COMPARING PSO-BASED CLUSTERING OVER CONTEXTUAL VECTOR EMBEDDINGS TO MODERN TOPIC MODELING

by

Samuel Miles

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Electrical and Computer Engineering

Indianapolis, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Zina Ben Miled, Chair

Department of Electrical and Computer Engineering

Dr. Paul Salama

Department of Electrical and Computer Engineering

Dr. Mohamed El-Sharkawy

Department of Electrical and Computer Engineering

Approved by:

Dr. Brian King

To my wife for all her patience.

ACKNOWLEDGMENTS

This research was supported in part by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA. I would like to thank Dr. Christopher M. Black, Mr. Weilin Meng and Dr. Lixia Yao from Merck & Co., Inc. and Dr. Zina Ben Miled, my thesis advisor, for their advice and guidance during this project. The support of Mr. Jarod Baker from the Regenstrief Institute is greatly appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

PSO	Particle Swarm Optimization
pPSO	Point-wise PSO
ETM	Embedded Topic Model
NVDM	Neural Variational Document Model
LDA	Latent Dirichlet Analysis
LSA	Latent Semantic Analysis
PWE	Pre-trained Word Embedding
BERT	Bidirectional Encoder Representations from Transformers
sBERT	sentence-BERT
SB	sBERT
SG	SkipGram
SVD	Singular Vector Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
DF-ICF	Document Frequency-Inverse Cluster Frequency
UMAP	Uniform Manifold Approximation and Projection
DC	Document Completion
PR	Parity
TC	Topic Coherence
TD	Topic Diversity

ABSTRACT

Efficient topic modeling is needed to support applications that aim at identifying main themes from a collection of documents. In this thesis, 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 three datasets across different domains. The first dataset consists of posts from the online health forum r/Cancer. The second dataset is a collection of NY Times abstracts and is used to compare the proposed model to LDA. The third is a standard benchmark dataset for topic modeling which consists of a collection of messages posted to 20 different news groups. It is used to compare state-of-the-art generative document models (i.e., ETM and NVDM) to pPSO. The results show that pPSO is able to produce interpretable clusters. Moreover, pPSO is able to capture both common topics as well as emergent topics. 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 20News-Groups 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 [1], personalizing users' experiences in recommender systems [2], to the analysis of the sentiments of patients towards their health care providers [3]. Extensive reviews of topic modeling applications and techniques are available in [4] and [5], however general overviews of modern techniques are provided for the purposes of a comparative background to the approach proposed in this thesis.

Most of the current research in topic modeling focuses on the use of probabilistic generative models [6, 7, 8, 9, 10]. 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 [11]. For instance, K-Means was shown to generate comparable clusters to those produced by LDA over a corpus of Twitter posts [12]. In [13], 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 thesis, an evolutionary clustering approach to topic modeling is also pursued.

An unstructured corpus of text documents is first processed using a text processing pipeline that applies common text standardization methods. A contextual representation of the documents in the form of vector embeddings is then generated using a general-purpose pre-trained language model. Particle Swarm Optimization (PSO) [14] is applied to the resulting embeddings to group the documents into semantically coherent clusters.

The proposed approach was motivated by the desire to gain a better understanding of cancer patients and caregivers experiences' outside of a clinical environment. Online health forums offer an alternate source that document patients' experiences. These forums are becoming ubiquitous and often chronicle the patient's journey. The /r/Cancer forum is a platform where users can post about their own experiences with cancer and seek support

from others in the community. This creates an opportunity to use concept identification methods to offer a better understanding of the patients experiences' outside of a clinical environment.

The approach was then extended to other domains in order to to demonstrate that it can be effectively applied across domains such as social media, news abstracts, and news groups. The proposed approach is also compared to current state-of-the-art generative topic models produced by ETM [11] and NVDM [15] using a collection of posts from the r/Cancer [16] health forum and the 20NewsGroups [17] dataset[1], a standard benchmark in topic modeling which was previously used to evaluate both ETM and NVDM.

Initially, the proposed approach to topic modeling pPSO is compared to LDA. In a second experiment the comparison is extended to ETM, ETM+PWE and NVDM. The r/Cancer social media dataset is used in the two experiments, the collection of NY Times article abstracts is used when comparing pPSO to LDA, and the 20NewsGroups dataset is used when comparing pPSO to ETM, ETM+PWE, and NVDM. The 20NewsGroups dataset is chosen for the second experiment because the NY Times article abstracts dataset consists of very short documents. Short text documents are known to return uninterpretable topics when considering key topic metrics specifically [11].

In summary, the primary goal of this thesis is to demonstrate that the proposed clustering-based topic modeling approach is effective at identifying latent topics across multiple domains. The secondary goal of this thesis is to demonstrate the competitive performance of the latest proposed approach in both a qualitative and quantitative evaluation compared to modern probabilistic topic models.

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

2. RELATED WORK

Topic extraction methodologies are continuously being improved and increasingly being used in various fields. For example, in market research, topic extraction is widely used to enhance our understanding of customer reviews [18] and for purposes such as brand name and product evaluation [19]. It is also used in the analysis of trends in news streams [9]. Recent research is extending topic extraction to health studies. In [10], it was used to extract clinically relevant information to the quality-of-life (QoL) of breast cancer patients and in [10] it was used to perform event extraction from a biomedical database.

The present thesis was also motivated by health studies. Specifically, a methodology was needed that is able to identify topics of interest to the r/Cancer online health community from the unstructured text of the subscribers' posts. There is strong evidence in the literature that social media data can help answer important health questions. For example, this type of data has been used for medication adherence prediction [20], the tracking of the spread of infectious diseases [21, 22, 23], the identification of common allergy types [24] and for the development of a better understanding of the consequences of a dementia diagnosis on the caregivers [25]. These previous studies suggest that online community health forums can improve our understanding of the patient experiences using data collected outside of a clinical environment.

The proposed approach was then extended to multiple domains and its performance was compared to modern state-of-the-art topic modeling techniques. A topic is defined as a "constellation of words that tend to come up in a discussion" [26]. Topic extraction consists of identifying these main constellations from a collection of documents and typically follows a two-step approach, especially when applied to a large collection of documents. This process is called topic modeling. 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.

In this chapter, some of the most commonly applied and modern approaches to the two aforementioned topic modeling steps are reviewed.

2.1 Text Processing

The goal of text processing 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 [27]. 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 [27] and topic modeling [28, 11] applications.

These text processing pipelines are highly customizable and flexible across different text domains. In some use-cases removal of dictionary specific keywords is required. For example, certain high frequency medical terms (e.g., patient, hospital) may need to be removed from a corpus of medical documents.

Additional techniques for text processing include the removal of terms based on part-of-speech (PoS), and using more advanced language standardization like lemmatisation. When considering removal by PoS tag, each term in each document of the corpus needs to be tagged and then filtered according to the text processing retention rules. Under lemmatisation, terms are reduced to a stanardized format where terms are singlularized and conjugation is removed. This is effectively maps different forms of a particular term to a single word.

2.2 Text Representation

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 [29]. When applied to a rich corpus, this encoding can lead to a very sparse matrix. These techniques include using singular value decomposition (SVD) or term-frequency inverse document-frequency (TF-IDF) [30] which emphasizes infrequent words in the documents. These techniques may still result in vector representation with a high dimension. Moreover, the underlying vector representation often ignores the context and the order in which the words appear in the

document [26]. Therefore, this text representation may suffer from multiple limitations including the ambiguity of polysemic words [31].

