Information Processing and Management 59 (2022) 102921

Contents lists available at ScienceDirect

Information Processing and Management

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

Comparing PSO-based clustering over contextual vector embeddings to modern topic modeling

Samuel Miles [a], Lixia Yao [b], Weilin Meng [b], Christopher M. Black [b], Zina Ben Miled [a,c,*]

[a] Department of Electrical and Computer Engineering (IUPUI) Indianapolis, IN 46202, USA
[b] Merck & Co., Inc., 2000 Galloping Hill Road, Kenilworth, NJ, 07033, USA
[c] Regenstrief Institute, Inc., 1101 W. 10th Street, Indianapolis, IN 46202, USA

ARTICLE INFO

Keywords:
Topic modeling
Clustering
Vector embedding
PSO
ETM
NVDM

ABSTRACT

Efficient topic modeling is needed to support applications that aim at identifying main themes from a collection of documents. In the present paper, a reduced vector embedding representation and particle swarm optimization (PSO) are combined to develop a topic modeling strategy that is able to identify representative themes from a large collection of documents. Documents are encoded using a reduced, contextual vector embedding from a general-purpose pre-trained language model (sBERT). A modified PSO algorithm (pPSO) that tracks particle fitness on a dimension-by-dimension basis is then applied to these embeddings to create clusters of related documents. The proposed methodology is demonstrated on two datasets. The first dataset consists of posts from the online health forum r/Cancer and the second dataset is a standard benchmark for topic modeling which consists of a collection of messages posted to 20 different news groups. When compared to the state-of-the-art generative document models (i.e., ETM and NVDM), pPSO is able to produce interpretable clusters. The results indicate that pPSO is able to capture both common topics as well as emergent topics. Moreover, the topic coherence of pPSO is comparable to that of ETM and its topic diversity is comparable to NVDM. The assignment parity of pPSO on a document completion task exceeded 90% for the 20NewsGroups dataset. This rate drops to approximately 30% when pPSO is applied to the same Skip-Gram embedding derived from a limited, corpus-specific vocabulary which is used by ETM and NVDM.

1. Introduction

Identifying and interpreting the latent topics present in a given collection of documents is an open area of research that has been active over the last few decades, with much of the most recent developments centered on probabilistic generative models. This process is traditionally known as topic modeling and is used in a wide variety of applications ranging from topic mining of electronic petitions (Hagen, 2018), personalizing users' experiences in recommender systems (Rajendran & Sundarraj, 2021), to the analysis of the sentiments of patients towards their health care providers (Shah, Yan, Tariq, & Ali, 2021). Extensive reviews of topic modeling applications and techniques are available in Chauhan and Shah (2021) and Churchill and Singh (2021).

Most of the current research in topic modeling focuses on the use of probabilistic generative models (Armstrong, Maupomé, & Meurs, 2021; Dieng, Ruiz, & Blei, 2019; Gupta & Zhang, 2021; Liu, Xia, Li, Yan, & Liu, 2020; Zhang, Liu, & Zhang, 2020). However, there are few topic modeling techniques that explore an alternative approach based on document clustering. The advantages of clustering techniques compared to generative models are often reduced computation complexity and increased scalability (Dieng,

* Corresponding author at: Department of Electrical and Computer Engineering (IUPUI) Indianapolis, IN 46202, USA.
E-mail address: zmiled@iupui.edu (Z.B. Miled).

https://doi.org/10.1016/j.ipm.2022.102921
Received 3 November 2021; Received in revised form 25 January 2022; Accepted 24 February 2022
Available online 14 March 2022

Ruiz, & Blei, 2020). For instance, K-Means was shown to generate comparable clusters to those produced by LDA over a corpus of Twitter posts (Tijare & Prathuri, 2022). In Viegas, Canuto, Gomes, Luiz, Rosa, Ribas, Rocha, and Gonçalves (2019), a weighted TF-IDF is first used to generate topic keywords. These keywords form the basis of a vector representation of the text which is in turn used to cluster documents according to their cosine similarity. In the present paper, an evolutionary clustering approach to topic modeling is also pursued.

A contextual representation of the documents in the form of vector embeddings is first generated from a general-purpose pre-trained language model. Particle Swarm Optimization (PSO) (Kennedy & Eberhart, 1995) is then applied to the resulting embeddings to group the documents into semantically coherent clusters. The proposed approach is compared to current state-of-the art generative topic models produced by ETM (Dieng et al., 2020) and NVDM (Miao, Yu, & Blunsom, 2016) using a collection of posts from the r/Cancer (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020) health forum and the 20NewsGroups (Lang, 1995) dataset,[1] a standard benchmark in topic modeling which was previously used to evaluate both ETM and NVDM.

The above described topic modeling methodology was initially introduced in Miles, Yao, Meng, Black, and Ben-Miled (2021), where the primary focus of the study was to understand the experiences of the cancer patients outside of a clinical environment. For the purpose of this earlier study, PSO was only applied to posts from r/Cancer, an online health forum where users can share their own experiences with cancer as a patient or a caregiver. While this application provided the initial motivation, the results in the present study show that the proposed methodology can be effective for other domains.

2. Related work

In general, topic modeling is often performed in two steps. The first step consists of translating the unstructured text to a numerical representation. The second step uses this representation to group together similar documents. There have been several research efforts towards improving both of these two steps.

2.1. Text processing and representation

The goal of text representation is to transform the unstructured documents into a structured form that can be used for topic modeling. An example of a document processing pipeline is described in Karl, Wisnowski, and Rushing (2015). This pipeline performs several normalization operations including removal of low frequency, high frequency and unimportant words (e.g., stop words and words with more than 30 characters in length), forcing the text to lowercase, and removal of numbers. These types of pre-processing operations are common to most text mining (Karl et al., 2015) and topic modeling (Blei, Ng, & Jordan, 2003; Dieng et al., 2020) applications.

