

# Visualization and performance measure to determine number of topics in twitter data clustering using hybrid topic modeling

R.M. Noorullah<sup>a,\*</sup> and Moulana Mohammed<sup>b</sup>

<sup>a</sup>*Department of CSE, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India*

<sup>b</sup>*Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India*

**Abstract.** Topic models are widely used in building clusters of documents for more than a decade, yet problems occurring in choosing the optimal number of topics. The main problem is the lack of a stable metric of the quality of topics obtained during the construction of topic models. The authors analyzed from previous works, most of the models used in determining the number of topics are non-parametric and the quality of topics determined by using perplexity and coherence measures and concluded that they are not applicable in solving this problem. In this paper, we used the parametric method, which is an extension of the traditional topic model with visual access tendency for visualization of the number of topics (clusters) to complement clustering and to choose the optimal number of topics based on results of cluster validity indices. Developed hybrid topic models are demonstrated with different Twitter datasets on various topics in obtaining the optimal number of topics and in measuring the quality of clusters. The experimental results showed that the Visual Non-negative Matrix Factorization (VNMf) topic model performs well in determining the optimal number of topics with interactive visualization and in performance measure of the quality of clusters with validity indices.

**Keywords:** Interactive visualization, visual non-negative matrix factorization model, an optimal number of topics, cluster validity indices, twitter data clustering

## 1. Introduction

Topic models are used for finding patterns of words in a document collection using statistics and machine learning techniques. Especially in the big data era, topic models are effective means in text mining and knowledge discovery. Topic modeling [23] is an effective means of data mining in machine learning to build models from unstructured textual data, where samples are treated as documents. Huge collection of documents organized by topic modeling or by clustering. In topic modeling, the topic is defined by a cluster

of words with the probability of occurrence of each word, and different topics have their respective cluster of words with corresponding probability. In clustering, the basic idea is to group documents into different groups based on some suitability measures. To perform grouping, each document is represented by a vector representing the weight assigned by tf-idf to words in documents. Topic modeling can project documents into a topic space which facilitates effective document clustering. Clustering helps us to identify the latent groups in a document collection and subsequently we can identify local and global topics. Researchers have applied topic models to cluster documents. Latent Dirichlet Allocation (LDA) [2] most widely used method in topic modeling for text mining that is based on an unsupervised statistical topic

\*Corresponding author. R.M. Noorullah, Department of CSE, Institute of Aeronautical Engineering, Dundigal, Hyderabad 500043, Telangana, India. E-mail: noorullah.rm@iaere.ac.in.

model where topics are fixed for the whole corpus and the number of topics is assumed to be known. As an unsupervised statistical model, LDA discovers underlying topics in unlabeled data. To find the optimal number of topics  $k$ , the LDA model applied for different values of  $k$  and pick the one which gives the highest coherence value. Topic coherence [19] is a relative measure to find the number of topics produced by a particular topic model. Usually, the perplexity measure used, which is an intrinsic evaluation metric [28] and [7] measured as the normalized log-likelihood. Recent studies showed that perplexity is not correlated, sometimes anti-correlated, and cannot compare uni-gram and  $n$ -gram models, so topic coherence is used to find the optimal number of topics. Hierarchical Dirichlet Process (HDP) [27] is an extension of LDA, which is a Bayesian non-parametric model used to solve in finding the number of topics for the whole collection but not for a specific document, in which the number of topics need not be specified in advance and determined by collection during posterior inference. If the corpus is large, then take a uniform sample and run HDP-LDA. The hLDA model [3], and [4] regardless of data consider the number of topics as a hyperparameter. Detection of the number of topics is complemented with interactive visualization. In previous works, t-SNE [16], is a variation of SNE that is easier and to optimize and visual representation and produces significantly better visualizations by reducing the tendency to crowd points, iVis Clustering method [13] provide a two-dimensional plot that visualizes cluster similarities and graph-based representation relationship among data items so that user can interact with them. The UTOPIAN method [15] and [13], enables users to interact with the topic modeling method and steer the result in a user-driven manner, LDAVis [6] model is a web-based interactive visualization of topics estimated using LDA, which allows users to explore the topic-term relationship. SenLDA [12] is an extension of LDA to overcome the limitation in the generative and inference processes in finding the number of topics. In pyLDA [25] which combines python library for creation and visualization of the topic model to focus on the analysis of modeling results and in word cloud [26], and [24] used to visualize topics and to evaluate the quality of derived documents. In this big data era, the amount of online document data is growing with high velocity due to the widespread and availability of social media for sharing views and discussions of users, the large corpus is dealt in topic modeling in which finding number of topics and