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 [32]. The use of these embeddings varies considerably. Some of the most commonly used modern approaches to generating contextual vector embeddings are Bidirectional Encoder Representations from Transformer models, or BERT-based models [32]. These language models are pre-trained to represent text contextually. Having large amounts of data to train these models can lead to a better representation of the text. This can be observed in the extension of BERT, RoBERTa [33] where a large amount of language data from Facebook is taken to train a BERT model. Also, for longer documents, a modified version known as Longformer was introduced. It uses a sliding window for creating the contextual vector embedding representation of long documents [34].

The embedding technique that is adopted in this thesis is sentence-BERT (sBERT) [35]. 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. Given that sBERT is a BERT-based model, it into consideration the context of the word within the sequence. Therefore, a word may have a different embedding depending on the context as expressed by the surrounding words for each instance. The entire document is processed with sBERT. This generates a context aware vector embedding for each document. sBERT is a sequence-level embedding. It uses the BERT tokenizer to split words in the document. Thus a word may be associated with multiple tokens and each token has a an embedding. For example, the word "complaining" will result in the tokens "complain" and "ing" each with a vector embedding. An element-wide average of all the vector embeddings of all the tokens is computed and used as the final representation of each document.

In general, 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 [11] 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. 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. This variant of ETM is labeled ETM_SG in the remainder of the present thesis. The predecessor of ETM, LDA use a BoW matrix representation of the input text.

2.3 Dimension Reduction

Reducing the number of features present in the embedded representation of the data is common practice since higher dimension often require higher number of sample data for model development. Moreover, a feature space with reduced dimension may entail less computation. Dimension reduction techniques include T-Distributed Stochastic Neighbor Embedding (T-SNE) [36] and Uniform Manifold Approximation and Projection (UMAP) [37]. In short, both approaches to dimension reduction take a higher dimensional graph and reconstruct it in a lower dimenstionality while retaining the overall structure of the vector. T-SNE does this through point-by-point mapping to the lower dimension, and UMAP compresses the overall input graph to generate the reduced embeddings.

Dimension reduction is important for topic extraction since the aim is to cluster groups of coherent documents. Clustering using high dimension vector embeddings was shown to have lower performance compared to using lower dimension vector embeddings [38]. In this thesis, UMAP is used [37]. This method 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 done on a range of dimensional values from the full 768 dimension embeddings generated directly from sBERT [35] down to 30 dimensions. Each reduced embedding is generated from the original size embedding provided by SBERT [35].

2.4 Topic Modeling

A topic is traditionally defined, by previous work in the field, 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 [28]. 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.

2.4.1 Generative Models

One of the earliest topic modeling technique is Latent Semantic Analysis/Indexing (LSA/LSI) [39]. LSA is based on the frequency of co-occurrence of words in a document. LSA was superseded by Latent Dirichlet Allocation (LDA), a topic modeling technique that views each text as a collection of topics [28]. Most current topic extraction applications are based on LDA [26, 18]. For example, LDA was used for topic modeling of scientific abstracts [27], US patent data [40] and online forums [41, 31].

Despite the fact that LDA improves on LSA in dealing with a larger corpus, it still suffers from the difficulty of application-specific hyper-parameter fine-tuning [41] and prohibitive time complexity when applied to a corpus with a rich vocabulary [11]. The limited ability of LDA to handle large vocabularies can be partially attributed to the high dimensionality of the BoW vector representation [27]. Up until recently, this limitation was addressed by using the dimension reduction techniques discussed earlier. An extensive survey on the studies involving LDA, their applications, and the tools available presently to implement LDA is described in [4, 42]. The authors in [4] compare numerous LDA variants, dating as far back as 2004, up to modern studies from as recently as 2019. In another example, a model that redefines the conditional distribution given by LDA is proposed in [43]. This model (ProdLDA) is shown to return consistently better topics than LDA, even in the case where LDA is trained using Gibbs sampling.

The embedded topic model (ETM) [11] builds on the combination of LDA and the variational auto-encoder (VAE). The resulting word embeddings are similar to the CBOW word embeddings [23]. Except in the case of ETM, the context vector expected in CBOW 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 thesis is NVDM [15]. 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.

Authors in [44] compare a wide range of modern topic models, including: ProdLDA, MetaLDA [45], NVDM, ETM_SG, and LDA. Each of the models is evaluated on 5 different datasets from various sources ranging from news articles, to Wikipedia pages, to archived Twitter data from 2011. The findings vary based on the dataset tested, with a modified version of ProdLDA developed by the authors performing best on average [44]. More relevant to the proposed thesis is the clear disadvantages in using topic coherence as a metric to evaluate ETM_SG, LDA, and NVDM when tested on short-text documents. In the case of the Stack Overflow forum post dataset, all but two of the tested topic models returned a negative topic coherence. This is expected since this dataset consists of short-text documents [11]. The authors did not provide a subjective analysis and comparison of topic words for each identified topic. Therefore, it is not possible to evaluate the effectiveness of the topic coherence metric in representing the quality of the topics.

2.4.2 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 results from evolutionary and generative topic models may be comparable [12, 31, 46].

Examples of evolutionary algorithms include K-Means, fuzzy clustering [47], and PSO [48]. As in the case of generative models, traditional evolutionary models for topic modeling relied on the BoW representation of the text [49, 50, 51]. Generally evolutionary models follow the process of generating a series of centroids within the search space, the centroids move throughout the search space based on minimizing the distance to a selection of sample data points using a distance measurement such as Cosine similarity or Euclidean distance. More recent studies investigate the use of vector embeddings for text representation with evolutionary topic models.

In [31], K-Means over doc2vec embeddings were found to outperform LDA with BoW. The evaluation performed in [31] includes multiple clustering methods and various feature representation techniques with the overall goal being to compare the results of each of the clustering and feature representation combinations to that of LDA with a BoW encoding. The best results were achieved using either hierarchical or k-means clustering, with word2vec and doc2vec being the best representations. It was found that any combination of the two best clustering and feature representation methods outperformed LDA with a BoW encoding. The evaluation was performed using social media datasets from Twitter and Reddit.

Despite the above described comprehensive comparison of different clustering methods to LDA [31], a comparison of evolutionary topic modeling techniques to more recent generative topic modeling methodologies such as ETM and NVDM was not explored.

3. METHODS

The primary objective of this thesis is to compare the performance of a topic modeling approach based on a modified PSO technique to the widely used generative models LDA, ETM and NVDM. Specifically, this thesis compares the proposed PSO approach to LDA, ETM, and NVDM under different text representations in the context of topic extraction from 1) a subject-specific collection of social media documents (cancer), 2) a collection of short-text documents from as NY Times abstracts, and 3) a diversified set of text from 20 news groups. The social media dataset is used in all of the experiments performed, whereas the short-text documents of the NY Times abstracts are only considered for the comparison with LDA, and the diversified set of text from 20 news groups is considered for the comparison with 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.

3.1 Datasets

The first dataset used in this thesis consists of r/Cancer posts from the Reddit archive [16], and the second dataset consists of documents submitted to 20 news groups [17]. 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 [20], the tracking of the spread of infectious diseases [21, 22], the identification of common allergy types [24] and the development of a better understanding of the consequences of a dementia diagnosis on the caregivers [25].

The posts in this dataset were extracted from the r/Cancer Reddit archive [16]. 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 second dataset considered is a collection of short-text documents from the NY Times. Specifically, it is a collection of article abstracts from the NY Times from 2020. This dataset was selected because it offers a more formal text corpus as opposed to social media text, and allows for analysis of the effects of a short-text format on the results of topic modeling for pPSO and LDA.