Once pre-processing is completed, each document is encoded as a vector of numerical values. Under the Bag of Words (BoW) representation, each entry in the vector corresponds to the presence or absence of a given word in the document (Zhang, Jin, & Zhou, 2010). This text representation ignores the context of the word and suffers from multiple limitations including the ambiguity of polysemic words (Curiskis, Drake, Osborn, & Kennedy, 2020).

Recent topic modeling techniques rely on a new vector representation that takes into account context and where the word encoding is learned from an encoder/decoder network. These encodings are commonly referred to as vector embeddings (Devlin, Chang, Lee, & Toutanova, 2019). The use of these embeddings varies considerably. Some topic modeling approaches develop the embeddings from a BoW representation during model training as in the case of the NVDM and ETM models. Others may start with a pre-trained embedding text representation which is fine-tuned during modeling training. For instance, ETM+PWE is a variant of ETM which starts with a pre-trained Skip-Gram embedding. In either of the above cases the embeddings are corpus-specific.

The embedding that is adopted in this paper is sentence-BERT (sBERT) (Reimers & Gurevych, 2019). This is a general-purpose language model that was pre-trained to capture semantic similarities between a pair of sentences. This language model was selected because its emphasis on sentence similarity aligns well with topic modeling. The sBERT is a sequence embedding model that takes into consideration the context of the word within the sequence. Therefore, a word may have a different embedding depending on the context. In contrast, Skip-Gram, the embedding model used in ETM+PWE, is a word embedding. That is, the embedding assigned to each word does not vary based on the context. For consistency, this variant of ETM is labeled ETM_SG in the remainder of the present paper.

2.2. Topic modeling

A topic is traditionally defined as a distribution of all the words in a vocabulary from which every term in every document of a given corpus is collected. In turn, documents are considered as a distribution of topics which can be used to model the topic proportions of an entire corpus (Blei et al., 2003). Collectively, these two distributions constitute the topic model for the corpus. Therefore, topic modeling aims at deriving an understanding of the distribution of a dataset of documents and how each sample document within the dataset relates to one another. This aim can be achieved using two main types of models: generative and evolutionary.

[1] http://qwone.com/~jason/20Newsgroups/.

2.2.1. Generative models

One of the earliest generative topic modeling technique is LDA (Blei et al., 2003). LDA suffers from the difficulty of application-specific hyper-parameter fine-tuning (Nzali, Bringay, Lavergne, Mollevi, & Opitz, 2017) and prohibitive time complexity when applied to a corpus with a rich vocabulary (Dieng et al., 2020). The limited ability of LDA to handle large vocabularies can be partially attributed to the high dimensionality of the BoW vector representation (Karl et al., 2015).

The embedded topic model (ETM) (Dieng et al., 2020) builds on the combination of LDA and the variational auto-encoder (VAE). The resulting word embeddings are similar to the CBOW word embeddings (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Except in the case of ETM, the context vector is replaced by an assigned topic vector. There are two implementations of ETM: The native ETM simultaneously learns its own embeddings for both topics and words; and ETM_SG uses the pre-trained Skip-Gram word embeddings.

The second generative topic model under consideration in the present paper is NVDM (Miao et al., 2016). This model starts with the BoW text representation and develops an embedding for the entire document. This step is followed by a softmax decoder that can infer the words in the document from its embedding. The embedding generated by NVDM is similar to sBERT since both models encode the entire document. In contrast, ETM encodes each word in the document independently.

2.2.2. Evolutionary models

The generative topic modeling algorithms described above are derived from the probability distribution of the words and topics across the corpus of available documents. These models tend to incur a high computational cost due to the calculation of the posterior joint probability distribution. In contrast, evolutionary algorithms use a heuristic approach to topic modeling. These algorithms search the input space for adequate topics using optimization techniques that are typically more computationally efficient than probabilistic models. Despite the computational advantage of evolutionary models, previous studies indicate that the quality of evolutionary and generative topic models may be comparable (Curiskis et al., 2020; Rashid, Shah, & Irtaza, 2019; Tijare & Prathuri, 2022).

Examples of evolutionary algorithms include K-Means, fuzzy clustering (George, Wang, Wilson, Epstein, Garland, & Suh, 2012) and PSO (Cui, Potok, & Palathingal, 2005). As in the case of generative models, traditional evolutionary models for topic modeling relied on the BoW representation of the text (Chen, Qin, Liu, Liu, & Li, 2010; Karol & Mangat, 2013; Rodrigues & Sacks, 2004). More recent studies investigate the use of vector embeddings for text representation with evolutionary topic models. In Curiskis et al. (2020), K-Means over Skip-Gram embeddings were found to outperform LDA with BoW. Comparison with ETM and NVDM was not explored. Other evolutionary techniques over vector embeddings have yet to be fully investigated. In a previous study (Miles et al., 2021), the authors of the present paper demonstrated that evolutionary techniques based on PSO and vector embeddings can lead to coherent and diverse topic models for a set of posts from r/Cancer. This evolutionary technique is extended in this study to the 20NewsGroups dataset and compared to the generative ETM and NVDM topic models.

3. Objectives

The primary objective of the present study is to compare the performance of a topic modeling approach based on a modified PSO technique to the widely used generative models ETM and NVDM. The secondary objective is to understand variances in the quality of the topic models developed by the proposed approach when applied to different domains and embeddings. Specifically, this study compares the proposed PSO approach to ETM and NVDM under different text representations in the context of topic extraction from (1) a subject-specific collection of documents (cancer) and (2) a diversified set of text from 20 news groups.

4. Methods

The present paper describes an evolutionary topic modeling technique that applies PSO to contextual vector embeddings of unstructured text documents. The technique is evaluated using two datasets and the results are compared to those produced by ETM and NVDM.

4.1. Datasets

The first dataset consists of r/Cancer posts from the Reddit archive (Baumgartner et al., 2020), and the second dataset consists of posts submitted to 20 news groups (Lang, 1995). Some of the differences between these two datasets include the primary domain and the vocabulary size.