measuring their performance are having limitations of the proposed methods. The study of stable metrics for measuring the quality of topics and the use of cluster analysis tools for analysis and in finding the optimal number of topics continues. This motivated us to develop the hybrid topic models in which the optimal number of topics will be visually represented in word cloud form of different keywords used in the dataset and to represent the optimal number of topics in the form of Visual Assessment Tendency images and used validity indices instead of perplexity or coherence as performance measures in assessing the optimal number of topics. The rest of the paper is organized as follows: Section 2 presents the methodology, section 3 deals with experimental evaluation and discussion, and section 4 presents the conclusion and future enhancement of the work.

## 2. Methodology

### 2.1. Datasets description

For the experiment, the datasets collected from Twitter on 20 topics of health-related documents, TREC2014, TREC2015 Keyword Phrases Tweets collected from Twitter as described in [21] and Tweets extracted from Twitter related to 25 keyword phrases of TREC2018 [10] as described in Table 1 and 24KeywordPhrases of Measures to decrease COVID-19 Spread Dataset as in Table 2 are used. Experiments are implemented with Intel core i7 processor@3.4 GHz, 8MB cache, 16GB RAM, 1TB HDD, Python IDLE 3.9 with 64bit, NLP for text preprocessing, sklearn, Numpy, Pandas, and cvindex packages used. In Table 1, Tweets collected from Twitter on Measures to decrease COVID-19 spread dataset covers five themes i.e. decreasing the spread of COVID-19 theme with six keyword phrases, the second theme covers with next five keyword phrases, the third theme covers availability of pharmaceutical interventions with next five keyword phrases, fourth theme prevention, and control measures in schools with four keyword phrases and the fifth theme covers training measures to reduce community spread with rest of four keyword phrases.

### 2.2. Process to find the optimal number of topics

In [8] and [11] instability of topics for the order of process, the document was discovered and presented in the paper. To eliminate the order of dependence of

Table 1  
Dataset keyword phrases description and size

Keyword Phrases	TREC2018 Dataset		Measures to decrease COVID-19 spread Dataset	
	Keyword Phrase Description	Size (No. of Documents)	Keyword Phrase Description	Size (No. of Documents)
2KeywordPhrases	Women in Parliaments, Black Bear Attacks	80	Wearing Mask, and Quarantine	150
3KeywordPhrases	2Keyword Phrases, and Airport Security	120	2Keyword Phrases, and Face Shield and Personal Protective Equipment	200
4KeywordPhrases	3Keyword Phrases, and Wildlife Extinction	160	3Keyword Phrases, and Environmental Measure	210
5KeywordPhrases	4Keyword Phrases, and Health and Computer Terminals	200	4Keyword Phrases, and Social/Physical Distancing	260
6KeywordPhrases	5Keyword Phrases, and human smuggling	240	5Keyword Phrases, and Hand Hygiene/Hand Washing	285
7KeywordPhrases	6Keyword Phrases, and transportation tunnel disasters	280	6Keyword Phrases, and Lockdown1.0 in India	310
8KeywordPhrases	7Keyword Phrases, and piracy	320	7Keyword Phrases, and Lockdown2.0 in India	340
9KeywordPhrases	8Keyword Phrases, and hydrogen energy	360	8Keyword Phrases, and Lockdown3.0 in India	355
10KeywordPhrases	9Keyword Phrases, and euro opposition	400	9Keyword Phrases, and Lockdown4.0 in India	370
11KeywordPhrases	10Keyword Phrases, and mercy killing	440	10Keyword Phrases, and Lockdown5.0 in India	420
12KeywordPhrases	11Keyword Phrases, and tropical storms	480	11Keyword Phrases, and Vaccine in progress	470
13KeywordPhrases	12Keyword Phrases, and women clergy	520	12Keyword Phrases, and Hydroxychloroquine	500
14KeywordPhrases	13Keyword Phrases, and college education advantage	560	13Keyword Phrases, and Remdesivir	550
15KeywordPhrases	14Keyword Phrases, and women driving in Saudi Arabia	600	14Keyword Phrases, and Avigan	590
16KeywordPhrases	15Keyword Phrases, and eating invasive species	640	15Keyword Phrases, and Drug evaluation	595
17KeywordPhrases	16Keyword Phrases, and protect Earth from asteroids	680	16Keyword Phrases, and Prevention and control in schools	615
18KeywordPhrases	17Keyword Phrases, and diabetes, and toxic chemicals	720	17Keyword Phrases, and Education policies	640
19KeywordPhrases	18Keyword Phrases, and car hacking	760	18Keyword Phrases, and school resources	670
20KeywordPhrases	19Keyword Phrases, and social media and teen suicide	800	19Keyword Phrases, and safety in schools	690
21KeywordPhrases	20Keyword Phrases, and federal minimum wage increase	840	20Keyword Phrases, and National coordination	710
22KeywordPhrases	21Keyword Phrases, and eggs in a healthy diet	880	21Keyword Phrases, and Logistics	740
23KeywordPhrases	22Keyword Phrases, and email scams	920	22Keyword Phrases, and essential services	770
24KeywordPhrases	23Keyword Phrases, and food prices	960	23Keyword Phrases, and Research	800
25KeywordPhrases	24Keyword Phrases, and bacterial infection mortality rate	1000	—	—

