HHS Public Access

Author manuscript

*Artif Intell Med*. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

*Artif Intell Med*. 2021 July ; 117: 102096. doi:10.1016/j.artmed.2021.102096.

# Evaluation of clustering and topic modeling methods over health-related tweets and emails

**Juan Antonio Lossio-Ventura**[*],
Stanford Center for Biomedical Informatics Research, Stanford University, 1265 Welch Road, 94305-5479, Stanford, California, USA

**Sergio Gonzales**,
Stanford Center for Biomedical Informatics Research, Stanford University, 1265 Welch Road, 94305-5479, Stanford, California, USA

**Juandiego Morzan**,
Universidad del Pacífico, Av. Salaverry 2020, Jesús María, 15072, Lima, Peru

**Hugo Alatrista-Salas**,
Universidad del Pacífico, Av. Salaverry 2020, Jesús María, 15072, Lima, Peru

**Tina Hernandez-Boussard**,
Stanford Center for Biomedical Informatics Research, Stanford University, 1265 Welch Road, 94305-5479, Stanford, California, USA

**Jiang Bian**
Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, 2004 Mowry Road, 32610, Gainesville, FL, USA

## Abstract

**Background:** Internet provides different tools for communicating with patients, such as social media (e.g., Twitter) and email platforms. These platforms provided new data sources to shed lights on patient experiences with health care and improve our understanding of patient-provider communication. Several existing topic modeling and document clustering methods have been adapted to analyze these new free-text data automatically. However, both tweets and emails are often composed of short texts; and existing topic modeling and clustering approaches have suboptimal performance on these short texts. Moreover, research over health-related short texts using these methods has become difficult to reproduce and benchmark, partially due to the absence

E-mail addresses: jlossio@stanford.edu, juan.lossio@nih.gov (J.A. Lossio-Ventura).

Author's contributions

JALV conceived and designed the study. JALV, HAS, JM, and SG collected the data, set up the applications, and performed the evaluation. JALV, HAS, JM, and SG wrote the initial draft and revised subsequent versions. JB and THB provided relevant feedback. JB and THB, senior investigators, led the research project. All authors read, revised and approved the final manuscript.

Conflict of interest

The authors declare that they have no conict of interest.

of a detailed comparison of state-of-the-art topic modeling and clustering methods on these short texts.

**Methods:** We trained eight state-of- the-art topic modeling and clustering algorithms on short texts from two health-related datasets (tweets and emails): Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA), LDA with Gibbs Sampling (GibbsLDA), Online LDA, Biterm Model (BTM), Online Twitter LDA, and Gibbs Sampling for Dirichlet Multinomial Mixture (GSDMM), as well as the *k*-means clustering algorithm with two different feature representations: TF-IDF and Doc2Vec. We used cluster validity indices to evaluate the performance of topic modeling and clustering: two internal indices (i.e. assessing the goodness of a clustering structure without external information) and five external indices (i.e. comparing the results of a cluster analysis to an externally known provided class labels).

**Results:** In overall, for number of clusters (*k*) from 2 to 50, Online Twitter LDA and GSDMM achieved the best performance in terms of internal indices, while LSI and *k*-means with TF-IDF had the highest external indices. Also, of all tweets (N=286,971; HPV represents 94.6% of tweets and lynch syndrome represents 5.4%), for $k = 2$, most of the methods could respect this initial clustering distribution. However, we found model performance varies with the source of data and hyper-parameters such as the number of topics and the number of iterations used to train the models. We also conducted an error analysis using the Hamming loss metric, for which the poorest value was obtained by GSDMM on both datasets.

**Conclusions:** Researchers hoping to group or classify health related short-text data can expect to select the most suitable topic modeling and clustering methods for their specific research questions. Therefore, we presented a comparison of the most common used topic modeling and clustering algorithms over two health-related, short-text datasets using both internal and external clustering validation indices. Internal indices suggested Online Twitter LDA and GSDMM as the best, while external indices suggested LSI and *k*-means with TF-IDF as the best. In summary, our work suggested researchers can improve their analysis of model performance by using a variety of metrics, since there is not a single best metric.

### Keywords

Topic modeling; Clustering; Internal cluster indices; External cluster indices; Natural language processing

## 1 Introduction

Several social networking and microblog platforms have emerged exponentially in the last decade. Social networks such as Twitter enable users to interact with each other and share information on a wide range of different topics. Twitter is one of the most popular social media platforms intersecting all types of contents, including health-related texts. Twitter enables users to write short messages, called “tweets”, composed of 280 characters (140 characters before September 2017). Tweets are often adopted to share opinions, feelings, thoughts, and personal activities. With over 500 million tweets posted each day, Twitter has become a very valuable data resource to get real-world insights. In healthcare domain, Twitter has also been adopted by users to share their personal health status, their experience with the care and treatment options with other users with similar conditions/diseases

and symptoms as well as more broadly sharing and seeking health information of their interest, attracting the attention of clinical and biomedical researchers with the ultimate goal to improve patients' outcomes [130,40,154,139]. There have been various existing studies that demonstrated the use of Twitter as a low-cost data source for public health surveillance [138,108], such as for influenza vaccination [61], mental health [32,155], human papillomavirus (HPV) vaccination [153], tobacco [99,31], opioid [83], public mood [104], suicide [23], etc.

Furthermore, email is becoming popular in health care to establish and improve interactions between patients and healthcare professionals [36]. Emails allow patients to participate more actively in their health care, which can improve the quality and accessibility of health services [20,18]. Indeed, patient-physician email communication has been addressed in various studies [13], such as for detection of depression [137], rural family health practice [27], multiple sclerosis [51], disease prevention [126], coordination of healthcare appointments [18], communication between healthcare professionals [106], among others. Several of these studies found positive effects in the use of emails such as the improvement of clinic efficiency and cost-effectiveness [39,48,27].

As a result, Twitter and emails have created a vast amount of short texts. Several natural language processing (NLP) methods, such as topic modeling and clustering, have been adopted to digest and assess these short texts, allowing us to infer patients' interests, track new health-related stories, and identify emerging health topics. Clustering seeks to split documents into a certain number of groups based on a similarity metric. Topic modeling seeks to discover latent topics that describe the collection of documents. A topic represents a group of words that frequently occur together. There are numerous works that have used classic clustering methods (e.g., *k*-means) on short texts such as tweets [120,86,90,148]. Diverse topic modeling methods also have been proposed to analyze short texts from different fields. Two of the most popular methods are the latent dirichlet allocation (LDA) [22] and latent semantic indexing (LSI) [56]. There exist various LDA-based techniques applied to text from various domains, including biomedicine [67]. Also, several recent approaches have adopted the Dirichlet Mixture Model for short text clustering [149,150,75]. Despite the abundance of NLP techniques available in the literature, there are several challenges when analyzing tweets [10]: significant noise and inconsistent tweeting behaviours of user prevent researchers from leveraging the full potential information carried in tweets.

Moreover, health research using Twitter and emails are difficult to measure because of the lack of comparisons between the various existing applications. As December 2020, we identified two recent studies that compared several topic modeling and clustering methods on several short text datasets. The first study [117] evaluated nine topic modeling based on DMM, global word co-occurrence, and self-aggregation. They found that simpler methods such as GSDMM [149] and BTM [147,30] were the most suitable with respect to effectiveness and efficiency. The second study [33] evaluated the performance of four classic clustering algorithms (with four different feature representations such as TF-IDF and Doc2Vec) and a topic modeling method (LDA). The experiments showed that the best performance was achieved by *k*-means with Doc2Vec representation. However, there exist

several gaps in these two studies: (1) [117] did not consider LSI or any LDA-based method, (2) [117] did not consider any classic clustering algorithm, (3) [33] considered only LDA as topic modeling, (4) both used small datasets ($\leq 30K$ docs), (5) both used external validity indices only (i.e., comparing the results of a cluster analysis to an externally known provided class labels), and (6) both used a predefined number of topics for the evaluation, since each dataset was previously annotated.

In this paper, we seek to fill the gaps previously mentioned in order to discover how effectively several standard topic modeling and clustering methods perform on health-related tweets and emails. Therefore, we evaluate the performance of several state-of-the-art topic modeling and clustering algorithms (including those suggested in [117,33]) on short texts from two health-related datasets. The first dataset is composed of tweets ($\leq 290K$ docs) and the second is composed of emails ($50K$ docs). We consider individual tweets and emails as single documents, respectively. We include seven topic modeling approaches including LSI, LDA, GibbsLDA [142], Online LDA [57], BTM [147, 30], Online Twitter LDA [76], and GSDMM; as well as the *k*-means clustering algorithm with two different feature representations: TF-IDF and Doc2Vec. We use cluster validity indices to evaluate the performance of topic modeling and clustering: two internal (i.e., assessing the goodness of a clustering structure without external information) and five external validity indices.

The remainder of the paper is organized as follows. We will review the literature in the "Related work" section. We will explain our approach in the "Methods" section. The results of the experiments and evaluations of the topic modeling and clustering applications will be presented in the "Experiments and results" section. We will discuss the obtained findings in the "Discussion" section. Finally, we will conclude the current work and present future directions in the "Conclusions" section.

## 2 Related work

In this section, we review related work from short text clustering, topic modeling, and validity indices.

### 2.1 Clustering of short texts

Clustering is an unsupervised machine learning method that seeks to partition objects into a certain number of clusters (i.e., groups or subsets) based on a similarity metric. Generally, clustering methods applied on text data are based on vector representations; such as bag-of-words (BoW) or term frequency-inverse document frequency (TF-IDF); and then grouping texts based on their similarity [85,89,90,21,84]. These techniques are frequently applied to several information retrieval tasks such as event detection [65,93,115,82] and text summarization [105,128,86,148]. There are several works using classic clustering methods on short texts, such as tweets, for instance, a study [120] compared three well-known clustering algorithms: *k*-means, Singular Value Decomposition, and Affinity Propagation on over 600 tweets, and found that Affinity Propagation [46] had better performance. However, its complexity associated with the number of documents was quadratic, thus Affinity Propagation is not suitable for larger datasets. Other approaches focused on variations of

classic clustering techniques considering several tweet components such as texts, hashtags, users, and temporal aspect (e.g., stream clustering) [129,86,75].

Short text clustering represents a big challenge due to the data sparsity, since most words co-occur once or twice in the dataset [10]. Several sparseness-resistant methods were proposed to face this challenge such as text augmentation [19,156], topic modeling [147,30], neural networks [145,52], and Dirichlet Mixture Model [149,150,75]. Data augmentation methods seek to enrich data representation with external resources, such as Wikipedia [19,146]; or similar words by exploiting related text documents [69,133,35,156]; or the incorporation of semantic features from ontologies, terminologies, and dictionaries, such as WordNet, DBpedia, Freebase [59,45,26,141].

Moreover, recent approaches based on low-dimension representations with neural network [145] proved to be effective to tackle the sparsity problem in short text clustering [135,38,47,52], for instance using word embeddings [96,110], sentences embeddings [77,72] and document embeddings [34]. Also, several studies explored sophisticated models for short text clustering. For instance, a work proposed a Dirichlet Multinomial Mixture model-based approach for short text clustering (GSDMM) [149] which also infers the number of clusters and obtained the best performance when compared with clustering algorithms such as *k*-means [92], Hierarchical Agglomerative Clustering (HAC) [94], and DMAFP [60].

## 2.2 Topic modeling

Topic Modeling is also an unsupervised machine learning method mainly based on statistical properties of the data to discover “topics” that describe the collection of documents. Topic modeling methods seek to extract topics from a set of documents based on statistical techniques. Each topic is defined as a distribution over a set of words. Diverse topic modeling methods have been proposed to analyze texts from different fields including politics, medicine, and psychology. Two of the most popular methods are the latent dirichlet allocation (LDA) [22] and latent semantic indexing (LSI, a.k.a. LSA) [56]. A recent work has exhaustively listed LDA-based techniques proposed from 2003 to 2016 applied to text from various domains, including biomedicine [67].

Some topic modeling algorithms were proposed to work on general health and medical text [70,71]. Moreover, other proposed methods specifically aimed to predict therapy outcomes from emails sent by patients under treatment for a social anxiety disorder [58]; predict protein-protein [17], gene-drug [143] relations from biomedical literature; discover concepts in patients’ health records [16]; detect depression [124,137]; recognize genuine suicide notes from notes written by healthy subjects [111]; classify patient issues from their experience and the result of using a particular drug [68]; improve the automating classification of patient portal messages through the use of semantic features and word context [132]; identify patterns of events from medical reports of brain cancer patients [15]; treatment behaviors [62,63]; treatment activities [28]; determine patient mortality [49]; discover models of disease and phenotypes [112]; extract biological terminology [97]; discover biological processes [81]; among others. Topic modeling methods have also been applied over health-related tweets to identify latent health topics [107]. Hybrid approaches have also allowed to extract health trends in tweets by the integration of visualization approaches

with classical topic models [114,113]. Also, diverse studies addressed specific tasks such as grouping opinions about HPV vaccines leveraging also community structure methods [134]; identify common obesity-related themes through a combination of geographic information systems and topic modeling methods [50]; identify the associations of Zika-related topics, such as attitudes, knowledge, and behaviors [44].

As clustering methods, topic modelling can also be used for clustering by giving a probability distribution over a number of topics for each document. Indeed, clustering and topic modeling methods have been used for clustering tasks and have been compared in different studies. For instance, the authors in [117] proposed three categories for topic modeling based on: (1) DMM, (2) global word co-occurrence, and (3) self-aggregation. Then they compared nine different topic modeling techniques from these categories: (1) GSDMM [149], LF-DMM [102], GPU-DMM [80], GPU-PDMM [79], (2) BTM [30], WNTM [158], and (3) SATM [118], PTM [157]. They found that: (i) strategies that use word embeddings (LF-DMM, GPU-DMM, and GPU- PDMM) are very promising in short text topic modeling, (ii) the highest computation costs were obtained with LF-DMM and GPU-PDMM (i.e., they are not suitable for large datasets), and (iii) simpler methods (GSDMM and BTM) are the most suitable with respect to effectiveness and efficiency. For the clustering task, GSDMM achieved the best results. Another related work [116] described an approach with Gibbs sampling called PYPM. This model was tested on four short text datasets and compared with five well-known techniques (Non-negative Matrix Factorization [78], LDA, DMAFP, GSDMM, and FGSDMM [151]). Results showed that PYPM had the best results followed by GSDMM.