Online health forums often chronicle the patient's journey and provide an alternative to surveys. Data from these forums were used for several previous applications including the prediction of medication adherence (Haas, Ben Miled, & Mahoui, 2019), the tracking of the spread of infectious diseases (Agbehadji, Awuzie, Ngowi, & Millham, 2020; Freifeld, Brownstein, Menone, Bao, Filice, Kass-Hout, & Dasgupta, 2014), the identification of common allergy types (Lee, Agrawal, & Choudhary, 2015) and the development of a better understanding of the consequences of a dementia diagnosis on the caregivers (Gkotsis, Mueller, Dobson, Hubbard, & Dutta, 2020).

The posts in this dataset were extracted from the r/Cancer Reddit archive (Baumgartner et al., 2020). The training dataset used for all models under investigation consists of all posts from January to December of 2020. The testing set consists of randomly

selected posts from 2021 and makes up approximately 20% of the entire dataset. For both test and train datasets, the title and the body of each post are combined.

The 20NewsGroups dataset was selected because it was also used in the evaluation of ETM and NVDM. The version of this dataset that is used in the present paper is originally split into a test and a train subsets with no headers, footers and quotes. The latter fields were omitted to reduce any potential bias.

Since topic modeling can be sensitive to pre-processing, a streamlined approach was adopted for both datasets and across all models. All instances of punctuation and number symbols are removed. The text is forced to lowercase. Finally, stop words as identified in Bird, Klein, and Loper (2009) and terms with a document frequency greater than 30% are also removed.

After preprocessing, the documents are transformed into an embedding representation using sBERT (Reimers & Gurevych, 2019). An element-wide average of all the vector embeddings is computed and used to represent the entire document. For a given document j, e_j is used to denote this vector embedding. The size of the vector embedding (e_j) produced by sBERT consists of 768 numerical features. Dimension reduction is performed on e_j. Topic clustering using high-dimension vector embeddings was previously shown to have lower performance compared to lower dimension embeddings (Aggarwal, Han, Wang, & Yu, 2004). In this study, the Uniform Manifold Approximation and Projection (UMAP) technique is used for dimension reduction (McInnes, Healy, Saul, & berger, 2018). UMAP reduces the number of features in an embedded vector based on the values of the surrounding terms given a specified global structure value. Testing was previously done on several dimension sizes ranging from the 768 directly generated by sBERT down to 30 (Miles et al., 2021). It was empirically established that a reduced embedding with a dimension of 300 had the best performance for the proposed PSO model (Miles et al., 2021). This dimension was also the dimension of choice for ETM (Dieng et al., 2020).

For comparison purposes, PSO is also applied to document embeddings generated by Skip-Gram with a dimension of 300 and context window of 4. The sBERT embedding is only restricted by the vocabulary of the general-purpose corpus originally used to train this language model. Therefore, it is not corpus-specific. However, for both ETM and NVDM the embeddings are corpus specific and limited to the most frequent words. The same vocabulary was used for ETM, ETM_SG, NVDM, and the Skip-Gram version of PSO. The vocabulary size was set to 2,000 words.

4.2. Models

PSO is an evolutionary algorithm. It is adapted in this study to the clustering of vector embeddings generated from a collection of text documents. PSO simulates the movement of individual particles in the embedding space in search of the best representative vector embedding for each cluster of documents. The vector embeddings of K documents are initially randomly selected from the entire set of m available documents. Each of these selected documents corresponds to the initial position of a particle i, and represents a cluster. Thus, the total number of particles/clusters is fixed to K. The position of the particle, $x_i(t)$, is initialized to the embedding of the randomly selected document from the corpus (i.e., $x_i(0) = e_i$). As the particles move through the embedding space, their ability to represent all of the m documents is evaluated using a fitness function. First, each document e_j in the dataset is assigned to one of the particles/clusters according to Eq. (1). A cluster C_i consists of the set of documents e_j that are assigned to it as defined in Eq. (2). The fitness of particle i is then computed according to Eq. (3) which compares the current position of the particle $x_i(t)$ to the embedded representation of each document assigned to C_i.

$$clust(e_j) = argmin_{1\le i\le K}\{d(e_j, x_i(t))\} \tag{1}$$

$$C_i(t) = \{clust(e_j)_{1\le j\le m} = i\} \tag{2}$$

$$f_i(t) = \frac{\sum_{j\in C_i(t)} d(e_j, x_i(t))}{\|C_i(t)\|} \tag{3}$$

The distance d in Eq. (1) and Eq. (3) can take on several forms including Euclidean distance and cosine similarity. In the present study, cosine similarity is used as a distance measure. During each iteration t, the next velocity $v_i(t+1)$ and next position $x_i(t+1)$ for each particle i are computed using Eq. (4) and Eq. (5), respectively. In Eq. (4), L_i represents the best position that the particle has discovered thus far. This term is often referred to as "local" best. The best position among all the particles is denoted G in Eq. (4) and called the "global best", where ω, μ and ρ are hyper-parameters. The latter two hyper-parameters are referred to as the local and global conscience, respectively.

$$v_i(t+1) = \omega v_i(t) + \mu(L_i - x_i) + \rho(G - x_i) \tag{4}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{5}$$

The present study introduces a method adapted to vector embeddings for calculating the "best" position for each particle. Each dimension in the embedding vector is evaluated independently and the best value is retained for a given dimension without any consideration to other dimensions in the vector. This point-wise update allows the particles to find the best fit for each dimension individually as opposed to forcing all dimensions to collectively move in the same direction. This modified point-wise PSO algorithm (pPSO) will generate K clusters where each cluster is a main theme represented by the final position of the particle (i.e., the centroid of the cluster).

4.3. Topic evaluation

Topic models are evaluated both quantitatively and qualitatively using topic words. The qualitative evaluation can be subjective. The quantitative evaluation can be difficult (Hailu, Yu, & Fantaye, 2020) and is an open research problem (Nikolenko, Koltcov, & Koltsova, 2017). Moreover, the extraction of topic words is different in generative compared to evolutionary topic models.