documents in a corpus, to find the optimal number of topics, and to measure the quality of metrics the following procedure is adopted and explained in the following process diagram. Extracted Tweets docu-

ments from Twitter based on topics mentioned above are preprocessed with the NLP tool. Derived topic-document matrix and number of topics are treated as input to partition matrix. Under Euclidean distance,

hybrid topic models [22] are applied for word cloud visualization and visual access tendency images in determining the optimal number of topics. The confusion matrix for this specified number of topics is displayed. Based on the values in the confusion matrix validity of the optimal number of topics is performed with validity indices and results are stored in the database. The process is repeated for the same value of  $k$  and treating  $k = 2$  to  $k = 25$ , and the process is repeated for all datasets of 20 topics of health-related topics and 25 topics of TREC2018 keyword phrases. Results are stored in a database and graphs are generated for five validity indices and the optimal number of topics are selected based on cluster validity indices results.

### Procedure for optimal number of topic selection

1. Extract tweets from Twitter based on keyword phrases or topics on different datasets.
2. Preprocess extracted tweets using the NLP tool.
3. Assign  $k$  value for a particular dataset
4. Apply hybrid topics models for clustering by using the Euclidean distance metric.
5. Confusion Matrix developed for different hybrid topic models under distance metric.
6. VAT images represented for different topics of a specific dataset.
7. Find an optimal number of clusters using validity indices on a specific dataset.
8. Repeat step 3 to step 7 for other specified datasets

