eliminated from the word cloud to generate insight.

We also suggest a complementary dimensionality-reduction technique, PET, which uses the number of words within a document set to produce a higher quality result from LDA. The resulting dimensionality reduction utilizes the LDA topic modeling in the evaluation criteria to test and analyze the effects of narrowing the BoW based on the TF-IDF values with the removal of stopwords, utilizing both premade and custom lists. Thus, analysts can effectively right-size the BoW to achieve a level of utility not previously possible, as shown in the following example.

2.1. Proposed methodology

We used the baseball subset of the 20newsgroup dataset [38], which includes a collection of 11,314 text files of seven subjects, labeled for topics and subtopics. We performed common pre-processing steps: lowercasing all letters; removing special characters, digits, and stopwords using python preloaded package; stemming [39] ; and lemmatizing [40] using the Porter Stemmer [41]. Lemmatizing algorithms are generally slower than stemming because rule-based methods proceed through the corpus to find relevant word associations [42]. The WordNetLemmatizer from the Natural Language ToolKit was used.

Some corpora are noisy, meaning they contain information that is irrelevant to the user's specific needs [43], negatively affecting model output. The initial step to reduce noise is to create a visualization of the word cloud, which provides the user with a means of identifying words that do not add value in providing insight into the data. CUP is used to identify irrelevant words in the corpus and to create a unique, data-specific stopwords list, thereby reducing noise.

Once the stopwords list is created and employed, an objective technique for narrowing the BoW, called PET, can be applied. According to Eassom [44], effective keywords should be mentioned every 100 to 200 words in a journal article. Therefore, the total word count divided by 100 and 200 is utilized in the equations for PET. Eqs. (1) and (2), respectively, show the calculation for the lower and upper bounds on word frequency:

$$\frac{w}{200} - (0.10 \cdot w) \tag{1}$$

$$\frac{w}{200} + (0.10 \cdot w) \tag{2}$$

where w = the total number of the words in the BoW.

Fig. 2. Word clouds from a baseball dataset.

Both the lower bound, Eq. (1), and the upper bound, Eq. (2), were rounded down and up, respectively, to the nearest whole number. After calculating a lower and upper bound for the word frequency, the minimum and maximum TF-IDF values within that word frequency range were used to create the narrowed BoW.

Varying the percentage of the BoW, either added to or subtracted from the upper and lower bounds, respectively, created a space-filling screening design. The design used percentages from 0 to 20, with increments of 0.025. After completing the analysis, an increment of 0.10 provided a reasonable calculation without overestimating the word count bounds used in determining the minimum and maximum TF-IDF values. Therefore, we chose an increment of 0.10 when creating the BoW for the LDA topic modeling technique.

2.2. Evaluation

We used word clouds, coherence scores, and the overall output of the LDA model as evaluation criteria for algorithm effectiveness. This analysis illustrates that our novel culling techniques provide more discrimination, with greater dataset interpretability and clarity. Appendix A provides the algorithm for CUP and PET. The TF-IDF files were exported to excel where the narrowing calculations were performed. The narrowing bounds were inputs into the python code.

The first step in the proposed process requires the user to create and analyze a word cloud for usability. Word clouds representing varying word frequency from a baseball dataset are shown in Fig. 2. Word cloud A is generated using the full data set and Python's stopwords package for the baseball dataset. As shown, the general topic of the dataset, baseball, is not evident because extraneous words relating to the data format (i.e., email) dominate. Words that appear larger are more general words, providing little additional information about the dataset. However, with closer inspection to less prominent words in the cloud, there is an indication that the dataset may be about a sport. Similarly, we conducted the topic modeling process without the additional TF-IDF narrowing process using only the prestored Python stopwords package. As was the case with the word cloud, the words assigned by the LDA topic modeling technique do not provide adequate clarity because the general words dominate the topic-specific words.

Applying PET to the baseball dataset did not narrow the TF-IDF range, that is, the entire BoW was still being used. Therefore, we moved directly into the CUP technique. By following the CUP technique, we added another 48 words to the baseball dataset stopwords list. We created a word cloud to ensure the CUP technique was beneficial to the overall analysis, shown in Fig. 2B. Because of our culling technique, the user can now identify more insightful details about the datasets prior to PET (i.e., TF-IDF narrowing). With the application of our CUP technique, the LDA output has also subjectively increased in fidelity.

Utilizing the unique stopwords list emerging from the CUP technique provided the user with more insight into the dataset. To continue providing more details, we combined the unique stopwords list with the BoW dimensionality reduction technique, PET. Prior to PET, the TF-IDF range was [0.000751, 0.047811], after applying PET, the TF-IDF range narrowed to [0.025573, 0.046691].

Fig. 2C shows the results of word cloud creation after utilizing the new process. For example, the baseball dataset now shows that teams such as the Braves, the Cubs, the Mets, and the Phillies are represented in the dataset, information that was not prevalent prior to employing our CUP and PET culling techniques. We believe that, when compared to the less filtered word cloud in Fig. 2A, our culling techniques provide more utility and insight into the dataset. The coherence score improved by 43.4%. Fig. 3 shows an overall improvement on the coherence scores for number of topics, k, ranging from one through five, when using CUP and PET.