The 20NewsGroups dataset was also selected as a third dataset because it was also used in the evaluation of ETM and NVDM. The version of this dataset that is used in this thesis 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. This dataset is considered for the comparison of pPSO to ETM and NVDM.

3.2 Data Representation

Since topic modeling can be sensitive to pre-processing, a streamlined approach was adopted for all datasets and across all models. First, instances of punctuation and number symbols are removed. Next, the text is forced to lowercase. Finally, stop words as identified in [52] and terms with a document frequency greater than 30% are also removed. Additionally, in the first experiment where pPSO is compared to LDA, processing based on part-of-speech (PoS) is also used. Specifically, terms identified to be nouns, verbs, adjectives, adverbs, proper nouns, and interjections are retained and all other terms are removed. This additional processing is relevant to the first experiment, but it is not used in the second experiment when pPSO is compared to ETM, ETM_SG, and NVDM.

3.2.1 Vector Embedding

After preprocessing, the documents are transformed into an embedding representation using sBERT [35]. 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 before clustering. Topic cluster-

ing using high-dimension vector embeddings was previously shown to have lower performance compared to lower dimension embeddings [38].

For comparison purposes, pPSO 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 LDA, ETM, and NVDM the embeddings are corpus specific and limited to the most frequent words. When considering LDA, a BoW embedding is used to represent each document in the corpus and the vocabulary consists of all terms in the corpus after text pre-processing has occurred. A vocabulary of the top 2,000 words is considered in ETM, ETM_SG, NVDM, and the Skip-Gram version of PSO. This limited vocabulary is identical for each of the models being investigated in this thesis. An unlimited size vocabulary was considered, however the significant increase in computational complexity made this option impractical. For the proposed pPSO approach this limitation is not applicable since the methodology uses a pre-trained language model.

3.2.2 Dimension Reduction

In this thesis, the Uniform Manifold Approximation and Projection (UMAP) technique is used for dimension reduction [37]. UMAP reduces the number of features in an embedded vector based on the values of the surrounding terms given a specified global structure value. Evaluation of several dimension sizes ranging from the 768 directly generated by sBERT down to 30 was also performed. Each reduced embedding is generated from the original size embedding provided by SBERT [35]. Varying values of the global structure were also investigated.

3.3 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, which represents the centroid of a cluster. Thus, the total number of particles/clusters is fixed to K. The position of the particle, $x_{\mathrm{i}}(t)$, is initialized to the embedding of the randomly selected document from the corpus (i.e., $x_{\mathrm{i}}(0) = \mathrm{e_i}$). 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 $\mathrm{e_j}$ in the dataset is assigned to one of the particles/clusters according to Equation 3.1. A cluster C_{i} consists of the set of documents $\mathrm{e_j}$ that are assigned to it as defined in Equation 3.2. The fitness of particle i is then computed according to Equation 3.3 which compares the current position of the particle $x_{\mathrm{i}}(t)$ to the embedded representation of each document assigned to C_{i}.

$$clust(\mathrm{e_j}) = argmin_{1 \leq \mathrm{i} \leq K}\{d(\mathrm{e_j}, x_{\mathrm{i}}(t))\} \tag{3.1}$$

$$C_{\mathrm{i}}(t) = \{clust(\mathrm{e_j})_{1 \leq \mathrm{j} \leq m} = \mathrm{i}\} \tag{3.2}$$

$$f_{\mathrm{i}}(t) = \frac{\sum_{\mathrm{j} \in C_{\mathrm{i}}(t)} d(\mathrm{e_j}, x_{\mathrm{i}}(t))}{\|C_{\mathrm{i}}(t)\|} \tag{3.3}$$

The distance d in Equation 3.1 and Equation 3.3 can take on several forms including Euclidean distance and cosine similarity. In the present thesis, cosine similarity is used as a distance measure. During each iteration t, the next velocity $v_{\mathrm{i}}(t+1)$ and next position $x_{\mathrm{i}}(t+1)$ for each particle i are computed using Equation 3.4 and Equation 3.5, respectively. In Equation 3.4, 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 Equation 3.4 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.

$$v_{\mathrm{i}}(t+1) = \omega v_{\mathrm{i}}(t) + \mu(L_{\mathrm{i}} - x_{\mathrm{i}}) + \rho(G - x_{\mathrm{i}}) \tag{3.4}$$

$$x_{\mathrm{i}}(t+1) = x_{\mathrm{i}}(t) + v_{\mathrm{i}}(t+1) \tag{3.5}$$

Algorithm 1: Point-wise Particle Swarm (pPSO)

```
Data: number of clusters n, embedded posts
Result: list of extracted topic clusters
def pPSO(n, embedded posts, cluster list):
    generate n centroid particles, one for each cluster;
    while fitness threshold is not met do
        for each post e_i do
            for each particle with position x_j do
                calculate the distance from e_i to x_j and
                  record the best global position of the
                  swarm on a dimension-by-dimension
                  basis
            end
            calculate fitness as the average distance
              from e_i to all posts
        end
    end
    for each post e_i do
        assign e_i to the cluster whose centroid is the
          closest by distance;
    end
    return cluster list
```

Figure 3.1. Point-wise Particle Swarm Optimization (pPSO) Algorithm.

The present thesis 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) as described in Figure 3.1.

A variant of the pPSO algorithm was also investigated. This variant consists of: a hierarchical pPSO (hPSO) that combines multiple layers of pPSO to identify sub-topics.

Algorithm 2: Hierarchical Particle Swarm (hPSO)

```
Data: number of clusters n, embedded posts, global
      cluster list
Result: list of extracted clusters
def hPSO(n, embedded posts, cluster list):
    generate n clusters using pPSO;
    for each cluster C_i do
        if cluster size ≥ 15% of the corpus then
            add parent cluster C_i to cluster list;
            expand with hPSO for 1 generation with
             just two child clusters;
            add generated child clusters to cluster list;
        else
            add cluster C_i to extracted cluster list;
        end
    end
```

Figure 3.2. Hierarchical Particle Swarm Optimization (hPSO) Algorithm.

Under hPSO, documents are assigned to each of the clusters in the exact same way as pPSO on the first iteration. In subsequent iterations, clusters that were generated in the first generation are evaluated as the input corpus for a second generation of pPSO based on their size. Clusters containing at least 15% of the documents in the corpus are expanded upon hierarchically. The second generation of pPSO is limited to two clusters because it is assumed that the first iteration of pPSO produces well balanced topic clusters from the corpus. Moreover, limiting the second generation to 2 clusters allows us to visualize the

differences in clustered documents for more granular topics. This algorithm is shown in Figure 3.2.

3.4 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 [53] and is an open research problem [54]. Moreover, the extraction of topic words is different in generative models 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 [28, 11, 15].