Moreover, another recent study [33] evaluated the performance of four clustering algorithms (*k*-means, *k*-medoids, Hierarchical Agglomerative Clustering, and Non-negative Matrix Factorization) and a topic modeling method (LDA) on short texts from social networks such as Twitter and Reddit. The paper also evaluated four different feature representations including TF-IDF, Word2Vec, Word2Vec weighted with the top 1,000 TF-IDF scores, and Doc2Vec. The experiments showed that the best performance was achieved by *k*-means with Doc2Vec on both datasets.

Of note, topic modeling methods can be evaluated from several aspects such as from cluster evaluation, topic coherence, and classification evaluation. To compare clustering and topic modeling methods, we need to apply cluster validity indices. For this purpose, after using topic modeling to compute topic probabilities, the maximum topic probability of each document is selected to get the cluster label of each document [24,88,144,147,116,117]. Then, cluster validity indices are applied to evaluate their performances.

### 2.3 Validity indices

Cluster validity indices are metrics to validate clustering results and to find natural structures for a given dataset [152,14]. In other words, validity indices seek to find optimal partitions that are well compacted and well-separated from other partitions [54]. There are two kinds of cluster evaluation metrics which are called external and internal validation indices [53]. External indices measure the quality based on ground-truth labels, for instance Rand [119], Adjusted Rand Index [64], Fowlkes–Mallows, Variation of Information [11]. Internal indices

evaluate the result on information intrinsic to the data alone. The latter is useful when there is no annotated dataset available and the usual approach focus on evaluate the compactness and separation of clusters, such as Dunn [37], Calinski–Harabasz [25]. Several studies on cluster validity indices concluded that there is not a single best metric [42,95,14]. Also, they found that the performance of cluster validity indices decrease considerably when there is noise or clusters overlap.

Recent works have compared topic modeling and clustering methods on short text clustering. Most of them used annotated datasets for the experiments, therefore, they mainly used external indices to measure the quality of clusters, such as Homogeneity (H) [122], Completeness (C) [122], V-Measure (V) [122], Adjusted Rand Index (ARI) [64], Normalized Mutual Information (NMI) [29], Adjusted Mutual Information (AMI), Accuracy (ACC), F-measure, Entropy, Purity. For instance, the authors that introduced GSDMM [149] used five external indices such as H, C, ARI, NMI, and AMI to evaluate their model. Another study proposed an online semantic-enhanced Dirichlet model for short text clustering (OSDM) [75] and considered several indices such as NMI, V, H, and ACC to validate their results. In [52] the authors reported the evaluation of various text representations and self-training methods with ACC and NMI. A recent study [117] compared nine topic modeling techniques in clustering tasks using two external metrics: NMI and Purity. Moreover, a recent work evaluated one topic modeling and fours classic clustering methods [33] using three external indices such as NMI, AMI, ARI.

On the other hand, there are also several commonly used internal indices when assessing the goodness of short text clustering such as Calinski-Harabasz (CH) [25], Silhouette Coefficient (SC) [123], Dunn [37], Duda [43], Elbow [136], among others. The internal indices can also be used to determine the optimal number of clusters in the data [98]. Internal indices in comparison to external ones usually detect improvements in the clustering distribution which have positive implications in the system evaluation [66]. In our previous work [87], we evaluated topic modeling and clustering methods using only tweets and two internal indices (CH and SC). However, most studies previously cited that compared topic modeling and clustering methods did not use internal validity indices to evaluate their results.

Therefore, in this paper, we use seven validity indices: five external (NMI, ARI, H, C, and V) and two internal (CH and SC). We selected the five most common external indices used in the literature that are also independent of the absolute values of the labels in comparison to F-measure, ACC, among others. Moreover, we included two internal indices: CH index which has demonstrated in several works to be effective [12], and SC which is one of the most well-known measures and provides graphical representations of how well each element has been classified. We used the implementation of these metrics in sklearn [109] in the experimental study.

## 3 Methods

This section describes our study for evaluating state-of-the-art topic modeling and clustering methods to automatically extract relevant topics from health-related tweets and emails.

Figure 1 outlines our approach with the basic steps for this evaluation. In this section we describe: (1) the two datasets used: tweets and emails; (2) the applications based on topic modeling and clustering algorithms; and (3) the validity indices used to assess the clusters defined by the algorithms we studied.

## 3.1 Datasets

**3.1.1 Tweets dataset—**Our tweets dataset is an unbalanced collection composed of two subsets: human papillomavirus (HPV) and lynch syndrome. The HPV represents 94.6% of all tweets while lynch syndrome represents 5.4% of all tweets. The extraction strategy considered keywords and hashtags[1] containing common generic HPV and lynch syndrome names and colloquial terms. Table 1 shows a description of our tweets collection. This dataset contains a total of 286,971 tweets containing at most 140 characters. Table 2 shows a sample of eight tweets related to HPV and lynch syndrome extracted from our tweets dataset. The average of number of characters per tweet is 60.6 and the average of number of tokens is 5.5. We applied several rules to preprocess the tweets collection: 1) text was changed to lowercase; 2) suppression of duplicated tweets; 3) suppression of stop-words; and 4) omission of links from the tweets.

**Annotation of tweets:** We only annotated tweets that contained hashtags, thus, a total of 126,083 tweets were labeled (115,859 HPV and 10,224 lynch syndrome). The annotation was semi-automatically performed. First, we selected the most frequent hashtags. We then manually selected the most relevant hashtags and grouped them by their semantic similarity, for instance “#vaccine”, “#vaccines”, and “#vax” represented a single group. We then manually selected the top 50 more frequent hashtag groups. Table 3 shows a sample of the top 10 groups of hashtags. Finally, we created a script to automatically annotate the tweets with the groups (labels) previously created. Of note, for our evaluation, the dataset was annotated into a different number of topics: first, the dataset was annotated into 2 topics, then 3, until 50 topics.

**3.1.2 Emails dataset—**We accessed the 50,000 available emails sent by patients with prostate cancer to their health care providers. The emails are apart of a clinical data warehouse at a tertiary academic care center from 2010 to 2019 [127]. Table 4 shows a description of the emails collection. We processed the emails to preserve the confidentiality, integrity, and availability of protected health information (PHI). Thus, we arbitrarily formatted text emails into uniformly formatted emails such that: 1) all text was lowercase; 2) generic tokens for dates, days, time, email address, and URLs replaced specific occurrences; 3) named entities such as people, organizations, and locations were replaced with generic tokens.

**Annotation of emails:** We annotated the vast unstructured, free-text emails by labeling each document with the topics with a similar method as we annotated the tweets. After thoroughly reading several hundred emails, we defined 2, then 3, up to 50 topics by grouping tokens with significant semantic information together (*e.g.* germ, infection) and

[1]Hashtag is a word or phrase preceded by a hash sign (#) to identify messages on a specific topic.

then labeled the emails based on the most frequent token(s) corresponding to the topics that we defined. Table 5 lists 10 of the most frequently occurring topics. We removed 13 emails from the analysis because they did not contain enough meaningful tokens after processing to assign them one of our external labels.

## 3.2 Applications

We used several available online implementation of topic modeling[2] and clustering methods. To cover every aspect, we briefly describe all used systems along with our experiments.

### 3.2.1 Topic modeling—

We set up seven well-known available methods used for short texts.

- **Latent Semantic Indexing (LSI):** a well-known information retrieval algorithm [41, 6]. LSI has been applied to a wide variety of learning tasks, such as search and retrieval, classification and filtering. LSI uses singular value decomposition of vector space spanned by the documents to describe latent semantics within the collection of documents.
- **Latent Dirichlet Allocation (LDA):** is a generative probabilistic model seeking to describe a set of observations as a mixture of distinct categories [22,5]. An observation is a document which represents a mixture of topics. Each topic is represented as a mixture of distributions of words. With these distributions, one can compute the probability that a document part of topic based the words used in that document. LDA uses a Bayesian [22] approach to learn the distributions: set of topics, word probabilities, etc.
- **LDA with Gibbs Sampling (GibbsLDA):** Gibbs Sampling is another technique for parameter estimation and inference of the distributions defined in the LDA model [142,3]. It was designed to analyze hidden and latent topic structures from large-scale datasets including large collections of documents from the Web. The LDA method with Gibbs Sampling has been shown to be comparable to *k*-means in terms of computational costs and execution time.
- **Online LDA:** the online variational Bayes algorithm for LDA (Online LDA) is based on online stochastic optimization [57,7], which has been shown to find good parameter estimates much faster than batch algorithms on large datasets. Online LDA analyzes a massive number of documents without storing the documents in a dataset; each can arrive in a stream and then be discarded after one look.
- **Biterm (BTM):** topics are learnt from short texts by directly modeling the generation of word co-occurrence patterns (i.e., biterms) in the text corpus [147,1]. Each latent topic is presented as a significance probability as well as a probability distribution over a vocabulary. The experimental results showed

[2]Most of the implementations were extracted using gensim version 3.8.3 [121], an opensource library for unsupervised topic modeling.

that BTM produces discriminative topic representations as well as more coherent topics for short texts.

- **Online Twitter LDA:** tracks emerging events in microblogs [76,8]. In contrast to other LDA algorithms, this method employs a built-in update mechanism that uses time slices and creates a dynamic vocabulary. In every update, words that do not reach a frequency threshold are removed and a new word is added when it reaches the threshold. This allows the model to study the topic evolution and detect emerging topics over time. The input is discretized time slices and documents partitioned into these slices. Thus, the model is able to process the input and update the model periodically; generate comparable topics throughout different time slices that allows topic shift evolution measurement; ensure sensitivity to the changes of topic over time. There are two main differences with Online LDA. First, the convergence is handled by the introduction of a new parameter called contribution factor, which reduces the influence from the previous model. Second, while Online LDA assumes a fixed vocabulary, Online Twitter LDA considers that it is not possible to calculate the vocabulary ahead of time and, thus, constructs a dynamic vocabulary.
- **GSDMM:** is a collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (DMM) applied to short text [149,101]. DMM is a probabilistic generative model for documents, and embodies two assumptions about the generative process [103]: (1) the documents are generated by a mixture model, and (2) there is a one-to-one correspondence between mixture components and clusters. Thus, GSDMM assumes each document belongs to a single topic, which is a suitable assumption for some short texts. Given an initial number of topics, this algorithm groups documents and extracts topic structures that are present in the datatset. If the number of topics is set to a high value, then the model will be able to automatically learn the number of topics.

#### 3.2.2 Clustering—

In this study, we use one of the most well-known algorithms, *k*-means (with *k*-means++ initialization) [91,4], with two different dataset representations: TF-IDF and Doc2Vec.

- **TF-IDF**: [125,9] term frequency-inverse document frequency, is a statistic measure intended to reflect how important a word is for a document in a collection of documents.
- **Doc2Vec:** is a simple extension of *word2vec* to include the learned embedding of word sequences rather than words alone [77,2]. It has been shown to out perform similar embedding techniques in terms of accuracy and computational cost.

### 3.3 Analysis of applications

Note that to evaluate the performance of clustering and topic modeling methods we need to apply cluster validity indices. Thus, after executing topic modeling to compute topic probabilities, the highest probability of each document is selected to get the cluster label of each document [117]. Then internal and external validity indices are applied to assess their

performances. Therefore, we performed an evaluation of all topic modeling and clustering applications in terms of: 1) experiments and results over the tweets dataset, a complete calculation of two internal and five external indices, and 2) experiments and results over the emails dataset, also a comparison of the internal and external indices. In the paragraph below, we explain the configuration for our evaluation, as well as the seven indices used.

**3.3.1 Configuration—**For topic modeling, "$k$" (i.e., number of topics) will range from 2 to 50. In our work, topic modeling results are used to classify tweets and emails to a particular topic. Each tweet/email is represented by a feature vector, where each component of the vector is the probability of the tweet/email to belong to a given topic. For instance, $k$=2 implies the size of the feature vector is 2 while for $k$=50 is 50. We then use the *argmax* function to determine the most prominent topic of each tweet.

The **clustering algorithm**, $k$-means, uses two document representations: TF-IDF and Doc2Vec. We tested various sizes of feature vectors (bag-of-words): 100, 200, 500, and 1,000 most frequent words after preprocessing (stop-words removal, deletion of punctuation, and correction of misspelled words) and considering only noun words. We determined that results were very similar using the four variations. Therefore, in this paper, we consider the 100 most frequents words as the number of features for TF-IDF and Doc2Vec, with a "$k$" (i.e., number of clusters) also ranging from 2 to 50.

Parameters of **topic modeling** are set as suggested in previous studies to get the optimal performance on short texts. For LDA, the hyper-parameters are set to $\alpha = 0.05$ and $\beta = 0.01$ as suggested in [147]. For GibbsLDA [147], $\alpha = 0.05$ and $\beta = 0.01$. For BTM [147], the parameters settings are $\alpha = 50/k$ and $\beta = 0.01$, where $k$ represents the number of topics. Online LDA [57] uses $\alpha = 1.0/k$, $\beta = 0.01$, and $\theta = 1$. Online Twitter LDA [76] sets $\alpha = 0.001$, $\beta = 0.01$, and $c$ (contribution factor) = 0.5. GSDMM [149] is set with $\alpha = 0.1$ and $\beta = 0.1$. Note that there is no distinction in the use of "$k$" when discussing the number of clusters and topics.

We evaluated all topic modeling and clustering algorithms using 100, 500, and 1,000 iterations. The initial number of iterations is recommended in [55] and is a default value in the applications.

**3.3.2 Internal indices—**We used two internal measures: Calinski-Harabasz index (CH) [25] and Silhouette Coefficient (SC) [123]. CH index has demonstrated in several works to be an effective measure for determining the most appropriate number of clusters [12]. On the other hand, SC is one of the most well-known measures and provides graphical representations of how well each element has been classified. Next, we explain the principles of the internal indices.