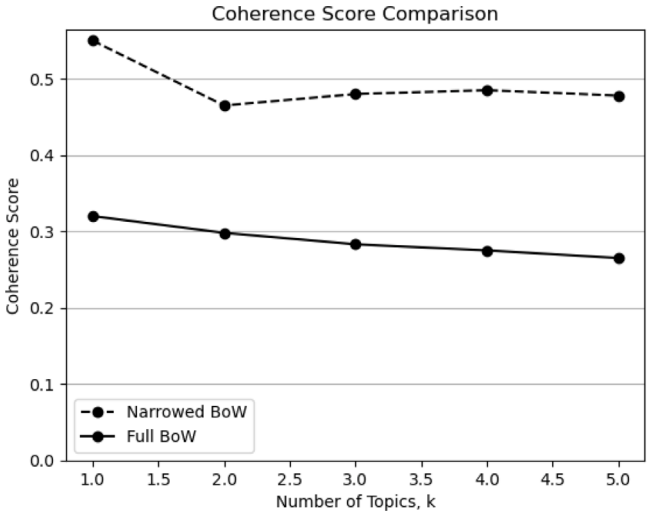


Fig. 3. Coherence score comparison.

3. Heuristic for determining number of topics, k

Many topic modeling approaches require the user to specify *a priori* the number of topics, k, contained in the corpus. Unfortunately, this approach requires the user to have advanced insight into the data, which is often not possible due to its volume and the competing demands. An additional complication manifests when the number of topics selected has a direct negative influence on the overall output of the model. This modeling flaw creates distortions that unintentionally influence the interpretability of the statistical model, thereby marginalizing its managerial utility [45].

A common way of modeling topics is to treat each topic as a probability distribution over words [46]. If there are T topics, then the probability of the ith word in a given document is written as

$$P(i_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \tag{3}$$

where

z_i is a latent variable indicating the topics from which the ith word was drawn;

$p(w_i|z_i = j)$ is the probability of the work w_i under the jth topic; and

$p(z_i = j)$ is the probability of choosing a word from topic j in the current document

3.1. Extant methodologies

Two assumptions common throughout most of the models are one, k is known and fixed and two, the words are infinitely exchangeable, as are the topics within the document [47]. Given the exponential growth of digital datasets and information extraction needs [25], research suggests different techniques to determine the number of topics for various topic models. This subsection provides a short overview of extant methodologies.

3.1.1. Graph dimensionality selection techniques

Graph-based dimensionality selection or the number of topics, k, has been used in methods like singular value decomposition (SVD) and PCA where the natural indicator is the singular value or eigenvalue, respectively. Fu et al. [48] showed that SVD and PCA produced comparable results when determining the optimal numbers of topics.

Fu et al. [48] used the elbow point in a scree plot to identify the optimal number of topics. The elbow method utilizes k-mean clustering on input data for a given number of clusters, k. The sum of squared errors is calculated for each cluster. The sum of squared errors is the distance of all data points to their respective cluster center. After plotting the number of clusters by the sum or squared errors, the point in which the sum of squares decreases abruptly is taken and one is added, resulting in the ideal number of topics. Fu et al. [48] noted that their findings were based on specific textual data. The heuristic proposed in this chapter is intended for a variety corpus and is based on the term-document matrix.

PCA is a multivariate technique that extracts information and represents that information as a set of new orthogonal variables, called principal components, and then displays a map that shows pattern(s) of similarity among the observations [49]. The goal of PCA is to identify major components embedded in the data matrix. This technique reduces noise data since the maximum variation source is selected and the small variations are ignored. In PCA, principal components are exact linear transformations of the data without considering residual error [50]. Our heuristic uses PCA.

3.1.2. Bayesian methods

A Bayesian classifier assumes all words in the document come from a single class [46]. This is not always the case. Input can come from multiple classes [51]. Griffiths and Steyvers [46] looked at the effects of changing the number of topics, utilizing the Gibbs sampling algorithm. The Gibbs sampling algorithm is a Markov chain Monte Carlo, a stochastic process for computing and updating α and β [52]. The Griffiths and Steyvers [46] dataset is comprised of 28,154 abstracts published in PNAS from 1991 to 2001. In LDA, two other input parameters are α and β. A high α indicates that every document is likely to contain a mixture of most topics and not a single topic. A low α indicates that a document is more likely to contain one or just a few topics. A high β indicates that each topic is likely to contain most of the words and not any word specifically. And a low β indicates that each topic may contain a mixture of only a few words. The value of α and β affect the optimal number of topics; therefore, during the experiment $\alpha = 50/k$ and $\beta = 0.1$ were fixed, and k was varied using Bayesian statistics. The optimal number for k is selected based on the log-likelihood of the data.

While Griffiths and Steyvers [46] proposed an approach to determine k, varying k and computing and graphing calculations was still required. This required the user to know a range in which to vary k and to know how to understand and interrupt the results of the graphs. There is potential for the optimal value of k to fall outside of the range in which the user selects to test. Our proposed heuristic does not require comparisons of various computations by varying k.

3.1.3. Stability analysis

Greene et al. [24] proposed a term-centric stability analysis strategy to address the issue of selecting the appropriate number of topics as an input to the nonnegative matrix factorization (NMF) topic modeling technique: k in $[k_{min}, k_{max}]$. S denotes the ith topic produced by the algorithm list R_i, that is, $S = R_1, \ldots, R_k$, where k is the number of ranked lists. In NMF this will correspond with the highest ranked values in each column of k basis vectors [24]. Jaccard similarity can be used to measure the similarity between the two top words of any two topics. If two topics have the same top word, then the Jaccard measure would be 1, and if all top words were different, then the Jaccard measure would be 0 [53]. The Jaccard index does not account for positional information. In other words, terms that are listed at the top of a ranked list will naturally be more relevant to a topic than those at the end of the list [24]. To alleviate this problem, Greene et al. [24] utilized a ranking distance measure proposed in [54].