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 structurally similar to term-frequency inverse document frequency (TF-IDF) [55, 56]. 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 at least one instance of w and ICF is then given by the following equation.

$$ICF(w, C) = log \frac{N}{\mid \{D \in C : w \in D\} \mid} \tag{3.6}$$

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

Evaluation metrics for topic models include coherence and diversity [11]. Topic coherence (TC) is a measure of how representative a collection of topic words is for a given set

of documents [57, 11]. TC is defined as the normalized summation of point-wise mutual information (PWI) [58] 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 thesis, the top 10 words are selected to represent a cluster and the corresponding TC expression is shown in Equation 3.7. This exact metric was used to evaluate ETM [11].

$$TC = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{45}\sum_{\mathrm{i}=1}^{10}\sum_{\mathrm{j}=\mathrm{i}+1}^{10} PWI({w_\mathrm{i}}^{(k)}, {w_\mathrm{j}}^{(k)}) \tag{3.7}$$

$$PWI(w_\mathrm{i}, w_\mathrm{j}) = \frac{log\frac{P(w_\mathrm{i},w_\mathrm{j})}{P(w_\mathrm{i})P(w_\mathrm{j})}}{-logP(w_\mathrm{i}, w_\mathrm{j})} \tag{3.8}$$

where $\{{w_1}^{(k)}, ..., {w_{10}}^{(k)}\}$ are the top-10 topic words for a given cluster C and K is the number of clusters. $P(w_\mathrm{i}, w_\mathrm{j})$ is the number of documents in a given cluster in which the pair of words w_i and w_j co-occur, whereas $P(w_\mathrm{i})$ and $P(w_\mathrm{j})$ are the number of documents within a given cluster in which the respective words appear at least once. The probabilities in equations 3.7 and 3.8 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 [11]. 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) [59]. 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, an exact 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.

$$PR = \frac{1}{n_{test}} \sum_{\mathrm{i}=1}^{n_{test}} 1 \qquad \text{if} \qquad clust(h_{1\mathrm{i}}) = clust(h_{2\mathrm{i}}) \tag{3.9}$$

where n_{test} is the number of documents in the test dataset.

The PR metric is introduced as an alternative to DC because evolutionary models assign documents to a single cluster. Moreover, during model development, the representation of the corpus is optimized such that each document belongs to a single topic. A probabilistic measure that would associate a given document with each topic needs to be developed in order to use the DC metric. This extension to the proposed topic modeling approach is subject to future work.

4. RESULTS

This chapter includes the evaluation of the proposed approach under different context. First, quantitative evaluation of the topic models of hPSO is performed. Second, pPSO is compared to LDA using the r/cancer and the NYtimes datasets. This comparison allows for the evaluation of the proposed topic modeling approach compared to a traditional generative model. The third evaluation compares the proposed approach to novel generative topic models over two dataset r/Cancer and 20NewsGroups.

Table 4.1. Example documents after preprocessing from the r/Cancer, 20NewsGroups, and NY Times Abstracts 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
NY Times	*perfect dolphins footballs ultimate toll nfl celebrating team go entire season without losing glory came cost several teams prominent players found cte*
	best valentines day recipes lobster steak chocolate skip wait crowds stay home cook people love

Table 4.1 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 pages, even after preprocessing.

UMAP is used to reduce the dimensionality of the vector embeddings to varying levels ranging from 300 to 30 dimensions. It was found that the optimal value for the global structure capture weight is 35. This value provided the greatest balance between global information capture and local information retention. It was empirically established that a reduced embedding with a dimension of 300 had the best performance for the proposed PSO model, as well as the LDA [28], ETM [11], and NVDM [15] Models. Due to these findings, a dimensionality of 300 is chosen for the purposes of generating all embeddings used in this thesis. Additionally, for visualization purposes only, a reduction of the vector embeddings to 3 dimensions is performed and then plotted with their respective clusters labeled to allow for quick visual evaluation of clusters.

4.1 Hierarchical pPSO (hPSO)

The primary goal of topic modeling is the identification of qualitatively clear topics. To illustrate this process topic modeling is performed on the r/Cancer dataset to generate 10 topics. The collection of topic words and topic clusters that are selected by the proposed approach are examined. Each cluster contains a variable number of posts from the 2020 r/Cancer dataset.

Table 4.2. Top-5 Topic Words for each Document Cluster

1	2	3	4	5	6	7	8	9	10
constipated	participants	goodbye	shave	atypia	tapping	iv	foods	ct	hes
cannabis	recruiting	dating	wig	benign	hardships	deteriorating	meals	iv	prostate
gin	massachusetts	angel	wigs	leg	payout	refused	oatmeal	guess	dads
eoe	helen	funeral	clumps	muscle	organisations	coma	sores	chemotherapy	fathers
edibles	coordinator	hug	shaving	pea	jejunostomy	mums	fruits	excision	grandfather
49 docs	627 docs	954 docs	416 docs	848 docs	34 docs	660 docs	475 docs	888 docs	1,500 docs

The top 5 topic words identified by DF-ICF for each cluster are shown in Table 4.2. Many of the clusters have interpretable topics based on the subjective similarity of the chosen topic

words. The size of each cluster is also shown to help give an understanding of the size of the domain from which these topic words have been extracted.

If a cluster contains at least 15% of all documents in the corpus, pPSO will expand into a second generation to generate two child clusters, each with a different sub-topic. An example of this hierarchical process is depicted in Table 4.3 where cluster 9, which has 888 documents in total, is expanded and new topic words are identified for each sub-topic.

Table 4.3. Top-5 topic words and cluster sizes for each sub-topic of cluster 9.

9-1	9-2
guess	recommended
ct	originating
margins	cohort
neutrophils	imrt
ports	ligament
516 docs	372 docs

It can be seen from the two sub-topics from cluster 9 in Table 4.3 that most of the topic words chosen are different from that of the parent cluster. This is because we account for topic words from the parent cluster so as to minimize the overlap between sub-topic words and the original topic words. The difference in terms chosen for each sub-cluster shows a clear separation, however the top terms for each sub-cluster do not appear to yield an obviously, human interpretable topic despite some thematic similarities in some of the chosen terms.

Clusters 4, 8, and 10 show the most clearly interpretable topics from the first level clusters where all other clusters show terms with some similarity that are biased towards a particular theme in many cases.

To better visualize the relationship between topic words in a given cluster, a graph representation of some sample clusters are shown. The first example is cluster 4 and can be observed in Figure 4.1. The graphical representation of the clusters has a few key elements: 1) the larger the node the greater the DF-ICF value is for that term, 2) the edges of the graph represent documents where the terms both occur together, and 3) the thickness of the edge represents the number of documents in the corpus in which those terms co-occur. Figure 4.2 shows the document overlap relationship between the topic words chosen for cluster 5 and visualizes the highest impact terms.