In generative topic modeling, topics are considered as a distribution of words and documents are considered as a distribution of the topics. These distributions are learned simultaneously through an iterative procedure that optimizes the likelihood of terms belonging to certain topics and the likelihood that a document expresses a particular set of topics. Therefore, topic words are identified as part of the learning process (Blei et al., 2003; Dieng et al., 2020; Miao et al., 2016).

In contrast, for evolutionary models such as PSO, topics correspond to positions in the embedding space and documents are assigned to a topic based on their relative distance to these positions. Therefore, once the clusters of documents are constructed, document frequency inverse cluster frequency (DF-ICF) is used to identify a set of words that represent the topic. DF-ICF is similar to term-frequency inverse document frequency (TF-IDF) (Ramos et al., 2003; Salton & Buckley, 1988). It assigns higher scores to words that appear often within a particular cluster, but not as often in the remaining clusters. For a given word w, DF is calculated as the frequency of documents within the corpus that contain w and ICF is then given by the following equation.

$$ICF(w, C) = log \frac{K}{|\{D \in C : w \in D\}|} \tag{6}$$

where $|\{D \in C : w \in D\}|$ is the number of clusters that include documents containing the word w.

Evaluation metrics for topic models include coherence and diversity (Dieng et al., 2020). Topic coherence (TC) is a measure of how representative a collection of topic words is for a given set of documents (Dieng et al., 2020; Mimno, Wallach, Talley, Leenders, & McCallum, 2011). TC is defined as the normalized summation of point-wise mutual information (PWI) (Lau, Newman, & Baldwin, 2014) for two terms drawn from the same document. TC values can range between −1 and 1, with values closer to 1 indicative of more coherent clusters and values closer to −1 indicative of less coherent clusters. In the present paper, the top 10 words are selected to represent a cluster and the corresponding TC expression is shown in Eq. (7). This exact metric was used to evaluate ETM (Dieng et al., 2020).

$$TC = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{45} \sum_{i=1}^{10} \sum_{j=i+1}^{10} PWI(w_i^{(k)}, w_j^{(k)}) \tag{7}$$

$$PWI(w_i, w_j) = \frac{log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-log P(w_i, w_j)} \tag{8}$$

where $\{w_1^{(k)}, \ldots, w_{10}^{(k)}\}$ are the top-10 topic words for a given cluster C and K is the number of clusters. $P(w_i, w_j)$ is the number of documents in a given cluster in which the pair of words w_i and w_j co-occur, whereas $P(w_i)$ and $P(w_j)$ are the number of documents within a given cluster in which the respective words appear at least once. The probabilities in Eqs. (7) and (8) are approximated using a count of the number of documents.

Topic diversity (TD) is a measure of the uniqueness of the topic words for a given cluster when compared to the top words for the entire corpus (Dieng et al., 2020). In other words, it evaluates the number of topic words associated with a given topic that are also selected for other topics. TD values closer to 0 indicate that the topic words are redundant and multiple clusters might be choosing similar topic words; whereas TD values closer to 1 suggest that the topic words are more varied and unique to a particular cluster.

Another metric that is used to evaluate generative topic models is the log likelihood on document completion (DC) (Wallach, Murray, Salakhutdinov, & Mimno, 2009). As opposed to TC and TD, this metric is calculated on a test dataset that is not observed during the development of the model. Each document in the test dataset is split in two halves and assigned to a separate test dataset. The topic assignments of the test dataset that contains the first portion of the document is compared to the topic assignment of the second dataset. Ideally, a complete match is expected since each of the two halves are extracted from the same document. Briefly, DC is the sum of the log loss of the difference between the topic distributions of the first half and the second half of the document. This metric can be applied to generative models since they produce a topic distribution for each document. However, clustering methods such as PSO assign a document to a single cluster. This hard assignment cannot be evaluated using the log likelihood. Therefore, an equivalent parity measure is introduced. Given a document which is split into two halves, h_1 and h_2, the document parity is set to one if both h_1 and h_2 are assigned to the same cluster and zero otherwise. The parity (PR) of the test dataset is the normalized sum of all the parities of the documents as shown below

$$PR = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} 1 \quad \text{if} \quad clust(h_{1i}) = clust(h_{2i}) \tag{9}$$

5. Results

This section includes a descriptive summary of the characteristics of the two datasets (i.e., r/Cancer and 20NewsGroups) under consideration in the present study as well as an evaluation of the topic models induced by these datasets using pPSO_SB, pPSO_SG, and the generative models ETM, ETM_SG and NVDM.

Table 1
Example documents after preprocessing from the r/Cancer and 20NewsGroups datasets.

Dataset	Text
r/Cancer	*run families dad grandpa different types dad lung grandpa prostate think risk coincidence*
	survivors first steps took youve received news youd stay positive things family friends worked broad question trying make sense fo whole thing dad position want help create support tools makes sense situation hes second te battling first prostate beat liver
20News Groups	*recently bought pack prospect hockey cards various players coming nhl got particular card russian named viktor kozlov says many scouts believe pick another guy quoted saying good mario lemieux anyone know guy*
	exactly space effects remain first rate even today later andersons tried shed reputation creators worst pseudo scientific shows tv history flying infinity one thing done part bbc educational sf series day tomorrow anderson episode dealt spaceship capable reaching speed light lightship altares four man crew eventually journeyed black hole ended far side galaxy think saw year old back liked much fan space guess easily satisfied days anyone know infinity released video space shows vhs know thunderbirds etc also available england

Table 2
Number of documents, unique words, words per document and characters per document for the r/Cancer and 20NewsGroups train and test datasets.

	r/Cancer		20NewsGroups	
	train	test	train	test
Number of documents	6,452	1,693	10,996	7,298
Number of unique words	20,415	11,552	69,881	52,677
Mean Length (words)	77.78	79.94	106.88	94.31
Std. Length (words)	73.55	76.81	392.21	280.55
Mean Length (characters)	521.82	539.28	717.43	659.42
Std. Length (characters)	664.73	693.62	2300.31	1,999.04

5.1. Datasets

Table 1 shows two sample documents from each dataset after preprocessing. Short documents were purposely selected as examples for ease of presentation. Some of the documents may actually span multiple page even after preprocessing. Aggregate descriptive statistics for the two datasets are included in Table 2. This table shows that the 20NewsGroups dataset includes more documents than the r/Cancer dataset. The mean number of words and characters of the 20NewsGroups documents are also larger than those of the r/Cancer posts.