Greene et al. [24] referred to Fagin et al.'s [54] approach as the average Jaccard (AJ) approach. The AJ approach is used to analyze the similarities between a pair of ranked lists (R_i, R_j). AJ is a top-weighted version of the Jaccard index, expressed as

$$AJ(R_i, R_j) = \frac{1}{\tau}\sum_{d=1}^{\tau}\gamma_d(R_i, R_j) \tag{4}$$

where

$$\gamma_d(R_i, R_j) = \frac{\left|R_{i,d} \cap R_{j,d}\right|}{\left|R_{i,d} \cup R_{j,d}\right|} \tag{5}$$

produces a value between $[0, 1]$ with

$$stability(k) = \frac{1}{\tau}\sum_{i=1}^{\tau} agree(S_0, S_i) \tag{6}$$

where

τ is the number of samples in dataset constructed by randomly selecting a subset of $\beta \cdot n$ documents without replacement.

$0 \le \beta \le 1$ is the sampling ratio controlling the number of documents in each sample, and

$$agree(S_x, S_y) = \frac{1}{k}\sum_{i=1}^{k}[AJ(R_{xi}), \pi(R_{xi})] \tag{7}$$

where

$$S_x = \{R_{x1}, \ldots, R_{xk}\}$$

$$S_y = \{R_{y1}, \ldots, R_{yk}\}$$

A plot of the stability scores is created. The final value of k will be based on the peaks of the plot. If more than one peak exists, then that may indicate that the corpus can be associated with more than one topic and the user still must decide on the value for k, thus no longer removing the decision-making requirement.

3.2. Coherence scores and perplexity

Topic coherence measures are a qualitative approach to automatically uncover the coherence of a topic [55]. They score a single topic by measuring the degree of semantic similarity between high-scoring words in the topic. The measures assist in differentiating between topics that are semantically interpretable and topics that are artifacts of statistical inferences [56]. Topics are "coherent" if all or most of the works are related, and if they support each other. Campagnolo, et al. [57] highlight the sensitivity of coherence scores (i.e., Cv) to noise in the data. The implementation of CUP and PET help to reduce noise by narrowing the BoW. The resulting effects on Cv are shown in

Fig. 5 and allow the coherence score to more accurately represent the documents within their respective topics.

Common topic coherence measures are the UCI measure [58], UMass measure [59], and Cv [60]. These measurements have been shown to reflect human judgement when referencing topic quality [56]. UCI and UMass measures compute the coherence of a topic as the sum of a pairwise distributional similarity scores, as in formula (8):

$$C_v = \sum_{(v_i, v_j) \in V} score(v_i, v_j, \epsilon) \tag{8}$$

where V is a set of words describing the topics, and ϵ is the smoothing factor to guarantee that score returns real numbers. The value of $epsilon$ is set to 1; however, Stevens et al. [56] looked at the effects of varying the value. Newman et al. [61] showed coherence scores based on pointwise mutual information (PMI) and normalized pointwise mutual information (NPMI) have the highest correlation with human judgement in topic evaluation [16].

The UCI measure defines the score as a PMI between two words, as shown in formula (9). The score can also be thought of as an external comparison to known semantic evaluations [56], such as

$$score(v_i, v_j, \epsilon) = log \frac{p(v_i, v_j) + \epsilon}{p(v_i, v_j)} \tag{9}$$

In contrast, the UMass measure defines the score based on document co-occurrence [56], as shown in formula (10). This measure uses the counts over the original corpus used to train the topic models, rather than the external corpus, as in the UCI measure, leading this metric to be more intrinsic in nature and expressed as

$$score(v_i, v_j, \epsilon) = log \frac{D(v_i, v_j) + \epsilon}{D(v_j)} \tag{10}$$

where $D(x, y)$ counts the number of documents containing x and y words, and $D(x)$ counts the number of documents containing x [56].

Aletras and Stevenson [62] showed that NPMI correlates with human judgement better than PMI. NPMI reduces the impact of low frequency counts in word co-occurrences thus utilizing more reliable estimates [63], thus leading to the improvement of NPMI over PMI. Rőder et al. [60] looked at the top word of a topic instead of defining probabilities over word pairs [64]. The Cv measure combines the indirect cosine measure with the NPMI and the Boolean sliding window [60].

Statistical measures of perplexity or likelihood of test data have been the method of choice for the evaluation of topic models [58]. Zhao et al. [35] used perplexity scores to assist in determining the optimal number of topics for the LDA model. Perplexity is defined as

$$p(D_{test}) = exp\left\{ -\frac{\sum_{d=1}^{M} log\left[p(w_d)\right]}{\sum_{d=1}^{M} N_d} \right\} \tag{11}$$

where p is the perplexity and D is the corpus containing M documents, d, having N_d words ($d \in \{1, \ldots, M\}$).

The point in which the rate of perplexity changed was determined to be the optimal number of topics. The perplexity measure does not reflect the semantic coherence of individual topics, nor does it provide indication to the user of the topic model's performance. It has been suggested that perplexity measures are contrary to human judgement [65].

While all these methods provided the researchers with promising results, the potential for multiple peaks still exists. Therefore, these techniques still required the user to decide on which peak they should select. Our proposed methodology introduces a heuristic that removes this requirement and provides the number of topics as an immediate input.

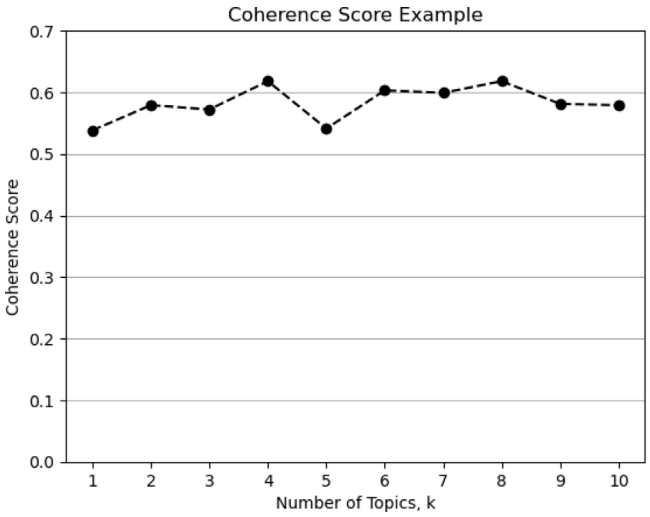


Fig. 4. Coherence score comparison.