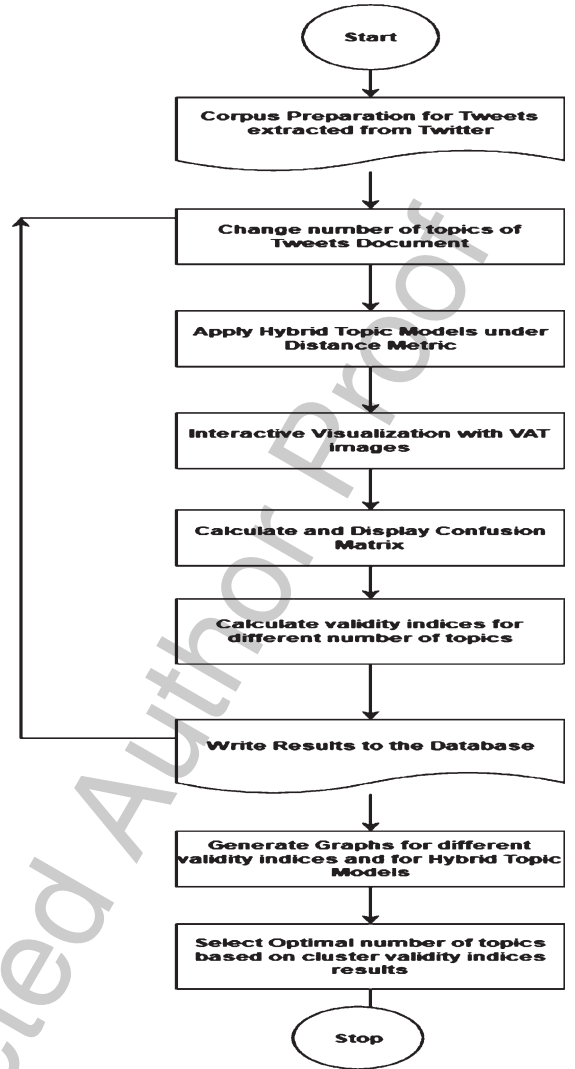


Fig. 1. Process diagram to find the optimal number of topics.

## 3. Experimental evaluation and discussion

### 3.1. Formation of word clouds for different topics

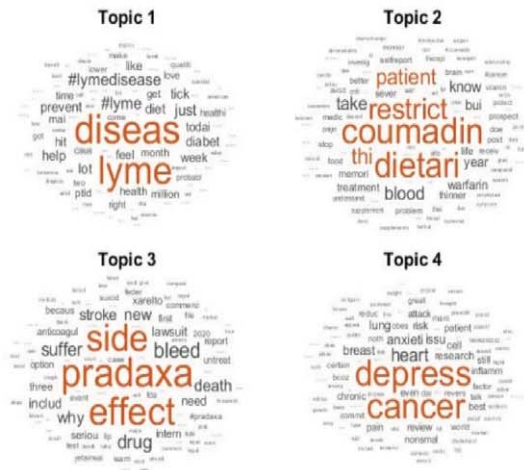
Document topic modeling is used to generate topics and word clouds from a large collection of textual information. In [14] MALLET topic modeling tool is used to generate topics and to give a list of topics in percentage. In this paper word clouds, the tool is used to describe results and trending keywords with the visual representation of word contents commonly used in the topic cluster with each word's frequency correlated with font size. Word clouds are generated for all datasets as described above. Few samples are shown in Fig. 2 in which four keyword phrases of TREC2014 and TREC2015 and two keyword phrases of TREC2014 and TREC2015 with word intensity are

displayed, which helps users to quickly evaluate the number of keywords used.

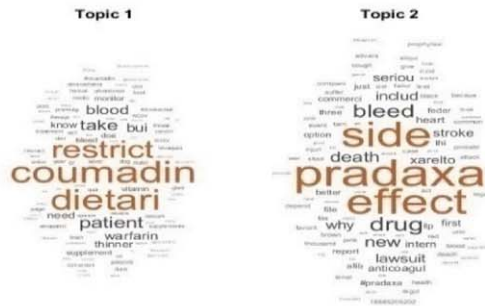
### 3.2. Assessment of number of topics with visual assessment tendency images

#### Steps for Tweets Document Clustering:

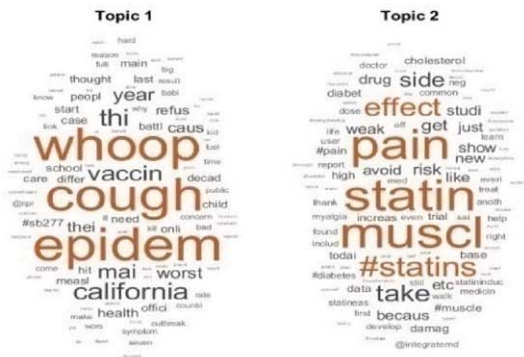
- 1: Preprocess the extracted tweet documents of different datasets to remove punctuations and special symbols, remove short words, tokenization, and stemming.
- 2: Determine the feature extraction of tweet documents using four traditional topic models NMF, LDA, LSI, and PLSI.



(a) TREC2015 Four Keyword Phrases Word Cloud



(b) TREC2015 Two Keyword Phrases Word Cloud



(c) TREC2014 Keyword Phrases Word Cloud

Fig. 2. Word clouds of twitter datasets.