5.2. Model evaluation

The evaluation of the models is performed using both qualitative and quantitative measures. First, a subjective qualitative analysis of the chosen topic words for each identified topic in both datasets is performed across the models. Second, a quantitative evaluation is reported using the topic modeling metrics TC, TD, DC, and PR.

Tables 3 and 4 show the top five topic words for each topic in the r/Cancer and 20NewsGroups datasets, respectively. A review of the results in Tables 3 and 4 indicates that many of the topics identified by pPSO_SB are coherent for both datasets. For instance, topics (1) and (3) generated by pPSO_SB for the r/Cancer dataset are centered around prostate cancer and food, respectively. In fact, the two example r/Cancer posts in Table 1 were assigned to topic (1). The focus of topic (5) is on hair loss. Topic (9) represents an emerging topic consisting of a large number of posts recruiting candidates for clinical trials. With the same dataset, ETM_SG produced a breast cancer topic (6) and a support topic (8). Some of the topics identified by ETM and ETM_SG are also similar (e.g., 9 & 5, 2 & 6). Several of the topics produced by NVDM are also interpretable but some are overlapping (e.g., 4 & 5, 7 & 9). This overlap is also observed for pPSO_SG (e.g., 9 & 10). However, the topics generated by pPSO_SG are in general different from those produced by pPSO_SB.

In the case of the 20NewsGroups dataset, the topics produced by pPSO_SB include hockey (2), space (8) and religion (9). The first example in Table 1 belongs to the hockey topic and the second example belongs to the space topic. Topic 3 contains documents of bitmap images converted into text from one of the computer newsgroups. These documents include several two-character symbols. This topic was also generated by ETM (8) and ETM_SG (2). However, it was neither identified by pPSO_SG nor NVDM.

As in the case of the r/Cancer dataset, the topics produced by ETM and ETM_SG share some similarities. Also the topics produced by NVDM and pPSO_SG are interpretable but tend to overlap (e.g., NVDM: 1 & 2, 4 & 6; pPSO_SG: 2 & 10, 5 & 7).

In summary, the qualitative evaluation of the topics indicates that the topics produced by pPSO_SB are relatively more interpretable than ETM and show less overlap than NVDM.

The topic coherence (TC) and topic diversity (TD) for the two datasets are shown in Tables 5 and 6, respectively. TC is calculated with the top 10 words and TD is calculated with the top 25 words. The TC produced by PSO and ETM are similar. The lowest TC is observed with NVDM for both datasets. These observations align with the qualitative results of Tables 3 and 4.

The lowest TD was achieved by ETM and the best TD was achieved by NVDM. The TD values of both variants of PSO are comparable. Moreover, they are similar to those of NVDM as the number of topics increases.

Table 3
Top 5 topic words identified in the r/Cancer dataset by pPSO_SB, pPSO_SG, ETM, ETM_SG and NVDM when the number of topics is set to 10.

1	2	3	4	5	6	7	8	9	10
pPSO_SB									
Dads	Hospice	Taste	Helpless	Wigs	Sensation	Hospice	Dads	Participants	Surgeon
Hes	Sister	Food	Medicine	Wig	Jaw	Moms	Hes	Survey	iv
Prostate	Passed	Meals	Nurses	Shave	Male	Metastasized	Hospice	Interview	Says
Father	Leave	Meal	Passed	Thick	Headaches	Palliative	Boyfriend	Adults	Experiences
Man	Daughter	Diet	Says	Shaving	Ear	Sister	Father	Email	Follow
pPSO_SG									
Sleeps	Graduate	Foods	Transplant	Infusion	Diarrhea	Survey	Mastectomy	Ultrasound	mm
Goodbye	Proud	Meal	Survived	Eating	Sore	Drive	Goodbye	Armpit	Tissue
Eating	ing	cbd	Higher	Appetite	Sweats	Link	ing	Lumps	Ultrasound
Gonna	Student	Eats	Stem	Session	Lower	Email	Grandmother	Private	Showed
Wake	Summer	Eating	Think	Transplant	Symptom	Resources	Supportive	Showed	Scheduled
ETM									
Pain	Scan	Back	im	Help	Feel	Stage	Mom	Chemo	Help
Also	Doctor	Hospital	Dont	Treatment	Life	Diagnosed	Dad	Treatment	Want
Symptoms	Biopsy	Told	ive	Would	Want	Liver	Time	Radiation	Family
Day	Breast	got	Feel	Anyone	People	Lung	Last	Anyone	Would
Started	ct	Surgery	Cant	Please	Really	Dad	Much	Hair	Advice
ETM_SG									
Would	Life	Mom	Chemo	Chemo	Scan	Pain	Help	Diagnosed	im
Want	Never	Dad	Back	Anyone	Tumor	Back	Would	Stage	Dont
Sure	One	Hospital	Time	Treatment	Surgery	Doctor	Support	Years	Feel
Told	Much	Home	First	Hair	Breast	Blood	Please	Ago	Really
Think	Years	Time	Last	Radiation	ct	Symptoms	Need	Year	Want
NVDM									
nsclc	Treatment	ent	Hair	Donate	Breast	us	Red	Eat	Want
Keytruda	School	Ultrasound	Happy	Leukemia	Young	Dad	Swelling	Eating	Doesnt
Cannabis	Meds	Ordered	Lucky	Symptoms	Mother	Hospice	Skin	Hospice	Immune
Survey	Name	Nose	Beautiful	Hair	Covid	Hospital	Increase	Dad	Wants
Trials	Pain	Contrast	Life	brca	Study	Father	Followed	Bed	Dont

When tested on a document completion task, NVDM has higher DC values than either variant of ETM for both datasets (Table 7). The metrics DC and PR cannot be directly compared. However, Table 7 shows that PSO_SB assigns the two halves of a single document to the same cluster with a percentage greater than 78% for the 20 NewsGroups. For the r/Cancer dataset, the parity increases as the number of topic increases. The topic parity for PSO_SG is significantly lower than the corresponding parity for PSO_SB.