3.3. Proposed methodology

We used a subset of datasets from 20 newsgroups [38], specifically, a varied combination of a collection of 11,314 text files of seven subjects labeled for topics and subtopics. The text documents were put through various pre-processing algorithms for stemming, lemmatization, removal of symbols, punctuation, and stopwords using preloaded python packages.

Röder et al. [60] introduce a coherence score measure, C_v, which achieves the highest correlation with all available human topic ranking. They chose LDA as the topic modeling technique, varying k to compute the coherence scores. After the coherence scores were calculated and plotted, the results were compared to the proposed technique in the analysis section.

The coherence score technique requires the user to input k to calculate the results, plot the various scores among a user-specified number of unique ks, and then determine the optimal number of topics. This is resource intensive and requires the user to interpret the plot or output of the coherence values. In addition, this approach presents a couple of immediate challenges: First, what range of k should the user specify to test for the optimal k? And second, what happens if more than one peak exists? Fig. 4 shows an example of a coherence score plot where the coherence score peak is the same for both values 4 and 8. In this case, the user would have to decide which number to use as an input into the model. The goal is to minimize the number and importance of user decisions, thus reducing opportunities for error.

Fig. 5 shows the coherence score plots for LDA prior to and after CUP/PET for the baseball data set. The figure illustrates the process used to determine k from the coherence value metric. Without CUP/PET, the highest coherence value overestimates the number of topics (2) within the corpus. Once CUP/PET is applied, the coherence value indicates the correct number of topics (1). Table 1 reflects the results of this process for each of the data sets. The location of the peak in each line was used to determine the number of topics the user would select when using the coherence score approach.

We propose a heuristic using the eigenvalues of the covariance matrix of the term-document matrix to determine the number of topics. A term-document matrix is a table showing the frequency of terms in a collection of documents. The rows correspond to terms and the columns correspond to documents, while the entries are the frequency of each term in a document.

The proposed heuristic utilizes the term-document matrix, providing an answer that will be fed directly into the LDA topic modeling technique. This eliminates the requirement for a user to manually enter the

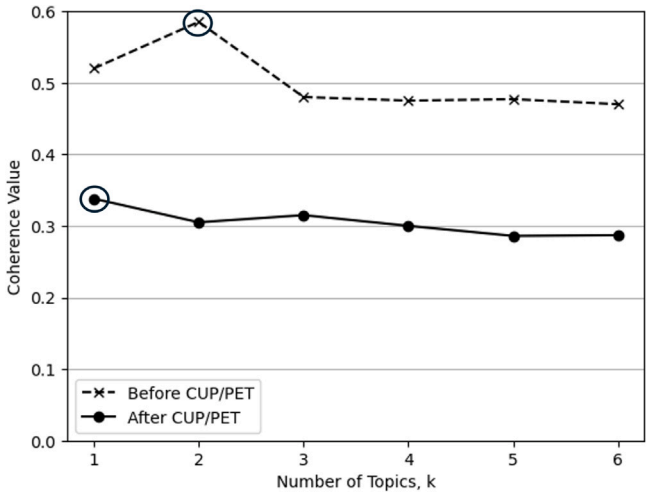


Fig. 5. Baseball dataset coherence value comparison.

Table 1
Number of categories predicted Coherence Score (CS) vs. Eigenvalue Heuristic (EH).

Topic(s)	CS w/o CUP/PET	CS with CUP/PET	EH w/o CUP/PET	EH with CUP/PET
Baseball	2	1	1	1
Baseball, Hockey	4	1	3	2
Baseball, Hockey, Space	2	1	3	3
Baseball, Hockey, Space, Auto	1	3	3	3
Baseball, Hockey, Space, Auto, Med	2	4	4	4
Space, Autos, Med	1	5	3	3
Autos, Med	5	3	2	2
Hockey, Autos, Med	1	1	3	3
Hockey, Space, Autos, Med	1	5	4	4

number of topics and make decisions based on a dataset that he or she may have little knowledge of.

Initially, the step of looking at the scree plot and finding the point of maximum curve was tested. This approach did not result in accurate results when tested on data of which the number of topics were known. The proposed heuristic identifies the number of topics being equated to the number of eigenvalues, of the covariance matrix of the term-document matrix, greater than one. Appendix B provides the algorithm for the eigenvalue heuristic as well as the LDA and coherence score algorithms used in the analysis.

The eigenvalue heuristic was applied to a variety of datasets containing one through five main topics. Table 1 displays the results of the eigenvalue heuristic, the number of topics the user would have selected if utilizing the method of selecting the largest coherence score, and the results when the two methods are used with CUP and PET. Both methods, the eigenvalue heuristic and the coherence score approach, had improved results when paired with our CUP/PET techniques. The eigenvalue heuristic provided a more reliable approach to determining k as an input into the LDA topic modeling technique. Since LDA is sensitive to a varying k, an effective and reliable approach is critical to increase model stability.

In 66.7% of the nine runs, the eigenvalue heuristic produced the correct number of topics, whereas using the coherence score did not correctly identify the number of topics for any corpus tested. The *Zimm Approach*, when the eigenvalue heuristic is used with CUP and PET, 77.8% of the nine runs produced the number of correct number of topics, versus 11.1% when using the coherence score approach with CUP and PET only.