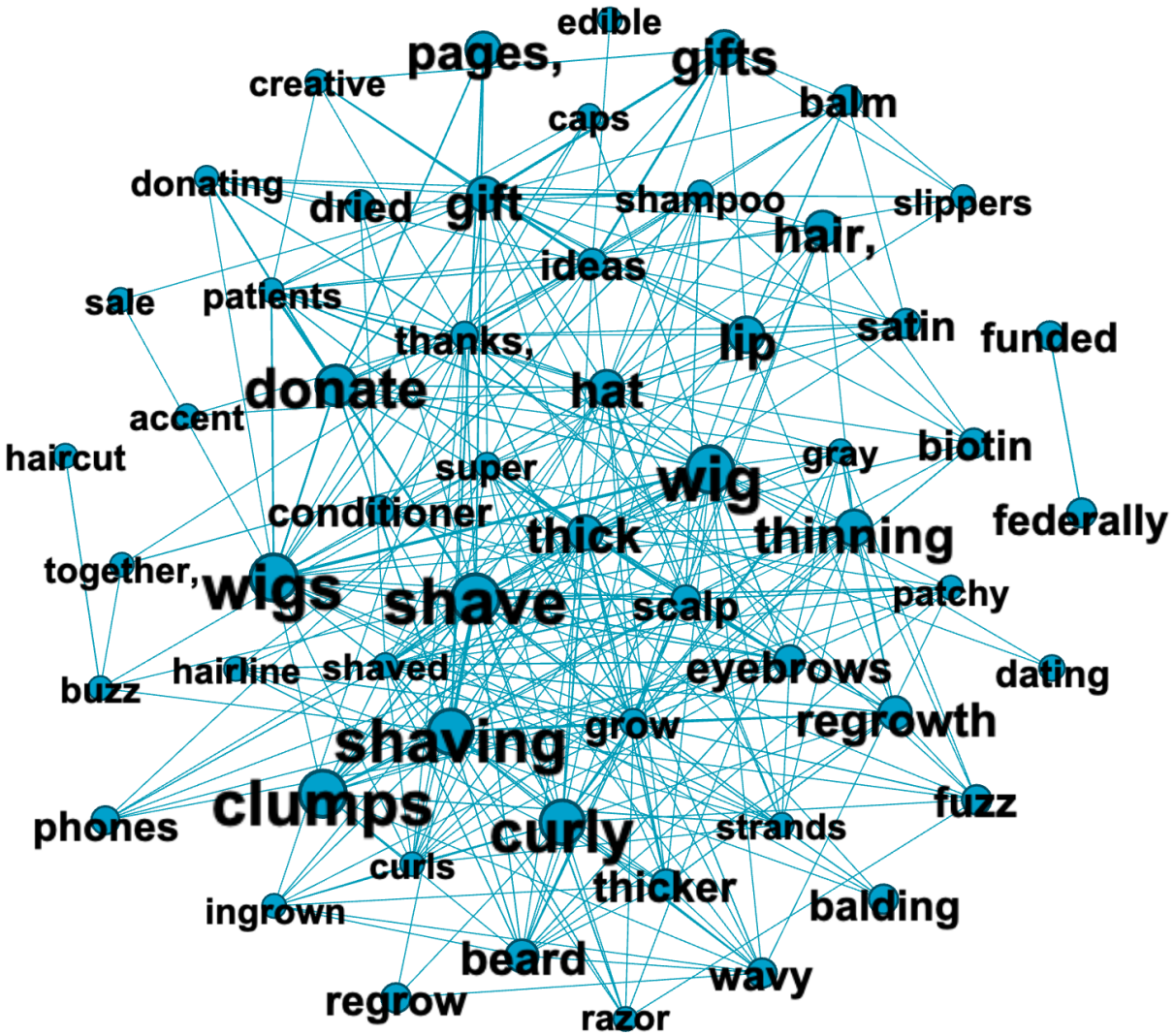

Figure 4.1. Graph representation of cluster 4 from Table 4.2.

The distribution of each generated cluster can be seen in Figure 4.3 with each color representing each respective cluster from left-to-right in Table 4.2. The results of clustering when projecting from 300 embedded dimensions down to 3 shown in Figure 4.3 indicates that PSO is functioning as expected to group similarly embedded submissions together using Euclidean distance calculated on each individual dimension. It can be clearly observed that each cluster defined by PSO is a collection of the most closely related submissions based on the their representations within the embedded space after dimension reduction. The

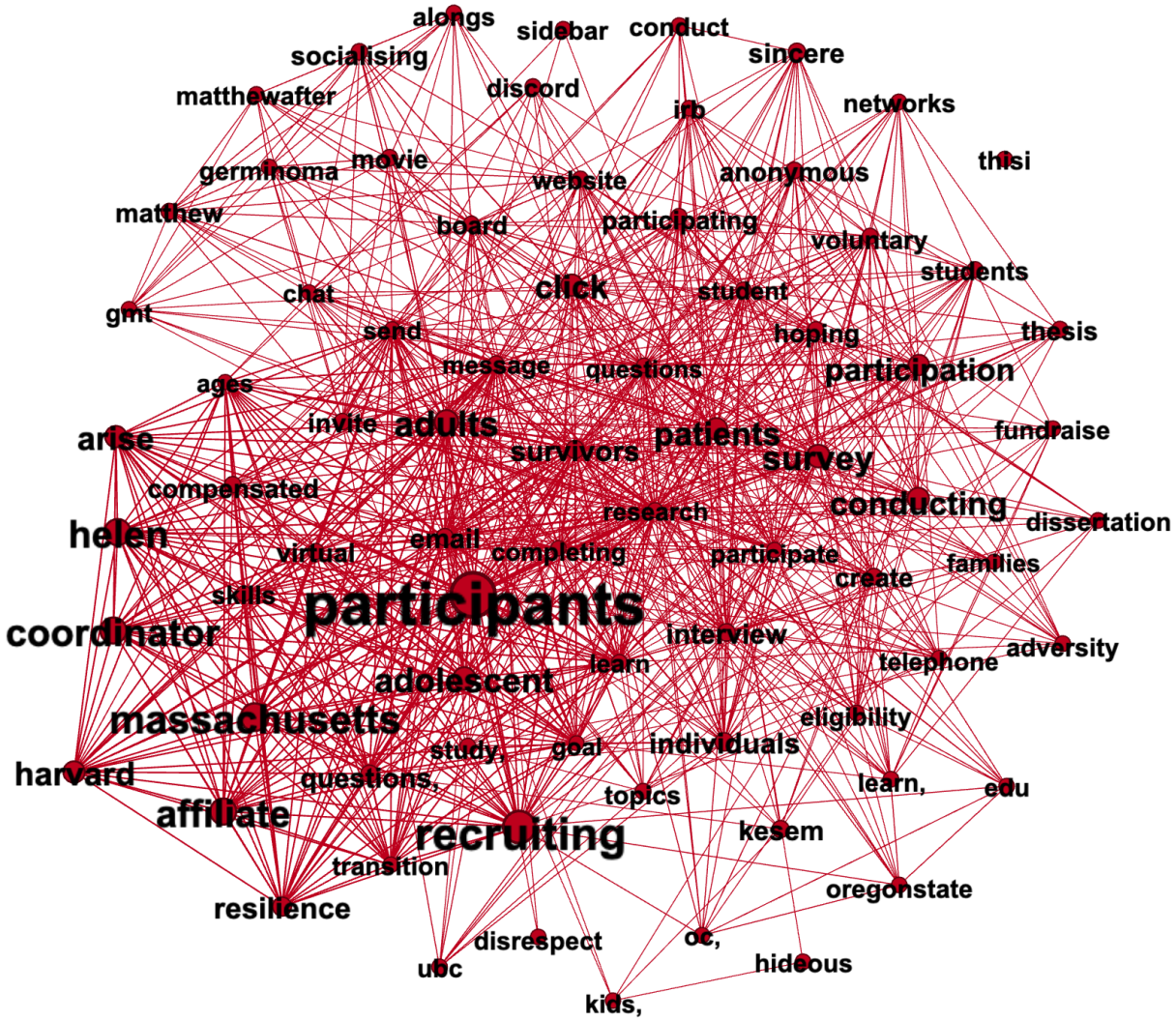

Figure 4.2. Graph representation of cluster 2 from Table 4.2.

clusters visualized in Figure 4.3 show how PSO functioned as desired, grouping documents with similar embedded representations together.

4.2 Comparing pPSO to LDA

The results discussed in this section demonstrate a subjective, qualitative analysis of the chosen topic words for each identified topic from two datasets for both pPSO and LDA.