6. Discussion

The present paper introduces an evolutionary topic modeling methodology based on PSO over a general-purpose sequence embedding representation of the input text. In general, evolutionary topic modeling techniques have not received adequate attention in the literature. PSO as a potential topic modeling technique was explored even less. Most previous topic modeling techniques are based on generative probabilistic models.

The topic models produced by PSO are compared to two generative topic models: ETM and NVDM. Several findings can be established from this comparison. First, previous studies show that embedded vector representation are more suitable for topic modeling than BoW for generative topic models (Armstrong et al., 2021; Bianchi, Terragni, & Hovy, 2021; Dieng et al., 2020; Gupta & Zhang, 2021). For evolutionary topic models, this aspect was only confirmed for K-Means (Curiskis et al., 2020). This paper extends this finding to PSO. Other evolutionary algorithms with embedded vector representation, such as fuzzy clustering, remain to be explored.

Second, the study confirms previous qualitative results established by other researchers. For example, ETM over SkipGram embedding outperforms ETM with respect to topic coherence, topic diversity and document completion. Moreover, NVDM outperforms ETM when evaluated on a document completion task (Dieng et al., 2020).

Third, some previous studies indicate that evolutionary models are strong contenders for traditional generative models. For example, K-Means with vector embedding was compared to LDA with BoW in Curiskis et al. (2020). Similarly, fuzzy clustering was shown to have better performance than LDA in Rashid et al. (2019) over BoW encoded input text. A comparison between a generative model and an evolutionary model over an embedded vector space was not established. The present paper shows that the TC of PSO over a general-purpose embedding space is similar to that of ETM over a corpus-specific embedding space. The TD of PSO is also comparable to the TD of NVDM. On a document completion task, PSO over a general-purpose embedding can have a parity that exceeds 90%.

Table 4
Top 5 topic words identified in the 20NewsGroup dataset by pPSO_SB, pPSO_SG, ETM, ETM_SG and NVDM when the number of topics is set to 10.

1	2	3	4	5	6	7	8	9	10
pPSO_SB									
Beauchaine	nhl	bq	dos	Bike	Encrypted	Israeli	Orbit	Bible	dos
Encryption	Hockey	bf	Graphics	Ride	Encryption	Jews	Shuttle	Christians	Monitor
Manhattan	Bruins	tq	Appreciated	Engine	Escrow	Israel	Launch	Christ	Computer
Sank	Rangers	qq	Advance	Car	Wiretap	Crime	Moon	Jesus	pc
Queens	Playoff	ei	Interested	Cars	Clipper	Weapons	Earth	Religion	Disk
pPSO_SG									
Morality	Modem	Test	Replies	Crypto	ide	Crime	Water	Heaven	Microsoft
Beliefs	Fax	Output	Hello	Government	mb	Guns	Nuclear	Armenian	dos
Christianity	dos	Research	Student	Firearms	Floppy	Society	Fuel	Armenians	Directory
Atheists	Shipping	Looks	ftp	Federal	scsi	Government	Cold	Turks	cica
Meaning	Directory	Common	Friend	Crime	meg	State	Food	Muslims	Edit
ETM									
Would	Space	People	Key	File	One	Drive	ax	God	Game
Anyone	New	Would	Use	Program	Like	Card	Max	One	Year
Know	University	Said	One	Image	get	Windows	gv	Would	Team
Thanks	Research	One	System	Window	Would	System	bf	People	First
Please	Also	Us	Public	Use	Good	Disk	pl	Think	Last
ETM_SG									
Would	ax	Year	File	Key	God	People	Drive	Car	Space
One	Max	Game	Program	Available	One	Said	Thanks	Good	Government
Know	gv	Team	Use	Information	Jesus	Armenian	Card	May	Gun
Like	bf	Play	Window	Mail	Believe	Israel	Windows	Used	President
Think	pl	First	Output	System	Would	War	System	Like	New
NVDM									
Israeli	Israeli	Printer	God	Security	God	Car	Program	Turks	Players
Arab	Israel	Windows	One	Bill	One	Drive	dos	Genocide	Jesus
Israel	Armenians	Mode	Would	Police	Someone	Bus	Build	Calgary	Clipper
Islamic	Armenia	Error	People	Federal	Believe	Team	package	pp	Team
rsa	Turks	vga	may	fbi	Something	Speed	Funds	Greek	Game

Table 5
Topic coherence for the r/Cancer and 20NewsGroups train datasets over the top 10 topics.

	r/Cancer			20NewsGroups		
K	10	20	30	10	20	30
PSO_SB	0.15	0.19	0.18	0.26	0.31	0.34
PSO_SG	0.14	0.14	0.17	0.28	0.32	0.32
ETM	0.14	0.16	0.16	0.28	0.33	0.35
ETM_SG	0.15	0.17	0.16	0.28	0.36	0.39
NVDM	0.05	0.14	0.15	0.14	0.23	0.17

Table 6
Topic diversity for the r/Cancer and 20NewsGroups train datasets over all topics.

	r/Cancer			20NewsGroups		
K	10	20	30	10	20	30
PSO_SB	0.76	0.69	0.73	0.82	0.74	0.75
PSO_SG	0.75	0.74	0.76	0.85	0.71	0.76
ETM	0.62	0.38	0.34	0.73	0.65	0.58
ETM_SG	0.70	0.60	0.37	0.83	0.76	0.64
NVDM	0.93	0.81	0.74	0.86	0.80	0.75

Fourth, using a sequence embedding can help improve the performance of the topic model on a document completion task. This is evident when comparing the PR of PSO_SB to that of PSO_SG as well the DC of NVDM compared to ETM_SG. In the latter case, both NVDM and ETM_SG use the same vocabulary. However, ETM_SG relies on a word-based embedding and NVDM uses a sequence embedding.