4. The *Zimm Approach* : A new topic modeling technique

We propose a new topic modeling methodology termed the *Zimm Approach*, which incorporates and extends the techniques described above. Unlike LDA, the proposed technique does not require the user to input α, β, the number of topics, nor the number of topic terms. This removes the requirement for prior knowledge of the dataset or access to someone who has knowledge of the dataset. Appendix C provides the full *Zimm Approach* algorithm and the LDA algorithm used in this analysis.

4.1. Proposed methodology

Factor analysis (FA) is an unsupervised learning method for discovering latent variables and has been used since the early 1960s to extract topics and automatically classify documents [50]. PCA and FA are similar dimensionality reduction techniques; however, there are some differences. PCA does not generate a model of underlying principal components similar to FA. While both PCA and FA take new dimensions as a hyperparameter, only the FA model should be rebuilt, while the principal components already computed in PCA are not affected. Therefore, PCA is used in this topic modeling technique.

We used the "auto" and "med" files from the 20newsgroup dataset [38]. This led to a dataset size of 1088 text files and 19,140 words after preprocessing. The preprocessing included the standard lowercasing of letters, removal of punctuation, lemmatizing, and stemming. For this dataset, email characters were also removed.

We employed the previously proposed heuristics and approaches to begin the *Zimm Approach*. Eigenvalues were computed and the number of topics assigned based on the number of eigenvalues greater than one. The associated eigenvectors were extracted, and the loadings were calculated using formula (12):

$$L_i = v_i \cdot \sqrt{\lambda_i} \tag{12}$$

where

v_i is the eigenvector associated with the eigenvalue λ_i for all i such that $\lambda_i > 1$.

The loadings for each topic were sorted and plotted. The maximum curvature in each plot was used to identify where the cut off for the terms to be associated with each topic was located. This allowed for the number of terms for each topic to vary based on the loadings for each topic. The loadings were then mapped back to the term matrix to output terms for the number of topics specified.

We initially created a word cloud to implement CUP. Fig. 6 displays the word clouds prior to and after CUP. After creating and implementing the custom stopwords list, the word cloud after CUP shows that noise, which previously saturated the main ideas of the data, was filtered out.

Based on the eigenvalues greater than one heuristic, the algorithm suggested 37 topics. The algorithm was fed a dataset with two main topics and multiple potential subtopics. Additionally, the algorithm was modified to look at the maximum curvature of the scree plot of eigenvalues. This provided a value of 13. The output of the algorithm of both $k = 37$ and $k = 13$ was used for the *Zimm Approach* and for LDA. The varying of k demonstrated another benefit of this algorithm.

In LDA, varying k varies the output. The terms in the grouping of each topic will change based on the user specified k. Additionally, in LDA the user must specify the number of terms to output with the topics. The number of terms with the topics will be the same. For example, if the user selects 10, then there will be 10 terms in the output for each topic. Appendix E shows all the LDA and *Zimm Approach* output when the number of terms is 10, and k is 13, and again when k is 37. As shown, the terms will vary when k varies in the LDA topic modeling technique.

In the *Zimm Approach*, whether selecting 37 or 13, the first thirteen groups of terms are the same. When k varied, the words associated with each topic did not change. Therefore, if an individual decided to manually select k, the output within the topics would not change. Furthermore, the number of terms selected for each output is not consistent and does not require user input, as discussed below.

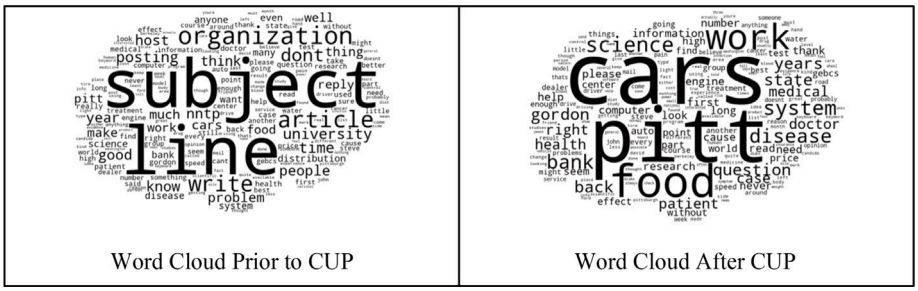


Fig. 6. Coherence score comparison.

After extracting each corresponding eigenvector and eigenvalue, the corresponding loading was calculated based on Eq. (12). The loading values were plotted, and the maximum curvature point of each plot was used to determine the number of terms for each topic. Then the vector values were mapped back to the term matrix to produce an output of k topics that contains the number of terms determined by the corresponding plot. This method allowed for a varying number of terms per topic since some terms may contribute more to the calculations than others. Appendix E shows a sample of the output for the *Zimm Approach* when $k = 13$ and again when $k = 37$. The number of terms per topic varies based on the heuristic of the algorithm; however, the terms do not vary when k changes and are thus consistent and stable. This stability is important when adding additional documents to the corpus. This approach will provide the user with a way to compare the impact of the new documents.

5. Conclusions

Data science results need to be interpretable if they are to aid in decision-making processes [66]. However, as we explained in this research, there are many varying factors that can influence the quality of the topic model output. Most importantly, the quality and quantity of the data fed into the models is a critical aspect towards maximizing the value and interpretability of the results. Technological improvements and advanced computing capacity have enabled vast amounts of data to be analyzed quickly; however, as the data becomes more complex and disparate, the quality of inputs can quickly and unintentionally degrade the model outputs.

As with any model, the quality of the output is highly dependent on the quality of the input. We sought to address shortcomings in pre-processing to include how many techniques simply employ a standard stopwords package, use the full BoW, and require users to estimate the number of topics a-priori.

We identified the need to have a customized stopwords list for any dataset. The word cloud is used as a visualization tool to assist the user in creating the custom stopwords list via CUP, which can be an irritative process to reduce as much noise as possible. Additionally, a technique for identifying a TF-IDF range to narrow the BoW is used as an input into the LDA. This proposed PET is based on the total words used in the document. The CUP and PET approaches allow the LDA topic modeling technique to achieve greater utility.