- **Calinski-Harabasz:** also known as the Variance Ratio Criterion. A higher $CH$ value indicates the model has well defined clusters. The $CH_k$ value is given by the ratio between average inter-cluster dispersion matrix ($B_k$) and intra-cluster dispersion matrix ($W_k$) as defined in Formula 1.

$$\mathbf{CH_k} = \frac{\mathrm{B_k}}{\mathrm{W_k}} \times \frac{n-k}{k-1} \quad (1)$$

where $n$ is the total number of points and $k$ the number of clusters. The $B_k$ value is based on the distance between clusters and is defined as:

$$\mathbf{B_k} = \sum_{i}^{k} n_i \cdot dist^2(c_i - c)$$

where $n_i$ is the number of elements of cluster $C_i$, $c_i$ is the center of $C_i$, and $c$ is the center of the complete dataset. $W_k$ is based on the distance within clusters and is defined as:

$$\mathbf{W_k} = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(c_i, x)$$

where $x$ is a point of cluster $C_i$. Note that to obtain well separated and compact clusters, $B_k$ is maximized and $W_k$ minimized. Therefore, the maximum value of $CH$ indicates a suitable partition for the dataset.

- **Silhouette Coefficient**: describes the separation distance between clusters. A width is computed for each point, which depends on its membership cluster. The widths are then averaged over all observations for each $k$. The $SC$ value has a range of [−1,1], where −1 represents poor clustering quality or poorly defined clusters and 1 high clustering quality or well-defined clusters. The $SC_k$ value for a single sample is defined in Formula 2.

$$\mathbf{SC_k} = \frac{1}{n} \times \sum_{i}^{n} \frac{b_i - a_i}{max(a_i, b_i)} \quad (2)$$

where $n$ represents the total number of elements in a cluster, $a_i$ is the average distance between an element $i$ of the cluster and all other elements within the same cluster, $b_i$ represents the average distance between the element $i$ of the cluster and all other elements in the nearest cluster.

In summary, higher clustering quality of a particular algorithm tends to yield higher predictive performance on information retrieval tasks. For this reason, we seek to identify the algorithms that maximize the overall clustering quality (i.e., internal indices).

**3.3.3 External indices—**Note that external evaluation measures can be applied when class labels for each data point in some evaluation set can be determined a priori. We used five well-known external measures for evaluation over the annotated dataset: Normalized Mutual Information (NMI) [131], Adjusted Rand Index (ARI) [64], V-measure (V) [122], Homogeneity (H) index [122], and Completeness (C) score [122].

- **Normalized Mutual Information**: is a normalization of the Mutual Information score, which scales the results between 0 (no mutual information) and 1 (perfect correlation) as defined in Formula 3.

$$\mathrm{NMI}(\mathbf{Y}, \mathbf{C}) = \frac{2 \times I(Y \mid C)}{[H(Y) + H(C)]} \quad (3)$$

where Y represents the class values, C the cluster labels, H the entropy, and I(Y,C) is the Mutual Information between Y and C, and is defined as:

$$\mathbf{I}(\mathbf{Y}, \mathbf{C}) = H(Y) - H(Y \mid C)$$

where H(Y) is the entropy of class labels, and H(Y|C) is the entropy of class labels within each cluster.

- **Adjusted Rand Index**: the rand index (RI) computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. RI gives a value between 0 and 1, where 1 indicates that the data in clusterings are the same. This measure can be seen as the percentage of correct decisions made by the algorithm, it can be expressed as:

$$\mathbf{RI} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. The Adjusted Rand Index rescales the RI, considering that random chances will cause some objects to occupy the same clusters. The ARI is calculated using the Formula 4.

$$\mathbf{ARI} = \frac{(RI - Expected\_RI)}{(max(RI) - Expected\_RI)} \quad (4)$$

- **V-measure**: is an entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied. V-measure is computed as the harmonic mean of distinct homogeneity and completeness scores,
- **Homogeneity**: a cluster has perfect homogeneity if all members of that cluster have the same external label. That is, the class distribution within each cluster contains only one class or equivalently has zero entropy. We determine how close a given clustering is to this ideal by examining the conditional entropy of the class distribution given the proposed clustering.
- **Completeness**: : is similar to homogeneity as it also describes how well the elements of the same external class are assigned to a single cluster. To evaluate

completeness, we examined the distribution of cluster assignments within each class. Completeness is formally defined as the conditional entropy of the proposed cluster distribution given the external class label.

## 4 Experiments and results

The focus of this study is to compare the performance of the applications cited below using internal and external indices over the tweets and emails datasets. Thus, next sections presents the results obtained for $k$={2,5,10,50}.

### 4.1 Results on the tweets dataset

**4.1.1 Internal indices—**We perform experiments using CH and SC to measure the performance of the topic modeling and clustering algorithms. Tables 6, 7, 8 and 9 show the CH and SC results for the seven topic modeling methods and *k*-means algorithm with Doc2Vec and TF-IDF for 2, 5, 10, and 50 number of clusters/topics (*"k"*) respectively. Of note, CH and SC seek to evaluate the clusters/topics based on two aspects: the similarity of tweets within the same cluster (cohesion), and the difference between the tweets of different clusters.

In addition, Figure 2 shows a general overview of the performance of the applications based on SC and CH for 100, 500, and 1,000 iterations; and for "*k*" ranging from 2 to 50. In all cases, the best values are obtained by Online Twitter LDA followed by GSDMM.

**4.1.2 External indices—**We perform experiments using the five external indices (NMI, ARI, V, H, and C) to measure the performance of the topic modeling and clustering algorithms. Tables 10, 11, 12 and 13 depict the NMI, ARI, V, H, and C values obtained with all topic modeling and clustering methods for 2, 5, 10, and 50 number of clusters/topics ("*k*") respectively. Of note, external indices measure the extent to which cluster labels match externally supplied class labels.

We also plotted the values obtained with NMI, ARI, V, H, and C with 100 iterations only as shown in Figure 3. In general, the best values are obtained by LSI followed by *k*-means with TF-IDF. Also, note that Online Twitter LDA significantly decreased its performance in comparison to the values obtained in the internal indices evaluation. It obtained the lowest performance, while other algorithms such as LSI, BTM, and techniques such as TF-IDF improved and in general, they are the three methods with the best performance.

### 4.2 Results on the emails dataset

**4.2.1 Internal indices—**We trained each model for integer values of $k$ ranging from 2 to 50, for 100 iterations and measured each model's performance with SC and CH scores. Online Twitter LDA was initially the best performing model both in terms of CH and SC scores of 8.2 million and 0.94, respectively, but in contrast to the results of the previous experiments, we saw a greater decrease in the model's performance, relative to LDA and Online LDA, as $k$ increased. The GSDMM model had the most stable, high level performance, with a SC that never dropped below 0.86, well above the next best performing model, LDA, which had a SC score of 0.65 (for $k$ = 50). LDA and Online LDA achieved the

second and third best SC scores once $k$ exceeded 30. LSI model had the worst performance, achieving negative SC values for almost all values of $k$.

The CH scores rapidly and substantially decreased with increasing values of $k$ for all models. The rates of decreases performance were not uniform as the Online Twitter LDA stated out well above the rest at 8.2 million for $k = 2$ and dropped to 5,719 for $k = 50$ behind the LDA model, which had a CH score of 8,062. The GSDMM model became the best performing model once $k$ exceeded 12 and was the only model for which all CH scores were greater than 10,000.

Table 14 lists the values of the SC and CH scores for the models with 2, 5, 10, and 50 clusters/topics while Figure 4 illustrates the SC and CH values for all values of $k$.

**4.2.2 External indices—**We also analyzed the same email clustering/topic models for external validity indices. Table 15 displays the NMI, ARI, V, H, and C for models with 2, 5, 10, and 50 clusters/topics. It is worth noting that all metrics of external validity range between 0 and 1. Figure 5 illustrates our findings for all values of $k$ between 2 and 50. Notably, we see that the $k$-means with TF-IDF embedding and LSI the are best performing algorithms across all metrics while Online Twitter LDA and $k$-means with Doc2Vec embedding showed the worst performance consistently across all 5 metrics of external validity. Overall, we see that all models improve along measures of external validity with increasing values of $k$. This trend is rather gradual starting at values 0.01–0.05 and increasing by one or two hundredths, with each value of $k$, reaching values between 0.04 and 0.08. Two exceptions to this trend are the best models: LSI and $k$-means with TF-IDF, which start at 0.01–0.02 for $k = 2$, quickly increase to vales 0.12–0.20 for $k = 10$, and then gradually increase to reach values 0.18–0.30. Another two exceptions are the worst models: Online Twitter LDA and $k$-means with Doc2Vec, which start at vales less than or equal to 0.01 and never grow larger than 0.02 for any value of $k$.

# 5 Discussion

## 5.1 Tweets dataset

A popular area of study is supervised algorithms using unbalanced datasets. However, skewed distributions also affect the learning process in unsupervised methods, especially in clustering [100] that are based on centroids [140,73]. Despite enormous solutions, there is a reduced effectiveness when the groups have highly different sizes [74], however, most of the models we used proved capable of creating a group of tweets bigger than the other that reflected the unbalanced nature of the tweets data set (94.6% and 5.4% of tweets related to HPV and lynch syndrome, respectively).

Online Twitter LDA followed by GSDMM obtained the highest values of SC and CH, which indicates that those clusterings were more compact, more dense (within the cluster), and better separated than all other models. In general, topic models out performed clustering methods in terms of CH and SC, which provides insights into the interconnected nature of medical communication. This problem is well suited for LDA proposed as improvements of LDA such as Online LDA.

For the evaluation of the external indices, a subset of tweets with hashtags was used. The external indices showed that LSI followed by *k*-means with TF-IDF obtained the best results. Note that Online Twitter LDA significantly decreased its performance in comparison to the values obtained in the internal indices evaluation. We did not find an obvious relationship between the number of iterations and the performance each of the different experimental configurations. In several cases, the performance is proportional to the number of iterations, although this is not a common pattern for all algorithms.

Several findings from the applications of topic modeling and clustering methods confirmed that many in society are using Twitter to share past and current experiences of a disease (HPV and lynch syndrome in our case), symptoms, treatment information, side effects, emotions, research, among others, as depicted in Table 2 and Figure 6. Figure 6 shows the most important topics extracted from two clusters created with *k*-means and Table 2 presents tweets extracted from these two clusters.

### 5.2 Emails dataset

In contrast to the tweets, the emails in our dataset characterize a smaller and more homogeneous domain of language. Each email was sent by a patient with prostate cancer (or their caregiver) to a health care provider. Careful consideration of the broader context of our modeling task can explain the findings of our experiments using emails as well as shed light on the results of our experiments with the tweets data.

The questions and concerns that arise as patients undergo treatment for prostate cancer, from scheduling procedures to managing a sudden crisis, are rarely discrete issues. This poses a substantial problem for clustering algorithms that search for perfectly separated clusters. *k*-means (with either embedding) is such an algorithm, which helps explain why it did not find internally meaningful clusters for any value of *k*. The nature of the emails may also help explain why the LSI model, which searches for a fixed low dimensional representation, generated internally inconsistent labels. In contrast, the LDA approach models document as a mixture of topics and seems to naturally represent an email that is primarily about family member introducing himself as the patient's new primary care giver while also mentioning several previously unreported health issues, for example. We see in our results that the three best performing models, with respect to internal indices, were a variation of LDA.

Despite the dramatic difference model performance over tweets and emails with respect to internal indices, we observed very similar patterns in performance with respect to external indices. Notably, LSI and *k*-means with TF-IDF, performed very well despite have mostly negative or near zero SC scores, respectively. BTM was always among the top 3 or 4 models while Online Twitter LDA had the best and worst measures of internal and external consistency, respectively in both experimental designs.

**Comparison to related studies:** Recent works compared topic modeling and clustering methods on short text clustering using the same external validity indices. In [117], GSDMM obtained the highest values, thus, one of the best suggested by external indices (NMI and Purity); while in our work GSDMM was suggested to be one of the best by internal indices (SC and CH) and the best for tweets only by an external index (C). In [117], GSDMM was

the best on 3 out of 6 datasets. NMI values for GSDMM varied from 0.3 to 0.8. Also, for a given partition, in several cases, GSDMM obtained the highest results in terms of NMI; while other external indices (e.g., Purity) obtained the highest results for another method.

In [33], *k*-means+Doc2Vec was suggested to be one of the best by external indices (NMI and ARI), while in our work it was suggested to be one of the worst with also external indices (NMI, ARI, H, C, and V). In [33], *k*-means+Doc2Vec was one of the best on the 3 datasets used. NMI and ARI values varied from 0.03–0.69 and 0.03–0.71 respectively. In [33] *k*-means+TF-IDF achieved the worst results, while in our work it was one of the best when evaluating external validity indices.

Both studies used small datasets ($\leq 30K$ docs). Both used a fixed number of topics for comparisons, since each dataset was already annotated. Both studies did not consider LSI (the best by external indices) or Online Twitter LDA (the best by internal indices).

Note that several studies showed that there is not a unique metric to validate clustering results [42,95,14], and the performance of metrics notably lowers with noise or overlapping clusters. Also, internal indices in comparison to external ones usually detect improvements in the clustering distribution which have positive implications in the system evaluation [66]

### 5.3 Error analysis

We also conducted an analysis of the types of error patterns found on short text clustering tasks. For this, we used the Hamming loss metric, which is a loss function, so the optimal value is zero (i.e., closer Hamming distance to the external classes and better performance) and its upper bound is one. Hamming loss measures the fraction of wrong labels to the total number of labels. Hamming Loss is relevant for an unbalanced classification tasks and relevant for multi-label classification. Thus, we computed the Hamming loss on the tweets and emails datasets. Note that this function depends on the labels of each document (tweet/email), thus, the interpretation of the output is very similar to the external validity indices.

For instance, Figure 7 illustrates the Hamming loss for emails for all models trained with 100 iterations. We found remarkably stable values of loss for all models for values of $k > 15$. The sable values of loss are consistent with the other external indices of validity: LSI and *k*-means with a TF-IDF representation being the first and second best performing models, and the variants of LDA along with GSDMM among the poorest performing models.