Some of the limitations of the proposed approach are with respect to the embedding representation. One of the benefits of the proposed evolutionary approach is that it uses a general-purpose embedding derived from a large vocabulary. Therefore, it can be directly applied to any corpus. In contrast, ETM and NVDM use a corpus-specific embedding. However, the dimension of the sBERT embedding had to be reduced using UMAP. This step can be avoided if sBERT is trained to produce an embedding with the

Table 7
Log likelihood on document completion and parity for the r/Cancer and 20NewsGroups test datasets .

		r/Cancer			20NewsGroups		
K		10	20	30	10	20	30
PR (%)	PSO_SB	46.37	39.34	61.96	78.51	99.01	99.75
	PSO_SG	24.63	19.73	17.60	33.28	28.91	23.91
DC	ETM	895.9	897.5	894.3	1006.7	995.6	979.5
	ETM_SG	925.9	916.8	908.1	1043.1	1004.7	989.6
	NVDM	1427.8	1587.0	1641.6	1348.7	1512.4	1954.0

appropriate dimension for topic modeling. The above limitation notwithstanding, a topic modeling technique that is based on a general-purpose sequence embedding can help support various applications including information retrieval, knowledge discovery and text summarization.

7. Conclusion

An evolutionary topic modeling technique (pPSO) which is based on PSO is applied to two very distinct datasets. The first, consists of posts from an online health forum (r/Cancer) and the second consists of messages posted to 20 different news groups. This modeling technique is compared to the generative models: ETM and NVDM. To our knowledge, this comparative analysis has not been previously explored for PSO.

The results indicate that pPSO can develop interpretable topics, with a TC and TD comparable to ETM and NVDM, respectively. The results also show that a sequence embedding is more adequate for topic modeling than a word-level embedding.

Further research efforts are needed to develop performance measures that can compare and identify overlapping topics produced by different topic modeling techniques. A better understanding of the impact of the vocabulary size and the general-purpose nature of the vocabulary on topic modeling is also needed. Finally, document completion metrics that can directly compare generative and evolutionary models should be investigated.

Data and related documentation developed in this study are available at osf.io/pz83w/.

Role of funding source

Authors are responsible for the work described in this paper. Study sponsor had no involvement in the collection, analysis and interpretation of the data.

CRediT authorship contribution statement

Samuel Miles: Development of the design methodology, implementation and analysis of the results, reviewed and approve the current manuscript. **Lixia Yao:** Development of the design methodology, implementation and analysis of the results, reviewed and approve the current manuscript. **Weilin Meng:** Development of the design methodology, implementation and analysis of the results, reviewed and approve the current manuscript. **Christopher M. Black:** Development of the design methodology, implementation and analysis of the results, reviewed and approve the current manuscript. **Zina Ben Miled:** Development of the design methodology, implementation and analysis of the results, reviewed and approve the current manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported in part by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA. The authors would also like to thank Jarod Baker of the Regenstrief Institute for his support.

References

Agbehadji, I. E., Awuzie, B. O., Ngowi, A. B., & Millham, R. C. (2020). Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. *International Journal of Environmental Research and Public Health*, *17*(15), 5330. http://dx.doi.org/10.3390/ijerph17155330.

Aggarwal, C. C., Han, J., Wang, J., & Yu, P. S. (2004). A framework for projected clustering of high dimensional data streams. *Vol. 30*, In *Proceedings of the thirtieth international conference on very large data bases* (pp. 852–863). http://dx.doi.org/10.5555/1316689.1316763.

Armstrong, M. D., Maupomé, D., & Meurs, M.-J. (2021). Topic modeling in embedding spaces for depression assessment. In *Proceedings of the Canadian conference on artificial intelligence*. http://dx.doi.org/10.1145/3133944.3133945.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *Vol. 14*, In *Proceedings of the international AAAI conference on web and social media* (pp. 830–839). http://dx.doi.org/10.5281/zenodo.3608135.

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. (pp. 759–766). http://dx.doi.org/10.18653/v1/2021.acl-short.96.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. http://dx.doi.org/10.5555/944919.944937.

Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys*, *54*(7), 1–35. http://dx.doi.org/10.1145/3462478.

Chen, Y., Qin, B., Liu, T., Liu, Y., & Li, S. (2010). The comparison of SOM and K-means for text clustering. *Computing and Information Science*, *3*(2), 268–274. http://dx.doi.org/10.5539/cis.v3n2p268.

Churchill, R., & Singh, L. (2021). The evolution of topic modeling. *ACM Computing Surveys*.

Cui, X., Potok, T. E., & Palathingal, P. (2005). Document clustering using particle swarm optimization. In *Proceedings 2005 IEEE swarm intelligence symposium, 2005. SIS 2005* (pp. 185–191). IEEE, http://dx.doi.org/10.1109/SIS.2005.1501621.

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*, *57*(2), Article 102034. http://dx.doi.org/10.1016/j.ipm.2019.04.002.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: PRe-training of deep bidirectional transformers for language understanding. *1*, In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 4171–4186). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423.

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2019). The dynamic embedded topic model. ArXiv abs/1907.05545. arXiv:1907.05545.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453. http://dx.doi.org/10.1162/tacl_a_00325.