We then explored extant methods used to help users determine the number of topics, k, for the topic modeling technique to populate. The requirement for the user to select a value for k assumes that the user has prior knowledge of the dataset. There are two challenges that exist with the current heuristics that were addressed with our heuristic: One, in graphical methods, which value should the user select if more than one peak exists? And two, users are expected to input different values of k to determine optimal scores; what range should the user select to test? LDA was selected to test our heuristic. The varying of k can cause the output to vary; therefore, it is important to provide a reliable method for the user to select k. Our developed heuristic based on the number of eigenvalues greater than one, using the term document matrix, provided more reliable results when compared to the popular graphing of coherence scores technique.

Finally, we proposed a new topic modeling technique called the *Zimm Approach*. LDA is a popular topic modeling technique; however, it requires the user to input the number of topics and the number of terms to output for the topics. In LDA, the number of terms per topic is the same. The *Zimm Approach* includes CUP, PET, and the eigenvalue heuristic to identify the keywords in each topic, which are derived from the eigenvector loadings. The *Zimm Approach* does not require the user to select a value for k and does not require the user to determine the number of terms for each topic. The new technique allows for a varying number of terms in each topic. Another advantage of the *Zimm Approach* is the stability of the algorithm. If you vary k, the terms do not change. For example, if k = 13, and then the user made $k = 37$, the first 13 terms of each topic for all ks will be the same. Whereas, when you vary k in LDA, the terms in the outputs will vary.

Existing techniques require the user to input parameters that have a direct impact on the output of the algorithm. This proposed topic modeling technique does not vary the terms associated with the topic, even if the user varies k. The number of terms the algorithm outputs with each term differs from term to term depending on the plot of the loadings. The topic modeling technique proposed in this article removes the requirement for those parameter inputs while providing a more stable output.

5.1. Limitations

Practitioners of NLP have powerful techniques available, but most rely on highly subjective parameter choices or *a priori* knowledge of the dataset. Subjectivity is a part of NLP. While our proposed techniques are designed to reduce some of the subjectivity involved with topic identification in a corpus, some will remain. Specifically, the word cloud portion of our process that informs the expanded stopword list remains subjective and requires some level of user interpretation. Future research may address this aspect of the topic identification process to further remove sources of subjectivity.

The analysis presented relies on multiple subsets of a single data set. While the analysis shows promising results, additional data sets should be explored to validate the robustness of the techniques.

Finally, the effectiveness of the *Zimm Approach* was measured through the coherence value metric. As indicated, there are multiple metrics available for the NLP practitioner, and the technique should be tested against additional metrics.

5.2. Future research

Topic modeling will continue to be an area of interest, and there are many areas for improvement. For example, techniques in this research used unigrams. Further research could look at multigrams to expand the concepts. In a similar vein, work initially presented in [67] and implemented through word2vec uses PCA in NLP large model training to predict words in a given context. Exploring the relationship to this work and similar works is a potential extension worthy of pursuit.

Additionally, this research focused on the LDA modeling technique. The techniques discussed could be applied among other topic modeling techniques such as latent semantic analysis and nonnegative matrix factorization.

The *Zimm Approach* outputs the topics and a list of terms for each topic. Future research could include creating a way other than a list for users to visualize the output. While the CUP technique retains the human in the data processing loop, requiring decisions to be made about the importance and usefulness of a word, future research should be conducted to create an algorithm to identify the words to enhance the stopwords list without the need for human entry.

Finally, the ultimate metric for evaluating topic modeling outputs is its usability. Coherence scores fluctuate and do not always align with human interpretability. Further research could develop or refine metrics for topic modeling with a greater focus on the usefulness of the output for decision-making and decision quality.

CRediT authorship contribution statement

Jamie Zimmermann: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Lance E. Champagne:** Writing – original draft, Visualization, Methodology, Conceptualization. **John M. Dickens:** Writing – original draft, Visualization, Methodology, Conceptualization. **Benjamin T. Hazen:** Writing – original draft.

Declaration of competing interest

The research meets all applicable standards for the ethics of experimentation and research integrity, and the following is being certified/declared true. None of the authors of this paper has a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the paper. It is to specifically state that "No Competing interests are at stake and there is No Conflict of Interest" with other people or organizations that could inappropriately influence or bias the content of the paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.dss.2024.114310.

References

[1] M. Mendoza, E. Alegría, M. Maca, C. Cobos, E. León, Multidimensional analysis model for a document warehouse that includes textual measures, Decis. Support Syst. 72 (2015) 44–59.

[2] A. Brahma, D. Goldberg, N. Zaman, M. Aloiso, Automated mortgage origination delay detection from textual conversations, Decis. Support Syst. 140 (2021) 113433.

[3] F. Tseng, A. Chou, The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence, Decis. Support Syst. 42 (2) (2006) 727–744.

[4] S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Decis. Support Syst. 50 (4) (2011) 680–691.

[5] C. Martins, M. Monard, E. Matsubara, Reducing the dimensionality of bag-of-words text representation used by learning algorithms, in: Proc of 3rd IASTED International Conference on Artificial Intelligence and Applications, 2003, pp. 228–233.

[6] H. Wallach, Topic modeling: beyond bag-of-words, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 977–984.

[7] Y. Zuo, H. Zhang, H. Lin, F. Wang, K. Xu, Topic modeling of short texts: A pseudo-document view, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery: Special Interest Group on Knowledge Discovery in Data, 2016, pp. 2105–2114.

[8] X. Li, Y. Wang, A. Zhang, C. Li, J. Chi, J. Ouyang, Filtering out the noise in short text topic modeling, Inform. Sci. 456 (2018) 83–96.