Tables 4.4 and 4.5 show the top five topic words for each topic in the r/Cancer and NY Times datasets, respectively. In addition to pPSO, these tables include the results obtained

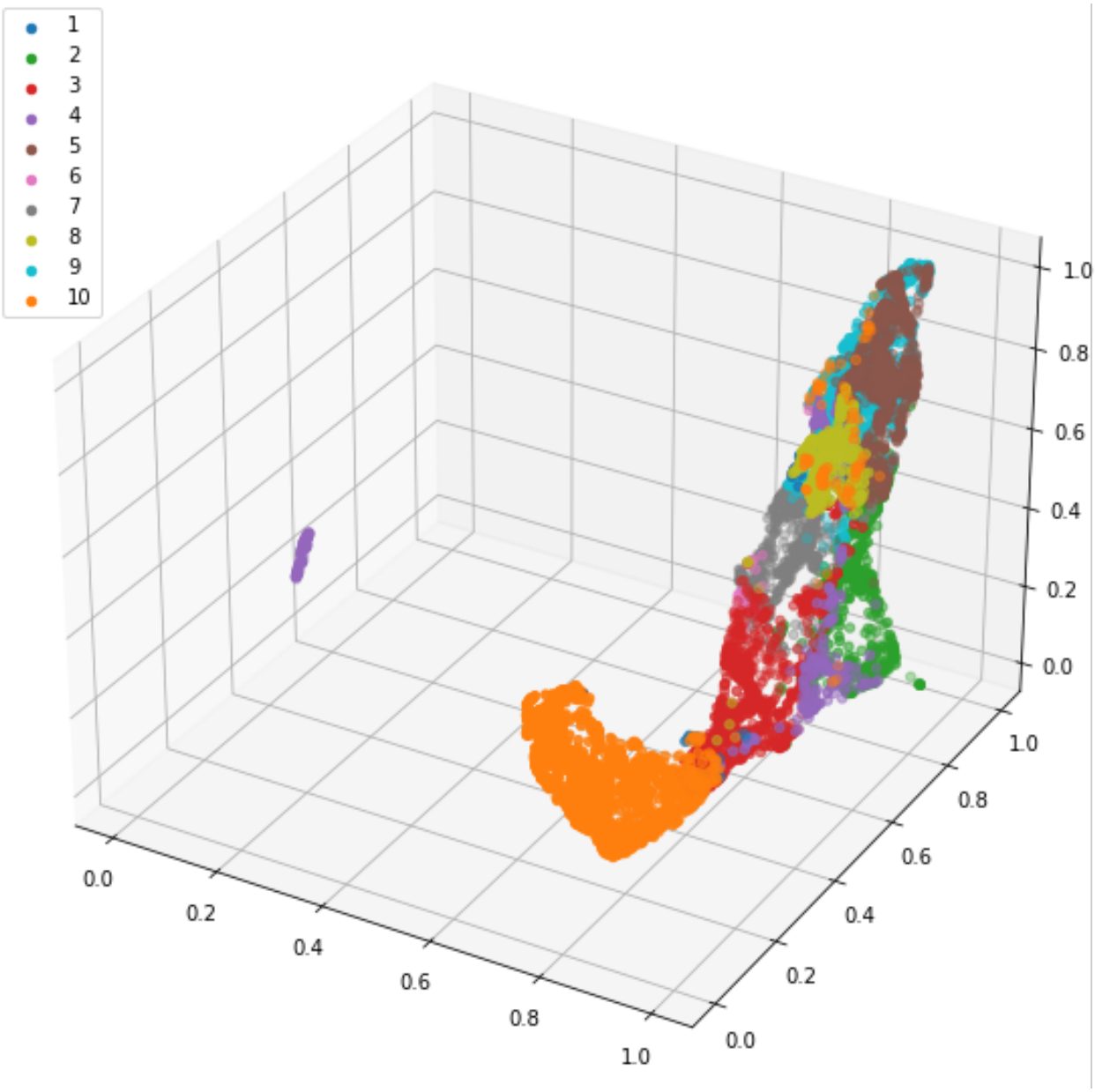

Figure 4.3. Three-dimensional visualization of the clusters in Table 4.2.

by using LDA. For the r/Cancer dataset, the number of topics was set to 10 across both of the evaluated models. In the case of the NY Times dataset, the number of topics was set to 15 since this dataset contains a large number of documents. Moreover, the implementation of LDA is the traditional generative model over words whereas pPSO uses a pre-trained vector embedding of the text as previously described.

Table 4.4. Top 5 topic words for each topic identified in the 2020 r/Cancer Reddit dataset.

1	2	3	4	5	6	7	8	9	10
					pPSO				
dads	friend	appetite	lynch	hair	uneasiness	insecticides	friend	recruiting	results
hes	goodbye	foods	ovarian	wigs	discomfort	canister	son	participants	looking
prostate	anymore	diet	situation	wig	siliar	appointmentthank	dads	massachusetts	tumor
looking	things	taste	screeching	shaving	doctorafter	spilled	man	coordinator	nodule
man	passed	meat	overheard	shave	uncomfortluckily	connectionshe	hes	affiliate	remove
					LDA				
get	chemo	chemo	chemo	like	dont	dont	dont	chemo	like
feel	like	get	dont	mom	like	chemo	like	want	dont
like	feel	would	like	te	told	would	te	te	want
treatment	really	like	get	feel	one	te	get	diagnosed	would
dont	dont	stage	dad	diagnosed	feel	get	feel	go	help

A review of the results in Tables 4.4 and 4.5 shows that many of the topics identified by pPSO are coherent for both datasets. LDA shows less specific topics especially for the r/Cancer dataset. For instance, topics (1) and (3) from pPSO for the r/Cancer dataset are centered around prostate cancer and food, respectively. Similarly, 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. The topics produced by LDA for the r/Cancer dataset are less distinct and are predominantly related to chemotherapy treatment.

A similar observation can be made for the NY Times dataset. The topics produced by pPSO are distinct such as football (1), renters (11) and education (14). There are also topics related to current events (e.g., antitrust and impeachment). Topic 9 in Table 4.5 for pPSO is interesting because it identifies a set of puzzles that are published in the NY Times. This is shown not only through the topic word choice of "puzzle", but also in the choice of "cox" and "rathvon", referring to the pair of well known crossword puzzle writers: Emily Cox and Henry Rathvon. The topics generated by LDA show less distinction and mostly focus on the impact of the pandemic as it relates to election, government and education.

Table 4.5. Top 5 topic words for each topic identified in the 2020 NY Times dataset.

1	2	3	4	5	6	7	8
				pPSO			
steelers	german	weinstein	recipes	italy	unemployment	invitation	antitrust
nfl	germany	harvey	meal	wuhan	economic	caption	facebook
colts	merkel	weinsteins	wines	infections	hurricane	ongoing	google
brady	infiltration	breonna	wine	outbreak	africa	runnersup	hackers
quarterback	farright	minneapolis	soup	animals	senegal	prompts	apple
9	10	11	12	13	14	15	
cox	vaccine	landlords	impeachment	donald	children	suleimani	
musical	vaccines	renters	bernie	republicans	classrooms	civilians	
puzzle	economic	renters	senate	republicans	childrens	progovernment	
cottage	beijing	tenants	sanders	trumps	exams	qassim	
rathvon	chinas	uber	ballots	administration	grades	afghan	
				LDA			
new	us	new	new	new	new	new	trump
coronavirus	coronavirus	coronavirus	coronavirus	coronavirus	york	coronavirus	president
many	new	trump	pandemic	trump	trump	one	new
one	trump	first	students	us	coronavirus	trump	coronavirus
city	could	pandemic	trump	president	pandemic	york	people
new	new	coronavirus	new	coronavirus	trump	new	
coronavirus	us	new	trump	new	president	trump	
pandemic	one	trump	coronavirus	us	coronavirus	people	
trump	coronavirus	pandemic	president	pandemic	new	pandemic	
york	house	could	pandemic	one	biden	many	

In summary, the qualitative evaluation of the topics produced by pPSO and LDA indicates that the pPSO topics are more cohesive than LDA, as it is difficult to subjectively identity clearly distinct topics from LDA. These observations are applicable to both the online health forum posts as well as the news abstracts from the NY Times.