We then manually evaluated a group of tweets and emails to assess the assignment of clusters. There are several factors that caused errors when assigning emails or tweets to their respective clusters: (1) most of the tweets and emails contained misspellings, that received a part-of-speech category of “noun” or “unknown”, thus, considered for the clustering tasks; (2) tweets and emails contain terms created by patients such as “onco” instead of “oncology” which also affected the groups of texts; (3) most of tweets contain hashtags composed of two or more words, for instance “#hpvvaccine”; (4) lack of more context (e.g., semantic information), indeed, *n*-gram terms ($n \geq 2$) as features provide more context for clustering than a single word terms, for instance “hpv vaccination” and ”flu vaccination”

rather than “vaccination”; (5) the subjectivity to tell for a tweet/email to what cluster it belongs, the more number of cluster the more subjective become this task; among others.

Our study has known limitations. First, the annotation of both datasets was semi-automatically performed which directly affected the values of the external indices compared to the internal indices. We selected the most frequent hashtags (tweets) and words (emails), we then manually selected the most relevant hashtags/words and grouped them by their semantic similarity, finally, we executed a script that automatically annotated the tweets/emails for each number of clusters, (i.e., dataset labeled with two groups only when $k = 2$, then, dataset labeled with three groups only when $k = 3$, until $k = 50$). Second, for the external validation of the tweets dataset, we have only considered those containing hashtags which potentially affected the different results between internal and external indices. A possible solution could be to also consider visualization methods that can intuitively reflect the validity of clustering results.

## 6 Conclusions

In this paper, we conducted a detailed comparison of different topic modeling techniques and a document clustering method on short texts from two health-related datasets. The first composed of tweets and the second of emails. We set up LSI, LDA, GibbsLDA, Online LDA, BTM, Online Twitter LDA, GSDMM, and *k*-means based on TF-IDF and Doc2Vec document vectorizations. We evaluated our models with two internal indices and five external indices. The two internal indices included Calinski-Harabasz index and Silhouette Coefficient. Online Twitter LDA obtained the best results, which indicates it created more consistent clusters of topics for tweets and emails. The five external indices included Normalized Mutual Information, Adjusted Rand Index, V-measure, Homogeneity, and Completeness. These indices were evaluated using a ground truth dataset. Methods based on term and document frequencies such as LSI and *k*-means with TF-IDF obtained the best performance. Overall, this comparison provides encouraging results towards the application of topic modeling and clustering over short health-related texts from tweets and emails.

As a rapidly growing number of machine learning methods for natural language processing are becoming easier to implement for experts and novices alike, our study showed us that thoughtful analysis of language models along several dimensions is essential to know if one has arrived at a significant result. We observed notable variation in performance metrics attributable to sometimes subtle differences in model assumptions or computational methods alone. Moreover, we showed additional variation in performance when using data generated from a different process but within the same domain. For us, one or two cutoffs would not have given us sufficient information to evaluate model performance. Our work suggests researchers can improve their analysis of model performance by using a variety of metrics.

We provide this benchmark over different datasets to help other researchers determine whether their topic modeling and clustering methods are well suited to investigate healthcare questions such as: what health topics are most often discussed in tweets and email threads or what kinds of conversations are occurring between healthcare professionals and patients.

As future work, different other methods could be considered for evaluation, such as recent methods based on data augmentation and deep neural networks. Given the error patterns found in the clustering process, we also shall further investigate how to better leverage the selection of more informative features. For example, we shall include $n$-gram terms ($n \geq 2$) and adjectives as features for the methods. For the emails dataset, it would be interesting to consider the features of the people writing the emails: is the patient, the patient's family/caretaker, their age, how topics vary with course of disease. Finally, given the limitations associated with the dataset annotation, we shall annotate a subset of tweets and emails, compute the inter-annotator agreement, in order to manually assess the validity indices as well as the error within the clustering process.

## Acknowledgements

A portion of the research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA183962. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Abbreviations

| | |
|---|---|
| **ARI** | Adjusted Rand Index |
| **BOW** | Bag-of-Words |
| **BTM** | Biterm |
| **C** | Completeness |
| **CH** | Calinski-Harabasz |
| **FN** | False Negatives |
| **FP** | False Positives |
| **H** | Homogeneity |
| **HPV** | Human Papillomavirus |
| **LDA** | Latent Direchlet Allocation |
| **LSI** | Latent Semantic Indexing |
| **NLP** | Natural Language Processing |
| **NMI** | Normalized Mutual Information |
| **PHI** | Protected Health Information |
| **RI** | Rand Index |
| **SC** | and Silhouette Coefficient |
| **TF-IDF** | Term Frequency-Inverse Document Frequency |
| **TN** | True Negatives |

| | |
|---|---|
| **TP** | True Positives |
| **V** | V-measure |

## References


1. Biterm. https://github.com/xiaohuiyan/BTM. [Online; accessed December 15, 2019].
2. Doc2Vec. https://radimrehurek.com/gensim_3.8.3/models/doc2vec.html. [Online; accessed December 15, 2019].
3. GibbsLDA. https://nlp.stanford.edu/static/software/tmt/tmt-0.4/. [Online; accessed December 15, 2019].
4. K-means. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. [Online; accessed December 15, 2019].
5. LDA. https://radimrehurek.com/gensim_3.8.3/models/ldamodel.html. [Online; accessed December 15, 2019].
6. LSI. https://radimrehurek.com/gensim_3.8.3/models/lsimodel.html. [Online; accessed December 15, 2019].
7. Online LDA. https://radimrehurek.com/gensim_3.8.3/models/ldamulticore.html. [Online; accessed December 15, 2019].
8. Online Twitter LDA. https://github.com/jhlau/online_twitter_lda. [Online; accessed December 15, 2019].
9. TF-IDF. https://radimrehurek.com/gensim_3.8.3/models/tfidfmodel.html. [Online; accessed December 15, 2019].
10. Aggarwal CC and Zhai C. A survey of text clustering algorithms. In Mining text data, pages 77–128. Springer, 2012.
11. Amigó E, Gonzalo J, Artiles J, and Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retr., 12(4):461–486, Aug. 2009.
12. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral ecology, 26(1):32–46, 2001.
13. Antoun J. Electronic mail communication between physicians and patients: a review of challenges and opportunities. Family practice, 33(2):121–126, 2016. [PubMed: 26711957]
14. Arbelaitz O, Gurrutxaga I, Muguerza J, PéRez JM, and Perona I. An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1):243–256, 2013.
15. Arnold C and Speier W. A topic model of clinical reports. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ‘12, page 1031–1032, New York, NY, USA, 2012. Association for Computing Machinery.
16. Arnold CW, El-Saden SM, Bui AA, and Taira R. Clinical case-based retrieval using latent topic analysis. In AMIA annual symposium proceedings, volume 2010, page 26. American Medical Informatics Association, 2010. [PubMed: 21346934]
17. Aso T and Eguchi K. Predicting protein-protein relationships from literature using latent topics. In Genome Informatics 2009: Genome Informatics Series Vol. 23, pages 3–12. World Scientific, 2009. [PubMed: 20180257]
18. Atherton H, Sawmynaden P, Sheikh A, Majeed A, and Car J. Email for clinical communication between patients/caregivers and healthcare professionals. Cochrane Database of Systematic Reviews, (11), 2012.
19. Banerjee S, Ramanathan K, and Gupta A. Clustering short texts using wikipedia. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ‘07, page 787–788, New York, NY, USA, 2007. Association for Computing Machinery.
20. Bergmo TS, Kummervold PE, Gammon D, and Dahl LB. Electronic patient–provider communication: Will it offset office visits and telephone consultations in primary care? International journal of medical informatics, 74(9):705–710, 2005. [PubMed: 16095961]

21. Bicalho P, Pita M, Pedrosa G, Lacerda A, and Pappa GL. A general framework to expand short text for topic modeling. Information Sciences, 393:66–81, 2017.
22. Blei DM, Ng AY, and Jordan MI. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.
23. Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, and Hanson CL. Validating machine learning algorithms for twitter data against established measures of suicidality. JMIR mental health, 3(2):e21, 2016. [PubMed: 27185366]
24. Cai D, Mei Q, Han J, and Zhai C. Modeling hidden topics on document manifold. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, page 911–920, New York, NY, USA, 2008. Association for Computing Machinery.
25. Caliński T and Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1):1–27, 1974.
26. Cano AE, Varga A, Rowe M, Ciravegna F, and He Y. Harnessing linked knowledge sources for topic classification in social media. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT '13, page 41–50, New York, NY, USA, 2013. Association for Computing Machinery.
27. Chang F, Paramsothy T, Roche M, and Gupta NS. Patient, staff, and clinician perspectives on implementing electronic communications in an interdisciplinary rural family health practice. Primary health care research & development, 18(2):149–160, 2017. [PubMed: 27995826]
28. Chen JH, Goldstein MK, Asch SM, Mackey L, and Altman RB. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. Journal of the American Medical Informatics Association, 24(3):472–480, 2017. [PubMed: 27655861]
29. Chen W-Y, Song Y, Bai H, Lin C-J, and Chang EY. Parallel spectral clustering in distributed systems. IEEE transactions on pattern analysis and machine intelligence, 33(3):568–586, 2010.
30. Cheng X, Yan X, Lan Y, and Guo J. Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12):2928–2941, 2014.
31. Chu K-H, Unger JB, Allem J-P, Pattarroyo M, Soto D, Cruz TB, Yang H, Jiang L, and Yang CC. Diffusion of messages from an electronic cigarette brand to potential users through twitter. PloS one, 10(12):e0145387, 2015. [PubMed: 26684746]
32. Coppersmith G, Harman C, and Dredze M. Measuring post traumatic stress disorder in twitter. In Proceedings of the AAAI Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1–4, 2014., 2014.
33. Curiskis SA, Drake B, Osborn TR, and Kennedy PJ. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. Information Processing & Management, 57(2):102034, 2020.
34. Dai AM, Olah C, and Le QV. Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998, 2015.
35. Dai Z, Sun A, and Liu X-Y. Crest: Cluster-based representation enrichment for short text classification. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 256–267. Springer, 2013.
36. Dash J, Haller DM, Sommer J, and Perron NJ. Use of email, cell phone and text message between patients and primary-care physicians: cross-sectional study in a french-speaking part of switzerland. BMC health services research, 16(1):549, 2016. [PubMed: 27716256]
37. Davies DL and Bouldin DW. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell., 1(2):224–227, Feb. 1979. [PubMed: 21868852]
38. De Boom C, Van Canneyt S, Demeester T, and Dhoedt B. Representation learning for very short texts using weighted word embedding aggregation. Pattern Recogn. Lett., 80(C):150–156, Sept. 2016.
39. de Jong CC, Ros WJ, and Schrijvers G. The effects on health behavior and health outcomes of internet-based asynchronous communication between health providers and patients with a chronic condition: a systematic review. Journal of medical Internet research, 16(1):e19, 2014. [PubMed: 24434570]

40. De Martino I, D'Apolito R, McLawhorn AS, Fehring KA, Sculco PK, and Gasparini G. Social media for patients: benefits and drawbacks. Current reviews in musculoskeletal medicine, 10(1):141–145, 2017. [PubMed: 28110391]
41. Deerwester S, Dumais ST, Furnas GW, Landauer TK, and Harshman R. Indexing by latent semantic analysis. J. of the American society for information science, 41(6):391–407, 1990.
42. Dimitriadou E, Dolničar S, and Weingessel A. An examination of indexes for determining the number of clusters in binary data sets. Psychometrika, 67(1):137–159, 2002.
43. Duda RO, Hart PE, et al. Pattern classification and scene analysis, volume 3. Wiley New York, 1973.
44. Farhadloo M, Winneg K, Chan M-PS, Jamieson KH, and Albarracin D. Associations of topics of discussion on twitter with survey measures of attitudes, knowledge, and behaviors related to zika: probabilistic study in the united states. JMIR public health and surveillance, 4(1):e16, 2018. [PubMed: 29426815]
45. Fodeh S, Punch B, and Tan P-N. On ontology-driven document clustering using core semantic features. Knowledge Information System, 28(2):395–421, Aug. 2011.
46. Frey BJ and Dueck D. Clustering by passing messages between data points. science, 315(5814):972–976, 2007. [PubMed: 17218491]
47. Ganguly D and Ghosh K. Contextual word embedding: A case study in clustering tweets about emergency situations. In Companion Proceedings of the The Web Conference 2018, WWW '18, page 73–74, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.
48. Garrido T, Meng D, Wang JJ, Palen TE, and Kanter MH. Secure e-mailing between physicians and patients: transformational change in ambulatory care. The Journal of ambulatory care management, 37(3):211, 2014. [PubMed: 24887522]
49. Ghassemi M, Naumann T, Doshi-Velez F, Brimmer N, Joshi R, Rumshisky A, and Szolovits P. Unfolding physiological state: Mortality modelling in intensive care units. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, page 75–84, New York, NY, USA, 2014. Association for Computing Machinery.
50. Ghosh D and Guha R. What are we 'tweeting'about obesity? mapping tweets with topic modeling and geographic information system. Cartography and geographic information science, 40(2):90–102, 2013. [PubMed: 25126022]
51. Haase R, Schultheiss T, Kempcke R, Thomas K, and Ziemssen T. Use and acceptance of electronic communication by patients with multiple sclerosis: a multicenter questionnaire study. Journal of medical Internet research, 14(5):e135, 2012. [PubMed: 23069209]
52. Hadifar A, Sterckx L, Demeester T, and Develder C. A self-training approach for short text clustering. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 194–199, 2019.
53. Halkidi M, Batistakis Y, and Vazirgiannis M. Cluster validity methods: Part i. SIGMOD Rec., 31(2):40–45, June 2002.
54. Halkidi M and Vazirgiannis M. A density-based cluster validity approach using multi-representatives. Pattern Recognition Letters, 29(6):773–786, 2008.
55. Halko N, Martinsson P-G, and Tropp JA. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review, 53(2):217–288, 2011.
56. Hinton GE and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. science, 313(5786):504–507, 2006. [PubMed: 16873662]
57. Hoffman M, Bach F, and Blei D. Online learning for latent dirichlet allocation. advances in neural information processing systems, 23:856–864, 2010.
58. Hoogendoorn M, Berger T, Schulz A, Stolz T, and Szolovits P. Predicting social anxiety treatment outcome based on therapeutic email conversations. IEEE journal of biomedical and health informatics, 21(5):1449–1459, 2016. [PubMed: 27542187]
59. Hu X, Sun N, Zhang C, and Chua T-S. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In Proceedings of the 18th ACM Conference on Information

and Knowledge Management, CIKM '09, page 919–928, New York, NY, USA, 2009. Association for Computing Machinery.