[9] I. Vayansky, S. Kumar, A review of topic modeling methods, Inf. Syst. 94 (2020) 101582.

[10] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[11] D. Slof, F. Frasincar, V. Matsiiako, A competing risks model based on latent Dirichlet Allocation for predicting churn reasons, Decis. Support Syst. 146 (2021) 113541.

[12] M. Gerlach, T. Peixoto, E. Altmann, A network approach to topic models, Sci. Adv. 4 (7) (2018) 1360.

[13] H. Misra, O. Cappé, F. Yvon, Using LDA to detect semantically incoherent documents, in: CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning, 2008, pp. 41–48.

[14] T. Hoffman, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (1) (2001) 177–196.

[15] S. Feuerriegel, N. Pröllochs, Investor reaction to financial disclosures across topics: An application of latent Dirichlet allocation, Decis. Sci. 52 (2021) 608–628.

[16] M. Hasan, A. Rahman, M. Karim, M. Khan, S. Islam, M. Islam, Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA), in: Proceedings of International Conference on Trends in Computational and Cognitive Engineering, Springer, 2021, pp. 341–354.

[17] R. Albalawi, T. Yeap, M. Benyoucef, Using topic modeling methods for short-text data: A comparative analysis, Front. Artif. Intell. 3 (2020) 42.

[18] D. Binkley, D. Heinz, D. Lawrie, J. Overfelt, Understanding LDA in source code analysis, in: Proceedings of the 22nd International Conference on Program Comprehension, ACM, 2014, pp. 26–36.

[19] G. Anthes, Topic models vs. unstructured data, Commun. ACM 53 (12) (2010) 16–18.

[20] M. Mustak, J. Salminen, L. Plé, J. Wirtz, Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda, J. Bus. Res. 124 (2021) 389–404.

[21] K. Singh, S. Devi, H. Devi, A. Mahanta, A novel approach for dimension reduction using word embedding: An enhanced text classification approach, Int. J. Inf. Manag. Data Insights 2 (1) (2022) 100061.

[22] Z. Yin, Y. Shen, On the dimensionality of word embedding, in: 32nd Conference on Neural Information Processing Systems, NeurIPS 2018, 2018, pp. 895–906.

[23] P. Kherwa, P. Bansal, A comparative empirical evaluation of topic modeling techniques, in: D. Gupta, A. Khanna, S. Bhattacharyya, A. Hassanien, S. Anand, A. Jaiswal (Eds.), International Conference on Innovative Computing and Communications, in: Advances in Intelligent Systems and Computing, vol. 1166, Springer, Springer, Singapore, 2021, pp. 289–297.

[24] D. Greene, D. O'Callaghan, P. Cunningham, How many topics? stability analysis for topic models, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2014, pp. 498–513.

[25] F. Hogenboom, F. Frasincar, U. Kaymak, F. De Jong, E. Caron, A survey of event extraction methods from text for decision support systems, Decis. Support Syst. 85 (2016) 12–22.

[26] M. Kraus, S. Feuerriegel, Decision support from financial disclosures with deep neural networks and transfer learning, Decis. Support Syst. 104 (2017) 38–48.

[27] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, H. Prendinger, Deep learning for affective computing: Text-based emotion recognition in decision support, Decis. Support Syst. 115 (2018) 24–35.

[28] B. Kim, J. Park, J. Suh, Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information, Decis. Support Syst. 134 (2020) 113302.

[29] J. Grimmer, A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases, Political Anal. 18 (1) (2010) 1–35.

[30] R. Zhao, K. Mao, Fuzzy bag-of-words model for document representation, IEEE Trans. Fuzzy Syst. 26 (2) (2017) 794–804.

[31] G. Lebanon, Y. Mao, J. Dillon, The locally weighted bag of words framework for document representation, J. Mach. Learn. Res. 8 (10) (2007).

[32] Y. Zhang, R. Jin, Z. Zhou, Understanding bag-of-words model: a statistical framework, Int. J. Mach. Learn. Cybern. 1 (1–4) (2010) 43–52.

[33] T. Geva, J. Zahavi, Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news, Decis. Support Syst. 57 (2014) 212–223.

[34] N. Passalis, A. Tefas, Entropy optimized feature-based bag-of-words representation for information retrieval, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1664–1677.

[35] W. Zhao, J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, W. Zou, A heuristic approach to determine an appropriate number of topics in topic modeling, BMC Bioinformatics 16 (13) (2015) 1–10.

[36] B. Ljungberg, Dimensionality reduction for bag-of-words models: PCA vs LSA, 2019, Retrieved from https://cs229.stanford.edu/proj2017/final-reports/5163902.pdf.

[37] C. Boulis, M. Ostendorf, Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams, in: Proceedings of the International Workshop in Feature Selection in Data Mining, Citeseer, 2005, pp. 9–16.