Freifeld, C. C., Brownstein, J. S., Menone, C. M., Bao, W., Filice, R., Kass-Hout, T., et al. (2014). Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug Safety*, *37*(5), 343–350. http://dx.doi.org/10.1007/s40264-014-0155-x.

George, C. P., Wang, D. Z., Wilson, J. N., Epstein, L. M., Garland, P., & Suh, A. (2012). A machine learning based topic exploration and categorization on surveys. *Vol. 2*, In *2012 11th international conference on machine learning and applications* (pp. 7–12). IEEE, http://dx.doi.org/10.1109/ICMLA.2012.132.

Gkotsis, G., Mueller, C., Dobson, R. J., Hubbard, T. J., & Dutta, R. (2020). Mining social media data to study the consequences of dementia diagnosis on caregivers and relatives. *Dementia and Geriatric Cognitive Disorders*, *49*(3), 295–302. http://dx.doi.org/10.1159/000509123.

Gupta, A., & Zhang, Z. (2021). Vector-quantization-based topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *12*(3), http://dx.doi.org/10.1145/3450946.

Haas, K., Ben Miled, Z., & Mahoui, M. (2019). Medication adherence prediction through online social forums: A case study of fibromyalgia. *JMIR Medical Informatics*, *7*(2), Article e12561. http://dx.doi.org/10.2196/12561.

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management*, *54*(6), 1292–1307. http://dx.doi.org/10.1016/j.ipm.2018.05.006.

Hailu, T. T., Yu, J., & Fantaye, T. G. (2020). A framework for word embedding based automatic text summarization and evaluation. *Information*, *11*(2), 78. http://dx.doi.org/10.3390/info11020078.

Karl, A., Wisnowski, J., & Rushing, W. H. (2015). A practical guide to text mining with topic extraction. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(5), 326–340. http://dx.doi.org/10.1002/wics.1361.

Karol, S., & Mangat, V. (2013). Evaluation of text document clustering approach based on particle swarm optimization. *Open Computer Science*, *3*(2), 69–90. http://dx.doi.org/10.2478/s13537-013-0104-2.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Vol. 4*, In *Proceedings of ICNN'95-international conference on neural networks* (pp. 1942–1948). IEEE, http://dx.doi.org/10.1109/ICNN.1995.488968.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995* (pp. 331–339). Elsevier.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics* (pp. 530–539). http://dx.doi.org/10.3115/v1/E14-1056.

Lee, K., Agrawal, A., & Choudhary, A. (2015). Mining social media streams to improve public health allergy surveillance. In *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015* (pp. 815–822). http://dx.doi.org/10.1145/2808797.2808896.

Liu, J., Xia, C., Li, X., Yan, H., & Liu, T. (2020). A BERT-based ensemble model for Chinese news topic prediction. In *Proceedings of the 2020 2nd international conference on big data engineering* (pp. 18–23). http://dx.doi.org/10.1145/3404512.3404524.

McInnes, L., Healy, J., Saul, N., & berger, L. G. (2018). UMAP: UNiform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861. http://dx.doi.org/10.21105/joss.00861.

Miao, Y., Yu, L., & Blunsom, P. (2016). Neural variational inference for text processing. In *International conference on machine learning* (pp. 1727–1736). PMLR.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miles, S., Yao, L., Meng, W., Black, C. M., & Ben-Miled, Z. (2021). Topic extraction from a cancer health forum. In *Proceedings of the 2021 IEEE 9th international conference on healthcare informatics* (pp. 491–492). http://dx.doi.org/10.1109/ICHI52183.2021.00085.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272). http://dx.doi.org/10.5555/2145432.2145462.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102. http://dx.doi.org/10.1177/0165551515617393.

Nzali, M. D. T., Bringay, S., Lavergne, C., Mollevi, C., & Opitz, T. (2017). What patients can tell us: topic analysis for social media on breast cancer. *JMIR Medical Informatics*, *5*(3), Article e7779. http://dx.doi.org/10.2196/medinform.7779.

Rajendran, D. P. D., & Sundarraj, R. P. (2021). Using topic models with browsing history in hybrid collaborative filtering recommender system: Experiments with user ratings. *International Journal of Information Management Data Insights*, *1*(2), Article 100027. http://dx.doi.org/10.1016/j.jjimei.2021.100027.

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. *Vol. 242*, In *Proceedings of the first instructional conference on machine learning* (pp. 29–48). http://dx.doi.org/10.5120/ijca2018917395.

Rashid, J., Shah, S. M. A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management*, *56*(6), Article 102060. http://dx.doi.org/10.1016/j.ipm.2019.102060.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing*. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1410.

Rodrigues, M. M., & Sacks, L. (2004). A scalable hierarchical fuzzy clustering algorithm for text mining. In *Proceedings of the 5th international conference on recent advances in soft computing* (pp. 269–274). Citeseer.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. http://dx.doi.org/10.1016/0306-4573(88)90021-0.

Shah, A. M., Yan, X., Tariq, S., & Ali, M. (2021). What patients like or dislike in physicians: Analyzing drivers of patient satisfaction and dissatisfaction using a digital topic modeling approach. *Information Processing & Management*, *58*(3), Article 102516. http://dx.doi.org/10.1016/j.ipm.2021.102516.

Tijare, P. V., & Prathuri, J. R. (2022). Correlation between K-means clustering and topic modeling methods on Twitter datasets. *Cyber Security and Digital Forensics*, 459–477. http://dx.doi.org/10.1007/978-981-16-3961-6_38.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., et al. (2019). CluWords: Exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 753–761). http://dx.doi.org/10.1145/3289600.3291032.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105–1112).

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52. http://dx.doi.org/10.1007/s13042-010-0001-0.

Zhang, J., Liu, M., & Zhang, Y. (2020). Topic-informed neural approach for biomedical event extraction. *Artificial Intelligence in Medicine*, *103*, Article 101783. http://dx.doi.org/10.1016/j.artmed.2019.101783.