60. Huang R, Yu G, Wang Z, Zhang J, and Shi L. Dirichlet process mixture model for document clustering with feature partition. IEEE Transactions on knowledge and data engineering, 25(8):1748–1759, 2012.
61. Huang X, Smith MC, Jamison AM, Broniatowski DA, Dredze M, Quinn SC, Cai J, and Paul MJ. Can online self-reports assist in real-time identification of influenza vaccination uptake? a cross-sectional study of influenza vaccine-related tweets in the usa, 2013–2017. BMJ open, 9(1):e024018, 2019.
62. Huang Z, Dong W, Duan H, and Li H. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. IEEE journal of biomedical and health informatics, 18(1):4–14, 2013.
63. Huang Z, Lu X, and Duan H. Latent treatment pattern discovery for clinical processes. Journal of medical systems, 37(2):9915, 2013. [PubMed: 23389419]
64. Hubert L and Arabie P. Comparing partitions. Journal of Classification, 2:193–218, 1985.
65. Ifrim G, Shi B, and Brigadir I. Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014. ACM, 2014.
66. Ingaramo D, Pinto D, Rosso P, and Errecalde M. Evaluation of internal validity measures in short-text corpora. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 555–567. Springer, 2008.
67. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, and Zhao L. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11):15169–15211, 2019.
68. Jiang Y, Liao QV, Cheng Q, Berlin RB, and Schatz BR. Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages. In AMIA Annual Symposium Proceedings, volume 2012, page 417. American Medical Informatics Association, 2012. [PubMed: 23304312]
69. Jin O, Liu NN, Zhao K, Yu Y, and Yang Q. Transferring topical knowledge from auxiliary long texts for short text clustering. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, page 775–784, New York, NY, USA, 2011. Association for Computing Machinery.
70. Karami A, Gangopadhyay A, Zhou B, and Karrazi H. Flatm: A fuzzy logic approach topic model for medical documents. In 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), pages 1–6. IEEE, 2015.
71. Karami A, Gangopadhyay A, Zhou B, and Kharrazi H. Fuzzy approach topic discovery in health and medical corpora. International Journal of Fuzzy Systems, 20(4):1334–1345, 2018.
72. Kiros R, Zhu Y, Salakhutdinov R, Zemel RS, Torralba A, Urtasun R, and Fidler S. Skip-thought vectors. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15, page 3294–3302, Cambridge, MA, USA, 2015. MIT Press.
73. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4):221–232, 2016.
74. Krawczyk B, Minku LL, Gama J, Stefanowski J, and Woźniak M. Ensemble learning for data stream analysis: A survey. Information Fusion, 37:132–156, 2017.
75. Kumar J, Shao J, Uddin S, and Ali W. An online semantic-enhanced Dirichlet model for short text stream clustering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 766–776, Online, July 2020. Association for Computational Linguistics.
76. Lau JH, Collier N, and Baldwin T. On-line trend analysis with topic models: twitter trends detection topic model online. In Proceedings of the 24th International Conference on Computational Linguistics, COLING '12, pages 1519–1534, 2012.

77. Le Q and Mikolov T. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14, page II–1188–II–1196. JMLR.org, 2014.

78. Lee DD and Seung HS. Algorithms for non-negative matrix factorization. In Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'00, page 535–541, Cambridge, MA, USA, 2000. MIT Press.

79. Li C, Duan Y, Wang H, Zhang Z, Sun A, and Ma Z. Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Trans. Inf. Syst., 36(2), Aug. 2017.

80. Li C, Wang H, Zhang Z, Sun A, and Ma Z. Topic modeling for short texts with auxiliary word embeddings. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, page 165–174, New York, NY, USA, 2016. Association for Computing Machinery.

81. Liu B, Liu L, Tsykin A, Goodall GJ, Green JE, Zhu M, Kim CH, and Li J. Identifying functional mirna–mrna regulatory modules with correspondence latent dirichlet allocation. Bioinformatics, 26(24):3105–3111, 2010. [PubMed: 20956247]

82. Lo SL, Chiong R, and Cornforth D. An unsupervised multilingual approach for online social media topic identification. Expert Systems with Applications, 81:282–298, 2017.

83. Lossio-Ventura JA and Bian J. An inside look at the opioid crisis over twitter. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1496–1499. IEEE, 2018.

84. Lossio-Ventura JA, Bian J, Jonquet C, Roche M, and Teisseire M. A novel framework for biomedical entity sense induction. Journal of biomedical informatics, 84:31–41, 2018. [PubMed: 29935347]

85. Lossio Ventura JA, Hacid H, Ansiaux A, and Maag ML. Conversations reconstruction in the social web. In Proceedings of the 21st International Conference on World Wide Web, WWW '12 Companion, pages 573–574, New York, NY, USA, 2012. ACM.

86. Lossio-Ventura JA, Hacid H, Roche M, and Poncelet P. Communication overload management through social interactions clustering. In Proceedings of the 31st Annual ACM Symposium on Applied Computing, SAC '16, page 1166–1169, New York, NY, USA, 2016. Association for Computing Machinery.

87. Lossio-Ventura JA, Morzan J, Alatrista-Salas H, Hernandez-Boussard T, and Bian J. Clustering and topic modeling over tweets: A comparison over a health dataset. In 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM'19. IEEE Computer Society, 2019 (in press).

88. Lu Y, Mei Q, and Zhai C. Investigating task performance of probabilistic topic models: An empirical study of plsa and lda. Information Retrieval, 14(2):178–203, Apr. 2011.

89. Lu Y, Zhang P, Liu J, Li J, and Deng S. Health-related hot topic detection in online communities using text clustering. Plos one, 8(2):e56221, 2013. [PubMed: 23457530]

90. Ma L, Wang Z, and Zhang Y. Extracting depression symptoms from social networks and web blogs via text mining. In International Symposium on Bioinformatics Research and Applications, pages 325–330. Springer, 2017.

91. MacQueen J et al. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA, 1967.

92. MacQueen JB. Some methods for classification and analysis of multivariate observations. In Cam LML and Neyman J, editors, Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press, 1967.

93. Manaskasemsak B, Chinthanet B, and Rungsawang A. Graph clustering-based emerging event detection from twitter data stream. In Proceedings of the Fifth International Conference on Network, Communication and Computing, ICNCC '16, page 37–41, New York, NY, USA, 2016. Association for Computing Machinery.

94. Manning CD, Schütze H, and Raghavan P. Introduction to information retrieval. Cambridge university press, 2008.

95. Maulik U and Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on pattern analysis and machine intelligence, 24(12):1650–1654, 2002.
96. Mikolov T, Sutskever I, Chen K, Corrado G, and Dean J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.
97. Millar JR, Peterson GL, and Mendenhall MJ. Document clustering and visualization with latent dirichlet allocation and self-organizing maps. In Twenty-Second International FLAIRS Conference, 2009.
98. Milligan GW and Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2):159–179, 1985.
99. Myslín M, Zhu S-H, Chapman W, and Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. Journal of medical Internet research, 15(8):e174, 2013. [PubMed: 23989137]
100. Nguwi Y-Y and Cho S-Y. An unsupervised self-organizing learning with support vector ranking for imbalanced datasets. Expert Systems with Applications, 37(12):8303–8312, 2010.
101. Nguyen DQ. jLDADMM: A Java package for the LDA and DMM topic models. arXiv preprint arXiv:1808.03835, 2018.
102. Nguyen DQ, Billingsley R, Du L, and Johnson M. Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 3:299–313, 2015.
103. Nigam K, McCallum AK, Thrun S, and Mitchell T. Text classification from labeled and unlabeled documents using em. Machine learning, 39(2–3):103–134, 2000.
104. Ofoghi B, Mann M, and Verspoor K. Towards early discovery of salient health threats: A social media emotion classification technique. In Biocomputing 2016: Proceedings of the Pacific Symposium, pages 504–515. World Scientific, 2016.
105. co/>Olariu A. Hierarchical clustering in improving microblog stream summarization. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 424–435. Springer, 2013.
106. Pappas Y, Atherton H, Sawmynaden P, and Car J. Email for clinical communication between healthcare professionals. Cochrane Database of Systematic Reviews, (9), 2012.
107. Paul MJ and Dredze M. Discovering health topics in social media using topic models. PloS one, 9(8), 2014.
108. Paul MJ and Dredze M. Social monitoring for public health. Synthesis Lectures on Information Concepts, Retrieval, and Services, 9(5):1–183, 2017.
109. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12:2825–2830, Nov. 2011.
110. Pennington J, Socher R, and Manning CD. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.
111. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, and Leenaars A. Suicide note classification using natural language processing: A content analysis. Biomedical informatics insights, 3:BII–S4706, 2010.
112. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, and Elhadad N. Learning probabilistic phenotypes from heterogeneous ehr data. Journal of biomedical informatics, 58:156–165, 2015. [PubMed: 26464024]
113. Prasad KR, Mohammed M, and Noorullah R. Visual topic models for healthcare data clustering. Evolutionary Intelligence, pages 1–18, 2019.
114. Prasad KR, Mohammed M, and Noorullah RM. Hybrid topic cluster models for social healthcare data. International Journal of Advanced Computer Science and Applications, 10(11), 2019.

115. Preo D¸tiuc-Pietro P. Srijith, Hepple M, and Cohn T. Studying the temporal dynamics of word co-occurrences: An application to event detection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4380–4387, 2016.
116. Qiang J, Li Y, Yuan Y, and Wu X. Short text clustering based on pitman-yor process mixture model. Applied Intelligence, 48(7):1802–1812, 2018.
117. Qiang J, Qian Z, Li Y, Yuan Y, and Wu X. Short text topic modeling techniques, applications, and performance: a survey. IEEE Transactions on Knowledge and Data Engineering, 2020.
118. Quan X, Kit C, Ge Y, and Pan SJ. Short and sparse text topic modeling via self-aggregation. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, page 2270–2276. AAAI Press, 2015.
119. Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336):846–850, 1971.
120. Rangrej A, Kulkarni S, and Tendulkar AV. Comparative study of clustering techniques for short text documents. In Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, page 111–112, New York, NY, USA, 2011. Association for Computing Machinery.
121. Rehurek R and Sojka P. Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. Citeseer, 2010.
122. Rosenberg A and Hirschberg J. V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
123. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65, 1987.
124. Rude S, Gortner E-M, and Pennebaker J. Language use of depressed and depression-vulnerable college students. Cognition & Emotion, 18(8):1121–1133, 2004.
125. Salton G and Buckley C. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5):513–523, 1988.
126. Sawmynaden P, Atherton H, Majeed A, and Car J. Email for the provision of information on disease prevention and health promotion. Cochrane Database of Systematic Reviews, (11), 2012.
127. Seneviratne MG, Seto T, Blayney DW, Brooks JD, and Hernandez-Boussard T. Architecture and implementation of a clinical research data warehouse for prostate cancer. eGEMs, 6(1), 2018.
128. Shou L, Wang Z, Chen K, and Chen G. Sumblr: continuous summarization of evolving tweet streams. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pages 533–542, 2013.
129. Shou L, Wang Z, Chen K, and Chen G. Sumblr: Continuous summarization of evolving tweet streams. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, page 533–542, New York, NY, USA, 2013. Association for Computing Machinery.
130. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, and Merchant RM. Twitter as a tool for health research: a systematic review. American J. of public health, 107(1):e1–e8, 2017.
131. Strehl A and Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research, 3(Dec):583–617, 2002.
132. Sulieman L, Gilmore D, French C, Cronin RM, Jackson GP, Russell M, and Fabbri D. Classifying patient portal messages using convolutional neural networks. Journal of biomedical informatics, 74:59–70, 2017. [PubMed: 28864104]
133. Sun A. Short text classification using very few words. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, page 1145–1146, New York, NY, USA, 2012. Association for Computing Machinery.
134. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, and Dunn AG. Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. Journal of Medical Internet Research, 18(8):e232, 2016. [PubMed: 27573910]

135. Tian F, Gao B, Cui Q, Chen E, and Liu T-Y. Learning deep representations for graph clustering. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI'14, page 1293–1299. AAAI Press, 2014.

136. Tibshirani R, Walther G, and Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2):411–423, 2001.

137. Van der Zanden R, Curie K, Van Londen M, Kramer J, Steen G, and Cuijpers P. Web-based depression treatment: Associations of clients ́word use with adherence and outcome. Journal of affective disorders, 160:10–13, 2014. [PubMed: 24709016]

138. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. Pharmacy and Therapeutics, 39(7):491, 2014. [PubMed: 25083128]

139. Vraga EK, Stefanidis A, Lamprianidis G, Croitoru A, Crooks AT, Delamater PL, Pfoser D, Radzikowski JR, and Jacobsen KH. Cancer and social media: A comparison of traffic about breast cancer, prostate cancer, and other reproductive cancers on twitter and instagram. Journal of health communication, 23(2):181–189, 2018. [PubMed: 29313761]

140. Wang Y and Chen L. Multi-exemplar based clustering for imbalanced data. In 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), pages 1068–1073. IEEE, 2014.

141. Wei T, Lu Y, Chang H, Zhou Q, and Bao X. A semantic approach for text clustering using wordnet and lexical chains. Expert Syst. Appl., 42(4):2264–2275, Mar. 2015.

142. Wei X and Croft WB. Lda-based document models for ad-hoc retrieval. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 178–185. ACM, 2006.

143. Wu Y, Liu M, Zheng WJ, Zhao Z, and Xu H. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. In Biocomputing 2012, pages 422–433. World Scientific, 2012.

144. Xie P and Xing EP. Integrating document clustering and topic modeling. In Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI'13, page 694–703, Arlington, Virginia, USA, 2013. AUAI Press.