4.3 Comparing pPSO to Modern Generative Models

The comparison of pPSO to modern generative 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.

Table 4.6 shows the distribution of each dataset in terms of number of documents, unique words, and document length.

Tables 4.7 and 4.8 show the top five topic words for each topic in the r/Cancer and 20NewsGroups datasets, respectively. A review of the results in Table 4.7 and Table 4.8

Table 4.6. Number of documents, unique words, words per document and characters per document for both the r/Cancer and 20NewsGroups 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

Table 4.7. 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 generated is set to 10 topics.

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

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 4.8. 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 4.1 were assigned to topic (1). The focus of topic (5) is on hair loss. Topic (9) represents an emergent topic consisting of a large number of posts recruiting candidates for clinical trials. With the same dataset, ETM_SG shows 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 4.1 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.

Table 4.9. 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

The topic coherence (TC) and topic diversity (TD) for the the two datasets are shown in Tables 4.9 and 4.10, 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 4.7 and 4.8.

Table 4.10. 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

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 4.11. 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

When tested on a document completion task, NVDM has higher DC values than either variant of ETM for both datasets (Table 4.11). The metrics DC and PR cannot be directly compared. However, Table 4.11 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.

5. DISCUSSION

The present thesis 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.

Topic modeling using hPSO shows that an evolutionary approach can generate clearly distinct topics. It is common for the larger clusters in the first generation to contain too many documents preventing the definition of a single topic that is common to all documents in the cluster. Using a hierarchical approach can improve the quality of the clusters, however additional improvement is necessary.

The topic models produced by pPSO are compared to three generative topic models: LDA, 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 [6, 7, 11, 44]. For evolutionary topic models, this aspect was only confirmed for K-Means [31]. This thesis extends this finding to PSO. Other evolutionary algorithms, such as fuzzy clustering, with contextual embedded vector representation remains to be explored.

Second, this 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 [11].

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 [31]. Similarly, fuzzy clustering was shown to have better performance than LDA in [46] 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 thesis 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%.

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 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.

6. CONCLUSIONS

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: LDA, 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 compared to all three models, 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.

REFERENCES

[1] L. Hagen, "Content analysis of e-petitions with topic modeling: How to train and evaluate lda models?" *Information Processing & Management*, vol. 54, no. 6, pp. 1292–1307, 2018.

[2] D. P. D. Rajendran and R. P. Sundarraj, "Using topic models with browsing history in hybrid collaborative filtering recommender system: Experiments with user ratings," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100027, 2021.

[3] A. M. Shah, X. Yan, S. Tariq, and M. Ali, "What patients like or dislike in physicians: Analyzing drivers of patient satisfaction and dissatisfaction using a digital topic modeling approach," *Information Processing & Management*, vol. 58, no. 3, p. 102516, 2021.

[4] U. Chauhan and A. Shah, "Topic modeling using latent dirichlet allocation: A survey," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–35, 2021.

[5] R. Churchill and L. Singh, "The evolution of topic modeling," *ACM Computing Surveys (CSUR)*, 2021.

[6] M. D. Armstrong, D. Maupomé, and M.-J. Meurs, "Topic modeling in embedding spaces for depression assessment," in *Proceedings of the Canadian Conference on Artificial Intelligence*, 2021.

[7] A. Gupta and Z. Zhang, "Vector-quantization-based topic modeling," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 3, 2021.

[8] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "The dynamic embedded topic model," *ArXiv*, vol. abs/1907.05545, 2019.

[9] J. Liu, C. Xia, X. Li, H. Yan, and T. Liu, "A bert-based ensemble model for chinese news topic prediction," in *Proceedings of the 2020 2nd International Conference on Big Data Engineering*, 2020, pp. 18–23.

[10] J. Zhang, M. Liu, and Y. Zhang, "Topic-informed neural approach for biomedical event extraction," *Artificial intelligence in medicine*, vol. 103, p. 101783, 2020.

[11] A. B. Dieng, F. J. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020.

[12] P. V. Tijare and J. R. Prathuri, "Correlation between k-means clustering and topic modeling methods on twitter datasets," *Cyber Security and Digital Forensics*, pp. 459–477, 2022.

[13] F. Viegas, S. Canuto, C. Gomes, W. Luiz, T. Rosa, S. Ribas, L. Rocha, and M. A. Gonçalves, "Cluwords: exploiting semantic word clustering representation for enhanced topic modeling," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 753–761.

[14] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[15] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *International conference on machine learning*. PMLR, 2016, pp. 1727–1736.

[16] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 830–839.

[17] K. Lang, "Newsweeder: Learning to filter netnews," in *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 331–339.

[18] M. Jin, X. Luo, H. Zhu, and H. H. Zhuo, "Combining deep learning and topic modeling for review understanding in context-aware recommendation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2018, pp. 1605–1614.

[19] T. A. Rana and Y.-N. Cheah, "Aspect extraction in sentiment analysis: comparative analysis and survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 459–483, 2016.

[20] K. Haas, Z. B. Miled, and M. Mahoui, "Medication adherence prediction through online social forums: A case study of fibromyalgia," *JMIR medical informatics*, vol. 7, no. 2, p. e12561, 2019.

[21] C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, and N. Dasgupta, "Digital drug safety surveillance: monitoring pharmaceutical products in twitter," *Drug safety*, vol. 37, no. 5, pp. 343–350, 2014.

[22] I. E. Agbehadji, B. O. Awuzie, A. B. Ngowi, and R. C. Millham, "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*, vol. 17, no. 15, p. 5330, 2020.

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

[24] K. Lee, A. Agrawal, and A. Choudhary, "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*, 2015, pp. 815–822.

[25] G. Gkotsis, C. Mueller, R. J. Dobson, T. J. Hubbard, and R. Dutta, "Mining social media data to study the consequences of dementia diagnosis on caregivers and relatives," *Dementia and Geriatric Cognitive Disorders*, vol. 49, no. 3, pp. 295–302, 2020.

[26] J. W. Mohr and P. Bogdanov, "Topic models: What they are and why they matter," *Poetics, 41(6)*, 2013.

[27] A. Karl, J. Wisnowski, and W. H. Rushing, "A practical guide to text mining with topic extraction," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 5, pp. 326–340, 2015.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[29] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[30] H. Patil and R. S. Thakur, "Document Clustering: TF-IDF Approach," pp. 264–281, 2016.

[31] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit," *Information Processing & Management*, vol. 57, no. 2, p. 102034, 2020.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1. Association for Computational Linguistics, 2019, pp. 4171–4186.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[34] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *ArXiv*, vol. abs/2004.05150v2, 2020.

[35] N. Reimers and I. Gurevych, "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, 2019.

[36] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[37] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[38] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional data streams," in *Proceedings of the Thirtieth international conference on Very large data bases*, vol. 30, 2004, pp. 852–863.