[38] Selva86, 20 Newsgroups training data, 2018, url: https://raw.githubusercontent.com/selva86/datasets/master/newsgroups.json.
[39] A. Schofield, D. Mimno, Comparing apples to apple: The effects of stemmers on topic models, Trans. Assoc. Comput. Linguist. 4 (2016) 287–300.
[40] V. Balakrishnan, E. Lloyd-Yemoh, Stemming and lemmatization: a comparison of retrieval performances, Lect. Notes Softw. Eng. 2 (3) (2014) 262–267.
[41] N. Razmi, M. Zamri, S. Ghazalli, N. Seman, Visualizing stemming techniques on online news articles text analytics, Bull. Electr. Eng. Inform. 10 (1) (2021) 365–373.
[42] A. Jivani, A comparative study of stemming algorithms, Int. J. Comput. Technol. Appl. 2 (6) (2011) 1930–1938.
[43] A. Rogers, A. Drozd, B. Li, The (too many) problems of analogical reasoning with word vectors, in: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, * SEM 2017, 2017, pp. 135–148.
[44] H. Eassom, How to choose effective keywords for your article, Discov. Future Res. (2017).
[45] B. Dahal, S. Kumar, Z. Li, Topic modeling and sentiment analysis of global climate change tweets, Soc. Netw. Anal. Min. 9 (1) (2019) 1–20.
[46] T. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (suppl 1) (2004) 5228–5235.
[47] Y. Xu, K. Heller, Z. Ghahramani, Tree-based inference for Dirichlet process mixtures, in: Artificial Intelligence and Statistics, PMLR, 2009, pp. 623–630.
[48] Q. Fu, Y. Zhuang, J. Gu, Y. Zhu, H. Qin, X. Guo, Search for K: assessing five topic-modeling approaches to 120,000 Canadian articles, in: 2019 IEEE International Conference on Big Data, IEEE, 2019, pp. 3640–3647.
[49] H. Abdi, L. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433–459.
[50] N. Péladeau, E. Davoodi, Comparison of latent Dirichlet modeling and factor analysis for topic extraction: A lesson of history, in: Proceedings of the 51st Hawaii International Conference on System Sciences, HICSS, 2018, pp. 615–623.
[51] K. Murphy, Naive Bayes Classifiers, Technical Report 18(60), University of British Columbia, 2006, pp. 1–8.
[52] A. Agrawal, W. Fu, T. Menzies, What is wrong with topic modeling? And how to fix it using search-based software engineering, Inf. Softw. Technol. 98 (2018) 74–88.
[53] M. Mantyla, M. Claes, U. Farooq, Measuring LDA topic stability from clusters of replicated runs, in: Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, IEEE, 2018, pp. 1–4.
[54] R. Fagin, R. Kumar, D. Sivakumar, Comparing top k lists, SIAM J. Discrete Math. 17 (1) (2003) 134–160.
[55] S. Syed, M. Spruit, Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation, in: 2017 IEEE International Conference on Data Science and Advanced Analytics, DSAA, IEEE, 2017, pp. 165–174.
[56] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler, Exploring topic coherence over many models and many topics, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 952–961.
[57] J.M. Campagnolo, D. Duarte, G. Dal Bianco, Topic coherence metrics: How sensitive are they? J. Inf. Data Manag. 13 (4) (2022) http://dx.doi.org/10.5753/jidm.2022.2181.
[58] D. Newman, J. Lau, K. Grieser, T. Baldwin, Automatic evaluation of topic coherence, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 100–108.
[59] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 262–272.
[60] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, 2015, pp. 399–408.
[61] D. Newman, Y. Noh, E. Talley, S. Karimi, T. Baldwin, Evaluating topic models for digital libraries, in: Proceedings of the 10th Annual Joint Conference on Digital Libraries, ACM, 2010, pp. 215–224.
[62] N. Aletras, M. Stevenson, Evaluating topic coherence using distributional semantics, in: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers, 2013, pp. 13–22.
[63] G. Bouma, Normalized (pointwise) mutual information in collocation extraction, in: Proceedings of GSCL, Vol. 30, 2009, pp. 31–40.
[64] D. Hamzeian, Using Machine Learning Algorithms for Finding the Topics of COVID-19 Open Research Dataset Automatically (Master's thesis), University of Waterloo, 2021.
[65] Y. Jiang, X. Song, J. Harrison, S. Quegan, D. Maynard, Comparing attitudes to climate change in the media using sentiment analysis based on Latent Dirichlet Allocation, in: Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism, 2017, pp. 25–30.
[66] K. Coussement, D. Benoit, Interpretable data science for decision making, Decis. Support Syst. 150 (2021) 113664.
[67] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, Vol. 26, Curran Associates, Inc., 2013.

Jamie Zimmermann (Ph.D. AFIT) is the 50th Force Support Commander at Schriever Space Force Base, Colorado. She received her Bachelor's degree in Applied Mathematics from Valdosta State University. She also attained her M.S. degree in Operations Research from the Air Force Institute of Technology. Her research efforts include textual analysis and machine learning. rjzim051609@gmail.com.

Lance E. Champagne (Ph.D. AFIT) is an Associate Professor of Operations Research at the Air Force Institute of Technology and Chair of the Modeling, Simulation, and Analysis Graduate Certificate Program. His research interests include simulation of autonomous system behavior, agent-based combat simulation, neural network design for image and video classification, and textual analysis. He is a member of the Cincinnati/Dayton Chapter of the Institute for Operations Research and Management Sciences (INFORMS) and the Military Operations Research Society (MORS). lance.champagne@afit.edu.

John M. Dickens (Ph.D. UNT) is an Associate Professor of Supply Chain Management at The Citadel, Charleston, South Carolina. He attained a B.S degree in History at the United States Air Force Academy and served in the Air Force on active duty for over 22 years. He received an M.S. degree in Logistics Supply Chain Management from AFIT, and a Ph.D. from the University of North Texas. His research interests include sustainment, value-creation, and supply chain modeling. E-Mail: jdicken2@citadel.edu.

Ben Hazen (Ph.D. Auburn) is an Assistant Professor of Operations and Supply Chain Management at University of Dayton and a retired Air Force maintenance and logistics officer. His expertise centers on organizational innovation diffusion and supply chain management, with a focus on how research and theory inform best practice applications. He is a founding editor emeritus of the *Journal of Defense Analytics and Logistics*and past editor of *International Journal of Physical Distribution and Logistics Management*. Ben serves at PW Communications as Chief Integration Officer of SHELDON, which leverages world-class data science tools to generate practical insights from open-source data to support US Government and Department of Defense decision-makers. He also serves on the boards of start-up technology companies that support defense research and applications. hazenscm@gmail.com