145. Xu J, Xu B, Wang P, Zheng S, Tian G, and Zhao J. Self-taught convolutional neural networks for short text clustering. Neural Networks, 88:22–31, 2017. [PubMed: 28157556]

146. Xu T and Oard DW. Wikipedia-based topic clustering for microblogs. Proceedings of the American Society for Information Science and Technology, 48(1):1–10, 2011.

147. Yan X, Guo J, Lan Y, and Cheng X. A biterm topic model for short texts. In Proc of the 22nd Int Conference on World Wide Web, WWW '13, pages 1445–1456, New York, NY, USA, 2013. ACM.

148. Yin J, Chao D, Liu Z, Zhang W, Yu X, and Wang J. Model-based clustering of short text streams. In Proc of the 24th ACM SIGKDD Int Conference on Knowledge Discovery & Data Mining, KDD '18, pages 2634–2642, New York, NY, USA, 2018. ACM.

149. Yin J and Wang J. A dirichlet multinomial mixture model-based approach for short text clustering. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, page 233–242, New York, NY, USA, 2014. Association for Computing Machinery.

150. Yin J and Wang J. A model-based approach for text clustering with outlier detection. In 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pages 625–636. IEEE, 2016.

151. Yin J and Wang J. A text clustering algorithm using an online clustering scheme for initialization. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 1995–2004, New York, NY, USA, 2016. Association for Computing Machinery.

152. Žalik KR and Žalik B. Validity index for clusters of different sizes and densities. Pattern Recognition Letters, 32(2):221–234, 2011.

153. Zhang H, Wheldon C, Dunn AG, Tao C, Huo J, Zhang R, Prosperi M, Guo Y, and Bian J. Mining twitter to assess the determinants of health behavior toward human papillomavirus vaccination

in the united states. Journal of the American Medical Informatics Association, 27(2):225–235, 2020. [PubMed: 31711186]

154. Zhang L, Hall M, and Bastola D. Utilizing twitter data for analysis of chemotherapy. International journal of medical informatics, 120:92–100, 2018. [PubMed: 30409350]
155. Zhao Y, Guo Y, He X, Huo J, Wu Y, Yang X, and Bian J. Assessing mental health signals among sexual and gender minorities using twitter data. In 2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W), pages 51–52. IEEE, 2018.
156. Zheng CT, Liu C, and San Wong H. Corpus-based topic diffusion for short text clustering. Neurocomputing, 275:2444–2458, 2018.
157. Zuo Y, Wu J, Zhang H, Lin H, Wang F, Xu K, and Xiong H. Topic modeling of short texts: A pseudo-document view. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 2105–2114, New York, NY, USA, 2016. Association for Computing Machinery.
158. Zuo Y, Zhao J, and Xu K. Word network topic model: A simple but general solution for short and imbalanced texts. Knowl. Inf. Syst., 48(2):379–398, Aug. 2016.

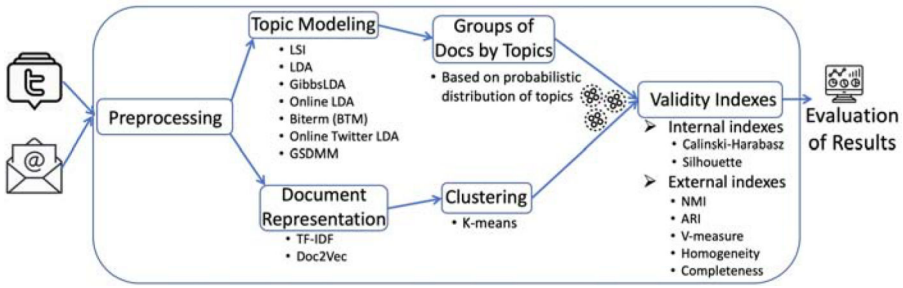


**Fig. 1.**

Workow of our approach to compare state-of-the-art topic modeling and clustering methods over health-related tweets and emails.

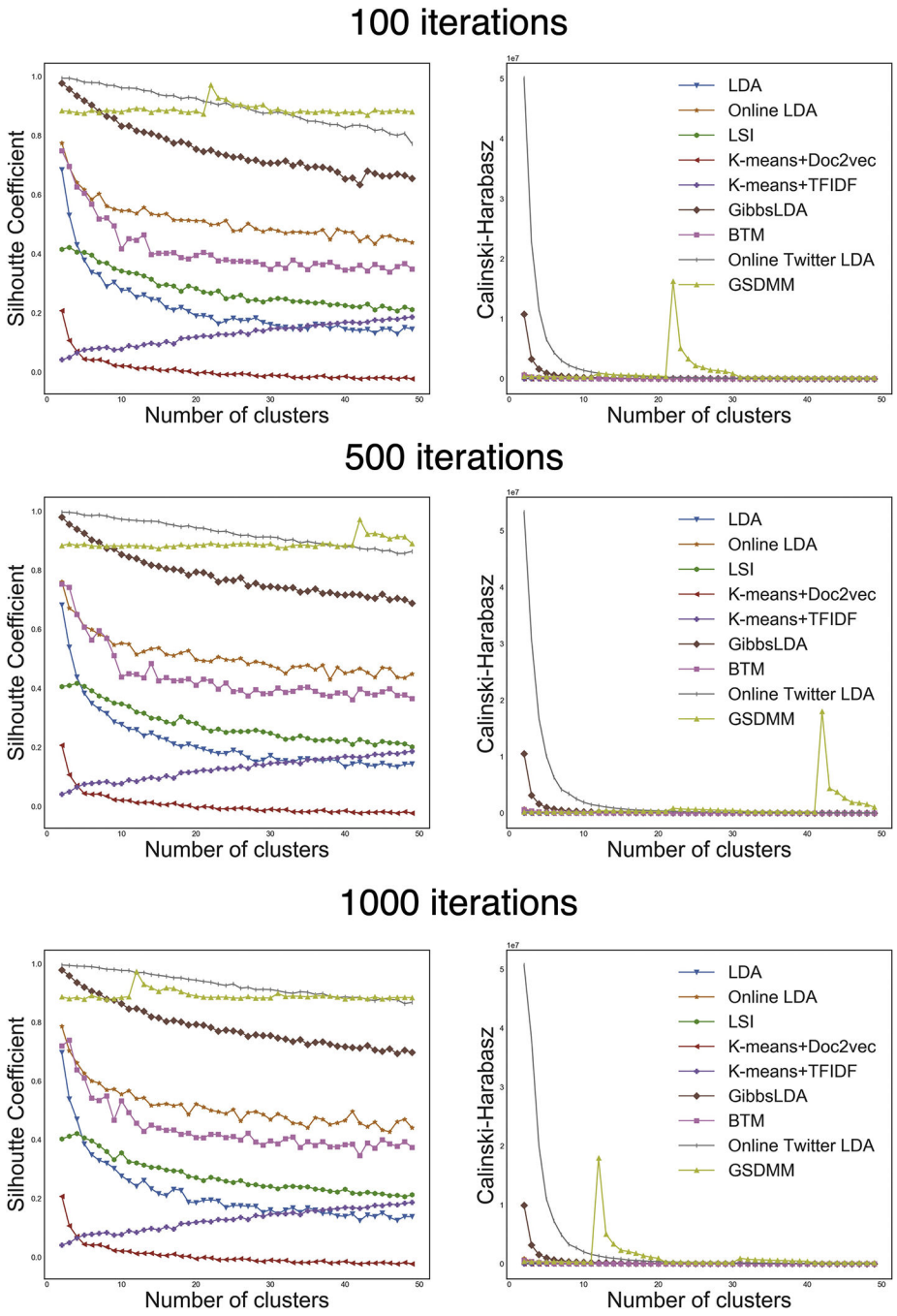


**Fig. 2.**

Silhouette Coefficient and Calinski-Harabasz metrics with 100, 500, and 1,000 iterations, for "*k*" ranging from 2 to 50.

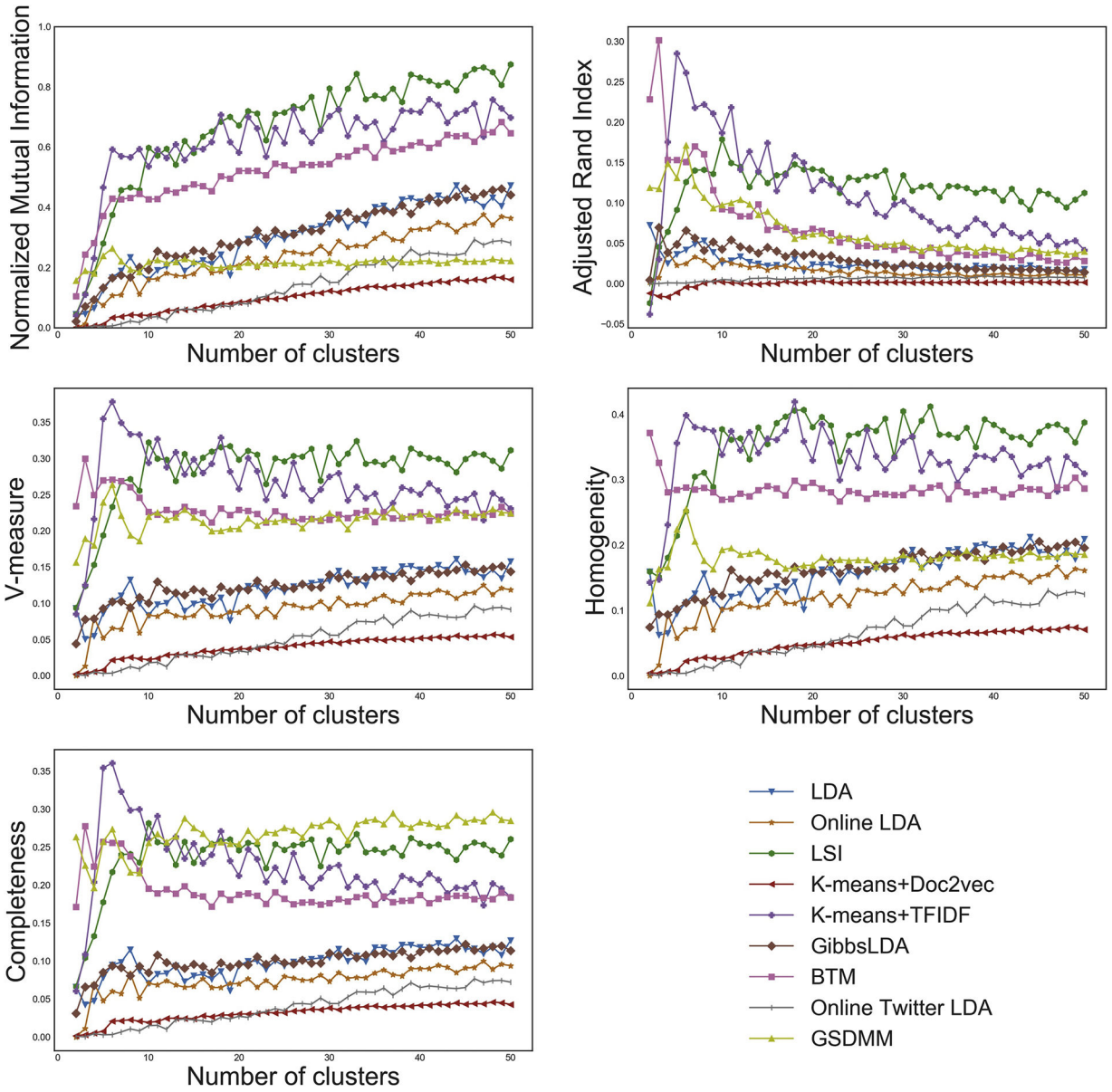


**Fig. 3.**

Normalized Mutual Information, Adjusted Rand Index, V-measure, Homogeneity, and Completeness metrics with 100 iterations for "*k*" ranging from 2 to 50 over the tweets dataset.

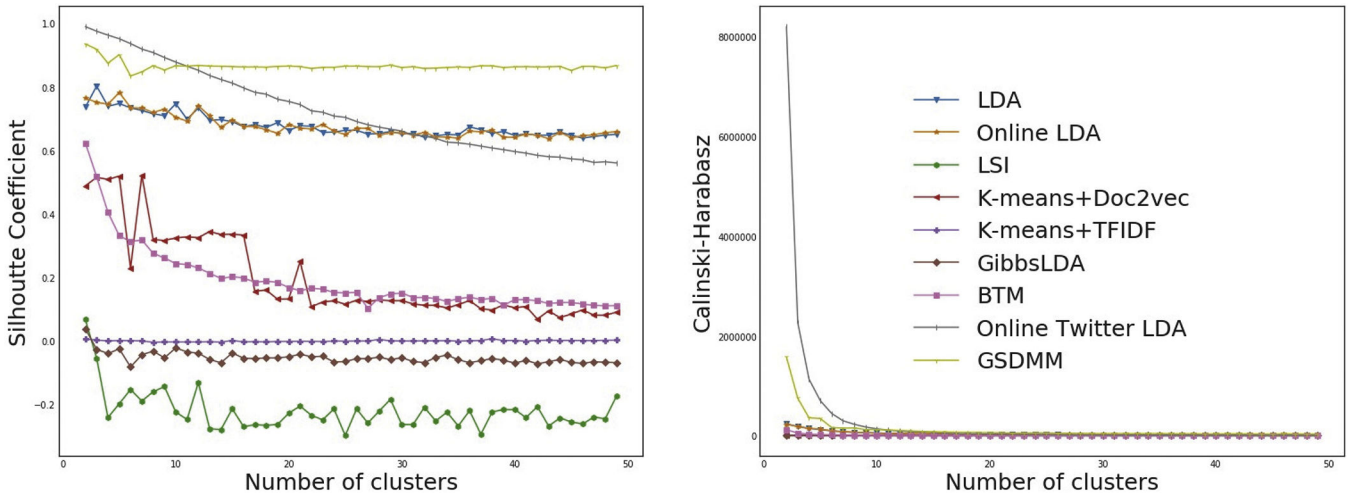


**Fig. 4.**

Silhouette Coefficient and Calinski-Harabasz metrics over the emails dataset with 100 iterations, for "*k*" ranging from 2 to 50.