[39] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic mdoels," *Journal of Information Science*, vol. 22, 2009.

[40] M. Huang, M. Zolnoori, J. E. Balls-Berry, T. A. Brockman, C. A. Patten, and L. Yao, "Technological innovations in disease management: Text mining us patent data from 1995 to 2017," *Journal of medical Internet research*, vol. 21, no. 4, p. e13316, 2019.

[41] M. D. T. Nzali, S. Bringay, C. Lavergne, C. Mollevi, and T. Opitz, "What patients can tell us: topic analysis for social media on breast cancer," *JMIR medical informatics*, vol. 5, no. 3, p. e7779, 2017.

[42] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, 2019.

[43] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," *stat*, vol. 1050, p. 4, 2017.

[44] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," 2021, pp. 759–766.

[45] H. Zhao, L. Du, W. L. Buntine, and G. Liu, "Metalda: A topic model that efficiently incorporates meta information," *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 635–644, 2017.

[46] J. Rashid, S. M. A. Shah, and A. Irtaza, "Fuzzy topic modeling approach for text mining over short text," *Information Processing & Management*, vol. 56, no. 6, p. 102060, 2019.

[47] C. P. George, D. Z. Wang, J. N. Wilson, L. M. Epstein, P. Garland, and A. Suh, "A machine learning based topic exploration and categorization on surveys," in *2012 11th International Conference on Machine Learning and Applications*, vol. 2. IEEE, 2012, pp. 7–12.

[48] X. Cui, T. E. Potok, and P. Palathingal, "Document clustering using particle swarm optimization," in *Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005.* IEEE, 2005, pp. 185–191.

[49] Y. Chen, B. Qin, T. Liu, Y. Liu, and S. Li, "The comparison of som and k-means for text clustering." *Comput. Inf. Sci.*, vol. 3, no. 2, pp. 268–274, 2010.

[50] M. M. Rodrigues and L. Sacks, "A scalable hierarchical fuzzy clustering algorithm for text mining," in *Proceedings of the 5th international conference on recent advances in soft computing.* Citeseer, 2004, pp. 269–274.

[51] S. Karol and V. Mangat, "Evaluation of text document clustering approach based on particle swarm optimization," *Open Computer Science*, vol. 3, no. 2, pp. 69–90, 2013.

[52] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc., 2009.

[53] T. T. Hailu, J. Yu, and T. G. Fantaye, "A framework for word embedding based automatic text summarization and evaluation," *Information*, vol. 11, no. 2, p. 78, 2020.

[54] S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *Journal of Information Science*, vol. 43, no. 1, pp. 88–102, 2017.

[55] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[56] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, 2003, pp. 29–48.

[57] D. Mimno, H. Wallach, E. Talley, M. Leenders, and 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.

[58] J. H. Lau, D. Newman, and T. Baldwin, "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*, 2014, pp. 530–539.

[59] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 1105–1112.

A. DATA IN BRIEF: A SOCIAL AND NEWS MEDIA BENCHMARK DATASET FOR TOPIC MODELING

The present chapter describes the datasets and the unique identifier that can be used to retrieve the original document, provides the pre-processing scripts used for each dataset, the topic assignment for each document produced by pPSO and the topic keywords generated by the different topic modeling techniques. These datasets allow direct comparison with other topic modeling techniques. In order to further facilitate comparison, the algorithm underlying the proposed evolutionary topic modeling technique (pPSO) is also provided.

Specifications Table

Subject Area	Machine Learning
Specific Subject Area	Topic Modeling; Document Clustering
Type of Data	Tables
How Data is Acquired	The topic assignment generated by pPSO are provided using the unique r/-Cancer and 20NewsGroups document identifiers. The keywords for each topic generated by the topic modeling techniques are also included.
Data Format	Comma separated value (CSV) files indexed by either the unique document identifier or the topic number as applicable.
Description of Data Collection	The raw documents are extracted from the r/Cancer [16] and 20NewsGroups [17] archives. Text pre-processing is then performed on the raw documents including removal of high frequency words, punctuation, numerical characters as well as forcing the text to lower case. Three topic modeling techniques, namely, pPSO, ETM and NVDM are then applied to the pre-processed documents. The resulting pPSO topic assignment, topic keywords for the topic modeling techniques, text pre-processing scripts and pPSO algorithm are provided.
Data Source Location	The raw r/Cancer and 20NewsGroups documents are obtained from the respective archives in [16] and [17]. They can be extracted using the unique document identifier.
Data Accessibility	The datasets are publicly available through the OSF framework. **Repository Name:** A Social and News Media Benchmark Dataset for Topic Modeling **URL:** *osf.io/pz83w/*
Related Research Article	S. Miles, L. Yao, W. Meng, C. M. Black, Z. Ben-Miled, Comparing PSO-based Clustering over Contextual Vector Embeddings to Modern Topic Modeling, Journal of Information Processing & Management (in press).

Value of the Data

- Public sharing of the fully labeled data rather than a subset of the data generated by the topic models allows the replication and validation of the results as well as enables direct comparison with competing topic modeling techniques.

- A well defined benchmark from two different domains (i.e., social and news media) is an opportunity for a shared baseline for various NLP applications. It also allows for the exclusion of the text pre-processing techniques as a source of difference between various studies. The current benchmark can also be combined with other benchmarks, such as the one offered in [31], to construct an extended dataset.
- The labeled documents can help promote retrospective studies that investigate topics important to the r/Cancer subscribers as well as mainstream media during the year 2020.

Data Description

The data release consists of two types of files: document and keyword tables. Two document tables are provided, one for each source of data: RC documents.csv and 20NG documents.csv. The document tables consist of two columns. The first column is the index document ID. It uniquely identifies the document in the source archive. The remaining columns correspond to the cluster ID generated by the pPSO model with different number of topics. ETM and NVDM are generative models and do not produce a single topic assignment for each document. A keyword table is provided for each pair of data source and topic modeling technique. The table is organized by topic and the top ten topic keywords are provided for each topic. Table A.1 shows an example keyword table.

Table A.1. Structure of the keyword tables.

Topic_Num	Keyword 1	Keyword 2	Keyword 3	...
1	steelers	nfl	colts	
2	german	germany	merkel	
3	weinstein	harvey	weinsteins	
...	...	...	...	

Experimental Design, Materials and Methods

The protocol used to generate the topic models is provided under methods in the Repository. This protocol defines a data processing pipeline that consists of multiple steps:

- Data cleaning: This step consists of the removal of stop words, high frequency words, low frequency words and punctuation from each raw document. The text is also forced to lower case.

- Vector Embedding: In this step, the initial embedding vector for each document is generated from the pretrained language model sBERT [35]. This embedding is then reduced to a dimension of 300 using UMAP [37].

- Topic Modeling: The source code for topic modeling using pPSO on the reduced embedded vectors allows for the clustering of the documents into distinct topics. The source code for ETM and NVDM can be found in [11] and [15], respectively.

- Data Analysis: This is the final step in the pipeline. The topic keywords are extracted for each topic cluster and the topic coherence and diversity metrics are computed.

Ethics Statements

The raw r/Cancer data are extracted from a public domain archive [1]. Only the document identifiers are included in the present dataset. Users can retrieve the text data directly from the archive as per the Reddit redistribution and data sharing policies. The raw 20NewsGroups data is publicly available. However, we still follow the same procedure as in the r/Cancer dataset of only including the document identifier in the current dataset.

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