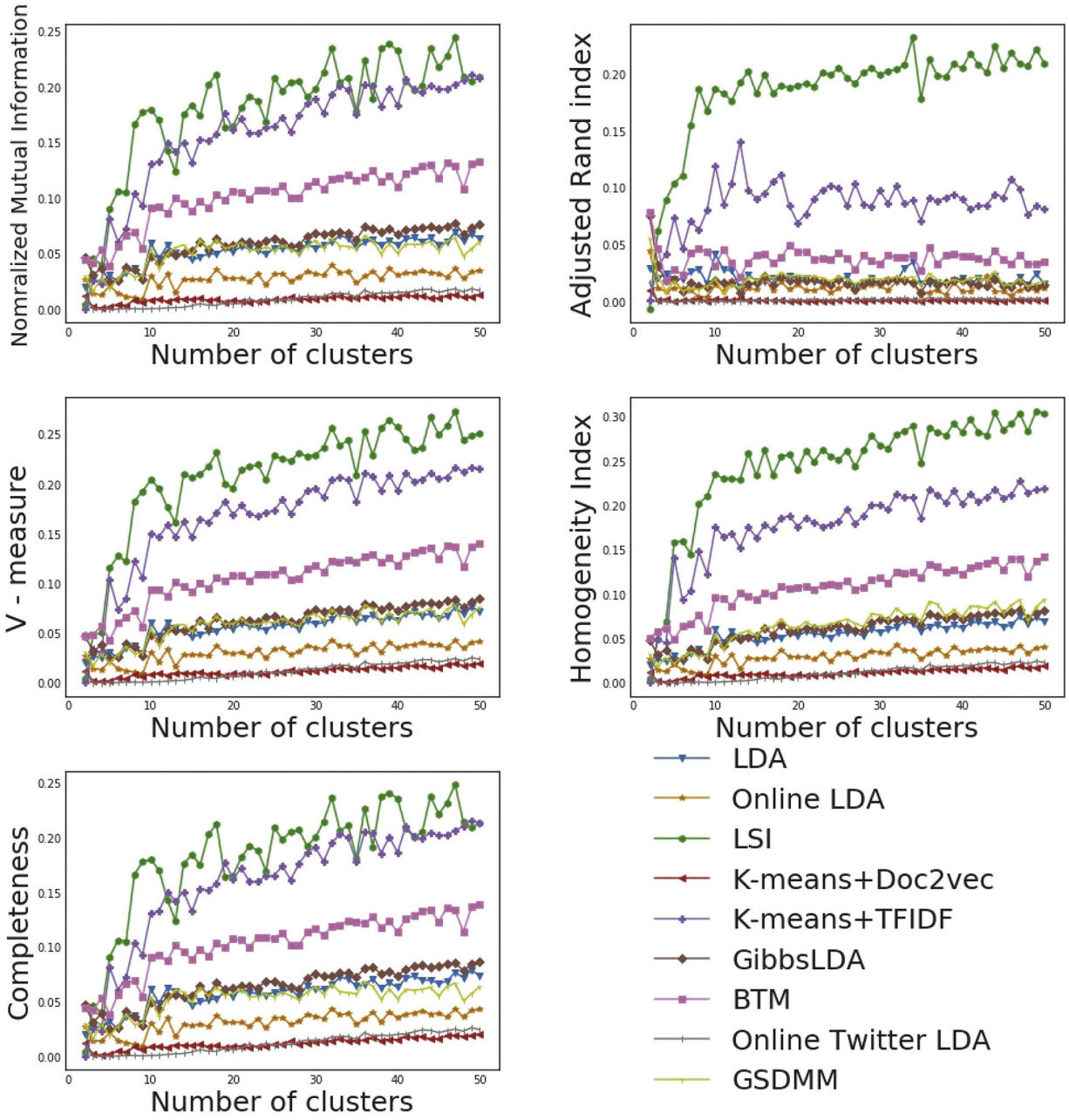


**Fig. 5.**

Normalized Mutual Information, Adjusted Rand Index, V-measure, Homogeneity, and Completeness indices with 100 iterations for "*k*" ranging from 2 to 50 over the email dataset.

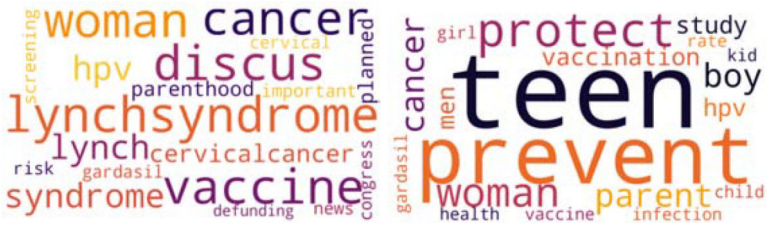


**Fig. 6.**

Most important topics extracted from two clusters created with *k*-means with TF-IDF.

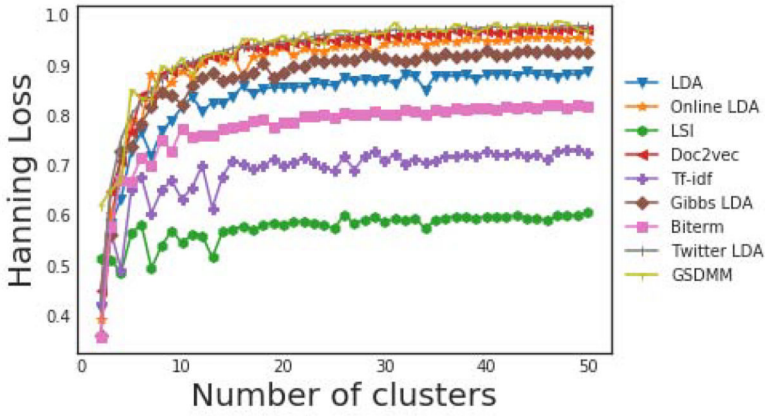


**Fig. 7.**

Hamming loss for topic modeling and clustering methods trained for 100 iterations and "*k*" ranging from 2 to 50 over emails dataset.

**Table 1**

Details of our health-related tweets dataset.

| Subset | HPV | Lynch syndrome |
|---|---|---|
| No. of tweets | 271,533 | 15,438 |
| No. of users | 99,227 | 4,492 |
| Collection period | Jan 2014 - Mar 2016 | Oct 2016 - Nov 2017 |
| No. of unique hashtags | 14,875 | 1,649 |
| No. of tweets with hashtag | 115,859 | 10,224 |
| No. of tokens before preprocessing | 1,767,920 | 147,144 |
| No. of tokens after preprocessing | 1,042,063 | 96,437 |
| Tokens per tweet after processing (mean ± SD) | 9.55 ± 3.85 | 8.86 ± 3.01 |

**Table 2**

Examples of tweets with HPV and lynch syndrome as contents in our tweets dataset.

| Dataset | Example of tweets |
|---|---|
| HPV | Continue to have incapacitating symptoms and remain unable to attend school or work |
| HPV | This is Cervical Health Awareness Month! Good news- HPV, the main cause of cervical cancer, is vaccine-preventable |
| HPV | You learn something new every day! Did you know that the cells of the cervix change with age in women? Immunity... |
| HPV | Earlier HPV vaccination may be more beneficial |
| Lynch | Research: Pain evaluation during gynaecological surveillance in women with Lynch syndrome |
| Lynch | #LynchSyndrome is an inherited condition which can predispose women to an increased risk of #endometrial #cancer #Lynchsyndr… |
| Lynch | Exact happened to my hubs– all symptoms kidney stone, no visual– get cultured urine test, test 4 Lynch Syndro… |
| Lynch | FDA apprvd Keytruda #immunotherapy- 1st drug to treat cancer based on #tumorgenetics |

**Table 3**

Top 10 groups of the most frequent hashtags from our tweets dataset.

| Top | Group of similar hashtags |
|---|---|
| 1 | #hpv, #hpvrelated, #hpvfacts, #humanpapillomavirus, #hpvassociated, #knowhpv |
| 2 | #lynch, #lynchsyndrome, #lynchsyndromeawareness, #lynchs |
| 3 | #vaccine, #vaccines, #vax, #vaxxed, #vaccination, #vaccinations |
| 4 | #cancer, #cancers |
| 5 | #cervical, #cervicalhealth, #cervicalcancer, #cervicalcancerawareness, #cervicalhealthawareness |
| 6 | #gardasil, #gardasil9, #gardasilvaccine |
| 7 | #health, #healthcare, #salud, #publichealth, #healthy |
| 8 | #learntherisk |
| 9 | #study |
| 10 | #cdc, #cdcwhistleblower |

**Table 4**

Details of our emails dataset.

| | |
|---|---|
| No. of emails | 50,000 |
| No. of patients | 4,535 |
| Collection period | Jan 2010 - Jan 2019 |
| No. of unique tokens before processing | 74,988 |
| No. of unique tokens after preprocessing | 42,868 |
| Tokens per email after preprocessing (mean ± SD) | 31.78 ± 32.72 |

**Table 5**

Top 10 groups of similar words of the most frequent topics in emails.

| Top | Groups of similar words |
|---|---|
| 1 | psa, test, results, scan, hba1c |
| 2 | surgery, prostate, urology |
| 3 | blood, pain, rash, nausea, symptoms, uti |
| 4 | prescription, mg, dose, ml, injection |
| 5 | obesity, nutritionist, weight |
| 6 | cancer, oncology, mass, masses |
| 7 | appointment, date, times |
| 8 | thank, thanks, dear, hello, hi, sincerely |
| 9 | germ, sick, infection |
| 10 | doctors, doctor, nurse |

**Table 6**

Internal index results for $k$=2.

| | 100 iterations | | 500 iterations | | 1,000 iterations | |
|---|---|---|---|---|---|---|
| | **CH** | **SC** | **CH** | **SC** | **CH** | **SC** |
| *LSI* | 86,260 | 0.41 | 86,260 | 0.40 | 86,260 | 0.40 |
| *BTM* | 634,412 | 0.74 | 661,242 | 0.75 | 604,629 | 0.72 |
| *LDA* | 515,737 | 0.68 | 486,522 | 0.68 | 521,198 | 0.69 |
| *GibbsLDA* | 10,767,060 | **0.97** | 10,514,730 | **0.98** | 9,932,722 | **0.97** |
| *Online LDA* | 849,068 | 0.77 | 834,351 | 0.76 | 938,428 | 0.78 |
| *Online Twitter LDA* | **50,110,500** | **0.99** | **53,291,260** | **0.99** | **50,730,260** | **0.99** |
| *GSDMM* | **16,232,540** | **0.97** | **18,030,810** | 0.97 | **17,961,280** | **0.97** |
| *k-means+Doc2Vec* | 31,196 | 0.20 | 31,196 | 0.20 | 31,196 | 0.20 |
| *k-means+TF-IDF* | 5,764 | 0.04 | 5,764 | 0.04 | 5,764 | 0.04 |

**Table 7**

Internal index results for $k$=5.

| | 100 iterations | | 500 iterations | | 1,000 iterations | |
|---|---|---|---|---|---|---|
| | **CH** | **SC** | **CH** | **SC** | **CH** | **SC** |
| *LSI* | 50,641 | 0.40 | 51,961 | 0.40 | 51,961 | 0.40 |
| *BTM* | 165,515 | 0.60 | 171,041 | 0.60 | 175,937 | 0.61 |
| *LDA* | 69,526 | 0.37 | 71,240 | 0.38 | 71,741 | 0.38 |
| *GibbsLDA* | 967,683 | **0.91** | 1,016,640 | **0.93** | 1,010,773 | **0.92** |
| *Online LDA* | 173,255 | 0.61 | 170,659 | 0.60 | 185,005 | 0.62 |
| *Online Twitter LDA* | **6,554,339** | **0.98** | **10,117,270** | **0.98** | **10,989,530** | **0.99** |
| *GSDMM* | **2,208,731** | **0.91** | **2,660,268** | 0.92 | **2,302,045** | 0.91 |
| *k-means+Doc2Vec* | 13,998 | 0.04 | 13,998 | 0.04 | 13,998 | 0.04 |
| *k-means+TF-IDF* | 4,722 | 0.07 | 4,722 | 0.07 | 4,722 | 0.07 |

**Table 8**

Internal index results for $k$=10.

| | 100 iterations | | 500 iterations | | 1,000 iterations | |
|---|---|---|---|---|---|---|
| | **CH** | **SC** | **CH** | **SC** | **CH** | **SC** |
| *LSI* | 25,346 | 0.34 | 25,539 | 0.34 | 25,532 | 0.35 |
| *BTM* | 55,110 | 0.41 | 61,790 | 0.43 | 67,856 | 0.53 |
| *LDA* | 24,498 | 0.27 | 24,102 | 0.27 | 24,860 | 0.27 |
| *GibbsLDA* | 239,457 | 0.83 | 266,065 | 0.85 | 272,035 | 0.86 |
| *Online LDA* | 70,824 | 0.54 | 69,418 | 0.55 | 69,892 | 0.55 |
| *Online Twitter LDA* | **1,400,045** | **0.96** | **1,925,903** | **0.97** | **2,035,547** | **0.97** |
| *GSDMM* | **878,294** | **0.88** | **957,856** | **0.89** | **955,849** | **0.89** |
| *k-means+Doc2Vec* | 7,617 | 0.02 | 7,617 | 0.02 | 7,617 | 0.02 |
| *k-means+TF-IDF* | 3,758 | 0.07 | 3,758 | 0.07 | 3,758 | 0.07 |

**Table 9**

Internal index results for $k$=50.

| | 100 iterations | | 500 iterations | | 1,000 iterations | |
|---|---|---|---|---|---|---|
| | **CH** | **SC** | **CH** | **SC** | **CH** | **SC** |
| *LSI* | 3,894 | 0.21 | 3,925 | 0.22 | 3,907 | 0.19 |
| *BTM* | 9,501 | 0.35 | 10,089 | 0.37 | 10,507 | 0.37 |
| *LDA* | 3,006 | 0.14 | 2,801 | 0.12 | 2,960 | 0.14 |
| *GibbsLDA* | 18,188 | 0.65 | 21,256 | 0.68 | 22,014 | 0.69 |
| *Online LDA* | 9,835 | 0.44 | 10,322 | 0.45 | 10,299 | 0.46 |
| *Online Twitter LDA* | **39,014** | **0.79** | **62,749** | **0.85** | **66,051** | **0.87** |
| *GSDMM* | **162,280** | **0.88** | **173,979** | **0.88** | **169,109** | **0.89** |
| *k-means+Doc2Vec* | 2,028 | −0.02 | 2,028 | −0.02 | 2,028 | −0.02 |
| *k-means+TF-IDF* | 2,200 | 0.17 | 2,200 | 0.17 | 2,200 | 0.17 |

**Table 10**

External index results for $k$=2.

| | 100 iterations | | | | | 500 iterations | | | | | 1,000 iterations | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **NMI** | **ARI** | **V** | **H** | **C** | **NMI** | **ARI** | **V** | **H** | **C** | **NMI** | **ARI** | **V** | **H** | **C** |
| *LSI* | 0.05 | −0.02 | 0.09 | **0.16** | 0.07 | 0.05 | −0.02 | 0.09 | **0.16** | 0.07 | 0.05 | −0.02 | 0.09 | 0.159 | 0.07 |
| *LDA* | 0.00 | 0.02 | 0.01 | 0.01 | 0.01 | 0.05 | 0.07 | 0.09 | **0.16** | 0.07 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| *GibbsLDA* | 0.02 | 0.01 | 0.03 | 0.05 | 0.02 | 0.02 | 0.01 | 0.04 | 0.08 | 0.03 | 0.02 | −0.00 | 0.04 | 0.07 | 0.03 |
| *Online LDA* | 0.03 | 0.06 | 0.07 | 0.12 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.06** | **0.09** | **0.12** | **0.21** | **0.09** |
| *BTM* | **0.11** | **0.23** | **0.24** | **0.38** | **0.17** | **0.11** | **0.23** | **0.23** | **0.37** | **0.17** | **0.09** | **0.14** | **0.18** | **0.30** | **0.13** |
| *Online Twitter LDA* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *GSDMM* | **0.17** | **0.13** | **0.17** | 0.12 | **0.28** | **0.16** | **0.12** | **0.16** | 0.11 | **0.26** | 0.05 | −0.06 | 0.05 | 0.04 | 0.08 |
| *k-means + Doc2Vec* | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 | 0.00 | −0.01 | 0.00 | 0.00 | 0.00 |
| *k-means + TF-IDF* | 0.04 | −0.04 | 0.09 | 0.14 | 0.06 | 0.04 | −0.04 | 0.09 | 0.14 | 0.06 | 0.04 | −0.04 | 0.09 | 0.14 | 0.06 |

**Table 11**

External index results for $k$=5.

| | 100 iterations | | | | | 500 iterations | | | | | 1,000 iterations | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **NMI** | **ARI** | **V** | **H** | **C** | **NMI** | **ARI** | **V** | **H** | **C** | **NMI** | **ARI** | **V** | **H** | **C** |
| *LSI* | 0.27 | 0.09 | 0.19 | 0.21 | 0.17 | 0.28 | 0.09 | 0.19 | 0.22 | 0.18 | 0.28 | 0.09 | 0.19 | 0.22 | 0.18 |
| *LDA* | 0.10 | 0.03 | 0.07 | 0.08 | 0.07 | 0.13 | 0.04 | 0.09 | 0.09 | 0.08 | 0.13 | 0.03 | 0.09 | 0.09 | 0.08 |
| *GibbsLDA* | 0.09 | 0.03 | 0.07 | 0.08 | 0.06 | 0.13 | 0.05 | 0.09 | 0.10 | 0.09 | 0.14 | 0.05 | 0.09 | 0.11 | 0.09 |
| *Online LDA* | 0.09 | 0.03 | 0.07 | 0.07 | 0.06 | 0.08 | 0.02 | 0.05 | 0.06 | 0.05 | 0.07 | 0.01 | 0.05 | 0.06 | 0.05 |
| *BTM* | **0.35** | **0.15** | **0.25** | **0.27** | **0.25** | **0.37** | **0.15** | **0.27** | **0.28** | **0.26** | **0.35** | **0.14** | **0.26** | **0.27** | **0.25** |
| *Online Twitter LDA* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 |
| *GSDMM* | 0.23 | 0.11 | 0.23 | 0.21 | **0.25** | 0.24 | 0.13 | 0.24 | 0.22 | **0.26** | 0.24 | 0.14 | 0.24 | 0.23 | **0.25** |
| *k-means + Doc2Vec* | 0.01 | −0.01 | 0.01 | 0.01 | 0.01 | 0.01 | −0.01 | 0.01 | 0.01 | 0.01 | 0.01 | −0.01 | 0.01 | 0.01 | 0.01 |
| *k-means + TF-IDF* | **0.47** | **0.29** | **0.36** | **0.36** | **0.35** | **0.47** | **0.29** | **0.36** | **0.36** | **0.35** | **0.47** | **0.29** | **0.36** | **0.36** | **0.35** |

**Table 12**

External index results for $k$=10.

| | 100 iterations | | | | | 500 iterations | | | | | 1,000 iterations | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **NMI** | **ARI** | **V** | **H** | **C** | **NMI** | **ARI** | **V** | **H** | **C** | **NMI** | **ARI** | **V** | **H** | **C** |
| *LSI* | **0.59** | **0.17** | **0.31** | **0.37** | **0.27** | **0.59** | **0.17** | **0.32** | **0.37** | **0.28** | **0.58** | **0.17** | **0.31** | **0.36** | **0.27** |
| *LDA* | 0.16 | 0.02 | 0.08 | 0.10 | 0.07 | 0.16 | 0.02 | 0.08 | 0.10 | 0.07 | 0.19 | 0.03 | 0.09 | 0.12 | 0.08 |
| *GibbsLDA* | 0.14 | 0.02 | 0.07 | 0.09 | 0.06 | 0.19 | 0.04 | 0.10 | 0.12 | 0.08 | 0.21 | 0.04 | 0.10 | 0.13 | 0.09 |
| *Online LDA* | 0.16 | 0.03 | 0.08 | 0.10 | 0.07 | 0.16 | 0.02 | 0.08 | 0.10 | 0.07 | 0.12 | 0.02 | 0.06 | 0.08 | 0.05 |
| *BTM* | 0.40 | 0.08 | 0.21 | 0.25 | 0.18 | 0.42 | 0.09 | 0.22 | 0.26 | 0.19 | 0.44 | 0.11 | 0.24 | 0.28 | 0.21 |
| *Online Twitter LDA* | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.03 | 0.00 | 0.01 | 0.02 | 0.01 | 0.04 | 0.00 | 0.02 | 0.02 | 0.01 |
| *GSDMM* | 0.24 | 0.12 | 0.24 | 0.21 | 0.27 | 0.22 | 0.10 | 0.22 | 0.19 | 0.26 | 0.22 | 0.12 | 0.22 | 0.20 | 0.25 |
| *k-means + Doc2Vec* | 0.04 | 0.00 | 0.02 | 0.02 | 0.01 | 0.04 | 0.00 | 0.02 | 0.02 | 0.01 | 0.04 | 0.00 | 0.02 | 0.02 | 0.01 |
| *k-means + TF-IDF* | **0.53** | **0.18** | **0.29** | **0.33** | **0.26** | **0.53** | **0.18** | **0.29** | **0.33** | **0.26** | **0.53** | **0.18** | **0.29** | **0.33** | **0.26** |

**Table 13**

External index results for $k$=50.

| | 100 iterations | | | | | 500 iterations | | | | | 1,000 iterations | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **NMI** | **ARI** | **V** | **H** | **C** | **NMI** | **ARI** | **V** | **H** | **C** | **NMI** | **ARI** | **V** | **H** | **C** |
| *LSI* | **0.87** | **0.10** | **0.30** | **0.38** | **0.25** | **0.87** | **0.11** | **0.31** | **0.38** | **0.26** | **0.84** | **0.09** | **0.29** | **0.37** | **0.24** |
| *LDA* | 0.40 | 0.01 | 0.13 | 0.17 | 0.10 | 0.47 | 0.01 | 0.15 | 0.20 | 0.12 | 0.44 | 0.01 | 0.15 | 0.19 | 0.12 |
| *GibbsLDA* | 0.40 | 0.01 | 0.13 | 0.17 | 0.10 | 0.44 | 0.01 | 0.14 | 0.19 | 0.11 | 0.45 | 0.01 | 0.14 | 0.20 | 0.11 |
| *Online LDA* | 0.36 | 0.01 | 0.11 | 0.16 | 0.09 | 0.36 | 0.01 | 0.11 | 0.16 | 0.09 | 0.34 | 0.00 | 0.11 | 0.15 | 0.09 |
| *BTM* | 0.62 | 0.03 | 0.21 | 0.27 | 0.17 | 0.64 | 0.02 | 0.22 | 0.28 | 0.18 | 0.63 | 0.02 | 0.21 | 0.28 | 0.17 |
| *Online Twitter LDA* | 0.19 | 0.00 | 0.06 | 0.08 | 0.04 | 0.28 | 0.00 | 0.09 | 0.12 | 0.07 | 0.34 | 0.00 | 0.11 | 0.15 | 0.08 |
| *GSDMM* | 0.22 | **0.04** | 0.22 | 0.18 | **0.28** | 0.22 | **0.04** | 0.22 | 0.19 | **0.28** | 0.23 | **0.04** | **0.23** | 0.19 | **0.30** |
| *k-means + Doc2Vec* | 0.15 | 0.00 | 0.05 | 0.07 | 0.04 | 0.15 | 0.00 | 0.05 | 0.07 | 0.04 | 0.15 | 0.00 | 0.05 | 0.07 | 0.04 |
| *k-means + TF-IDF* | **0.69** | **0.04** | **0.23** | **0.30** | 0.18 | **0.69** | **0.04** | **0.23** | **0.30** | 0.18 | **0.69** | **0.04** | **0.23** | **0.30** | 0.18 |

**Table 14**

Internal index results after 100 iterations.

| | k = 2 | | k = 5 | | k = 10 | | k = 50 | |
|---|---|---|---|---|---|---|---|---|
| | **CH** | **SC** | **CH** | **SC** | **CH** | **SC** | **CH** | **SC** |
| *LSI* | 1,316 | 0.06 | 529 | −0.19 | 696 | −0.22 | 517 | −0.24 |
| *BTM* | 121,814 | 0.62 | 19,884 | 0.33 | 9,200 | 0.24 | 1,173 | 0.11 |
| *LDA* | 237,684 | 0.74 | 118,955 | 0.75 | 54,848 | 0.75 | **8,062** | **0.65** |
| *GibbsLDA* | 59 | 0.04 | 1,128 | −0.02 | 706 | −0.02 | 103 | −0.07 |
| *Online LDA* | 222,296 | 0.76 | 128,584 | 0.78 | 47,653 | 0.70 | 7,974 | 0.64 |
| *Online Twitter LDA* | **8,197,541** | **0.99** | **699,231** | **0.95** | **136,077** | **0.88** | 5,719 | 0.56 |
| *GSDMM* | **1,576,441** | **0.94** | **342,109** | **0.90** | **122,952** | **0.87** | **27,889** | **0.86** |
| *k-means+Doc2Vec* | 13 | 0.48 | 13 | 0.52 | 21 | 0.32 | 77 | 0.09 |
| *k-means+TF-IDF* | 556 | 0.01 | 401 | 0.01 | 260 | −0.01 | 88 | 0.01 |

**Table 15**

External index results after 100 iterations.

| | ***k = 2*** | | | | | ***k = 5*** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *NMI* | *ARI* | *V* | *H* | *C* | *NMI* | *ARI* | *V* | *H* | *C* |
| *LSI* | 0.02 | 0.01 | 0.01 | 0.02 | 0.01 | **0.09** | **0.10** | **0.16** | **0.09** | **0.12** |
| *LDA* | 0.02 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 | 0.03 | 0.03 | 0.03 |
| *GibbsLDA* | **0.05** | **0.08** | **0.05** | **0.05** | **0.05** | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 |
| *Online LDA* | 0.03 | 0.05 | 0.03 | 0.03 | 0.03 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 |
| *BTM* | **0.04** | **0.08** | **0.05** | **0.04** | **0.05** | 0.04 | 0.03 | 0.05 | 0.04 | **0.04** |
| *Online Twitter LDA* | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| *GSDMM* | 0.03 | 0.06 | 0.03 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| *k-means + Doc2Vec* | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
| *k-means + TF-IDF* | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | **0.08** | **0.07** | **0.14** | **0.08** | 0.1 |

| | ***k = 10*** | | | | | ***k = 50*** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *NMI* | *ARI* | *V* | *H* | *C* | *NMI* | *ARI* | *V* | *H* | *C* |
| *LSI* | **0.18** | **0.19** | **0.24** | **0.18** | **0.20** | **0.21** | **0.21** | **0.30** | **0.21** | **0.25** |
| *LDA* | 0.06 | 0.04 | 0.06 | 0.06 | 0.06 | 0.06 | 0.02 | 0.07 | 0.07 | 0.07 |
| *GibbsLDA* | 0.05 | 0.02 | 0.05 | 0.05 | 0.05 | 0.08 | 0.02 | 0.08 | 0.09 | 0.08 |
| *Online LDA* | 0.03 | 0.01 | 0.03 | 0.03 | 0.03 | 0.03 | 0.01 | 0.04 | 0.04 | 0.04 |
| *BTM* | 0.09 | 0.03 | 0.10 | 0.09 | 0.09 | 0.13 | 0.04 | 0.14 | 0.14 | 0.14 |
| *Online Twitter LDA* | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | 0.02 | <0.01 | 0.02 | 0.02 | 0.02 |
| *GSDMM* | 0.05 | 0.02 | 0.05 | 0.05 | 0.05 | 0.02 | 0.06 | 0.09 | 0.06 | 0.08 |
| *k-means + Doc2Vec* | 0.01 | <0.01 | 0.01 | 0.01 | 0.01 | 0.01 | <0.01 | 0.02 | 0.02 | 0.02 |
| *k-means + TF-IDF* | **0.13** | **0.12** | **0.18** | **0.13** | **0.15** | **0.21** | **0.08** | **0.22** | **0.21** | **0.22** |