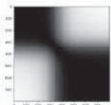
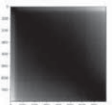
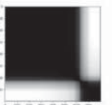
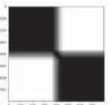
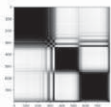
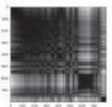
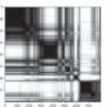
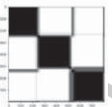
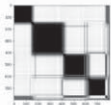
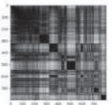
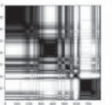
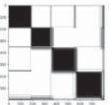
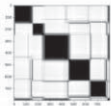
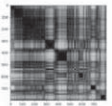
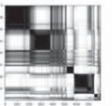
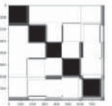
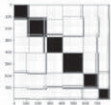
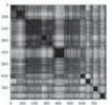
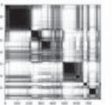
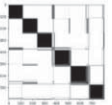
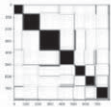
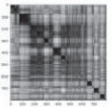
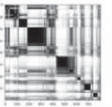
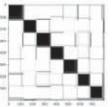
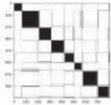
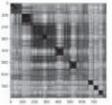
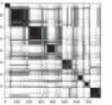
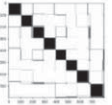
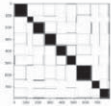
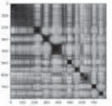
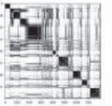
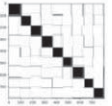
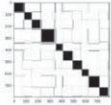
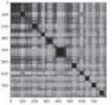
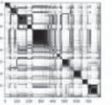
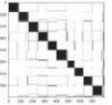
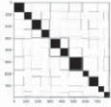
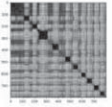
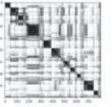
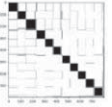
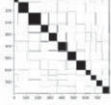
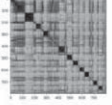
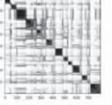
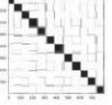
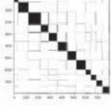
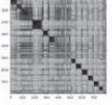
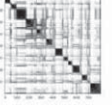
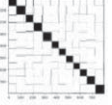
- [illegible]

Visual assessment of the optimal number of topics is shown in Fig. 3 for TREC2018 twenty keyword phrases for proposed hybrid topic models viz., Visual LDA (VLDA), Visual LSI (VLSI), Visual NMF (VNMF), and Visual PLSI (VPLSI) models under Euclidean distance metric as a sample. We have generated VAT images for all datasets of various topics used in Twitter Health Topics datasets, TREC2014, TREC2015, TREC2018, and Measures to decrease COVID-19 spread keyword phrases datasets of tweets collected from Twitter. From the above VAT images, every individual topic is represented as dark square-shaped blocks along diagonal. The quality and number are recognized with more quality of shape in the VAT image. We experimented by taking a sample total number of tweets based on the number of topics and repeated by changing values of  $k$  from 2 to 25 and all generated images for all four hybrid topic models are stored. In the above figure, a sample of TREC2018 for 20 keyword phrases dataset is considered and VAT images are presented for four models by varying  $k$  values from 2 to 25. From the visual evidence of Fig. 3, it is observed as  $k$  values vary overlapping of tweets occurring which is visual evidence by enhancing VAT images. Among these models, Visual LSI and Visual NMF are generating good results in identifying the optimal number of topics when compared to Visual LDA and Visual PLSI models. Visual NMF best performed in finding the optimal number of topics not only in Visual representation and also in measuring quality by using cluster validity indices, which is shown in the ensuing section.

### 3.3. Performance measure and finding number of topics using validity indices

In [18] and [5] qualitative evaluation of topic-author and topic-word are evaluated in terms of perplexity, although it may generate meaningful results in some cases, it is not stable and results vary with selected seeds for the same dataset. Recent studies showed that perplexity is not correlated and

- 3: Features of tweet documents are labeled as  $D_1, D_2 \dots D_n$  for the total  $n$  documents in each dataset.
- 4: Compute the distance using Euclidean and update values of  $D_1, D_2 \dots D_n$ .
- 5: Dissimilarity matrix of tweet documents is computed.

K Value	Visual LDA	Visual LSI	Visual PLSI	Visual NMF
K=2				
K=3				
K=4				
K=5				
K=6				
K=7				
K=8				
K=9				
K=10				
K=11				
K=12				
K=13				



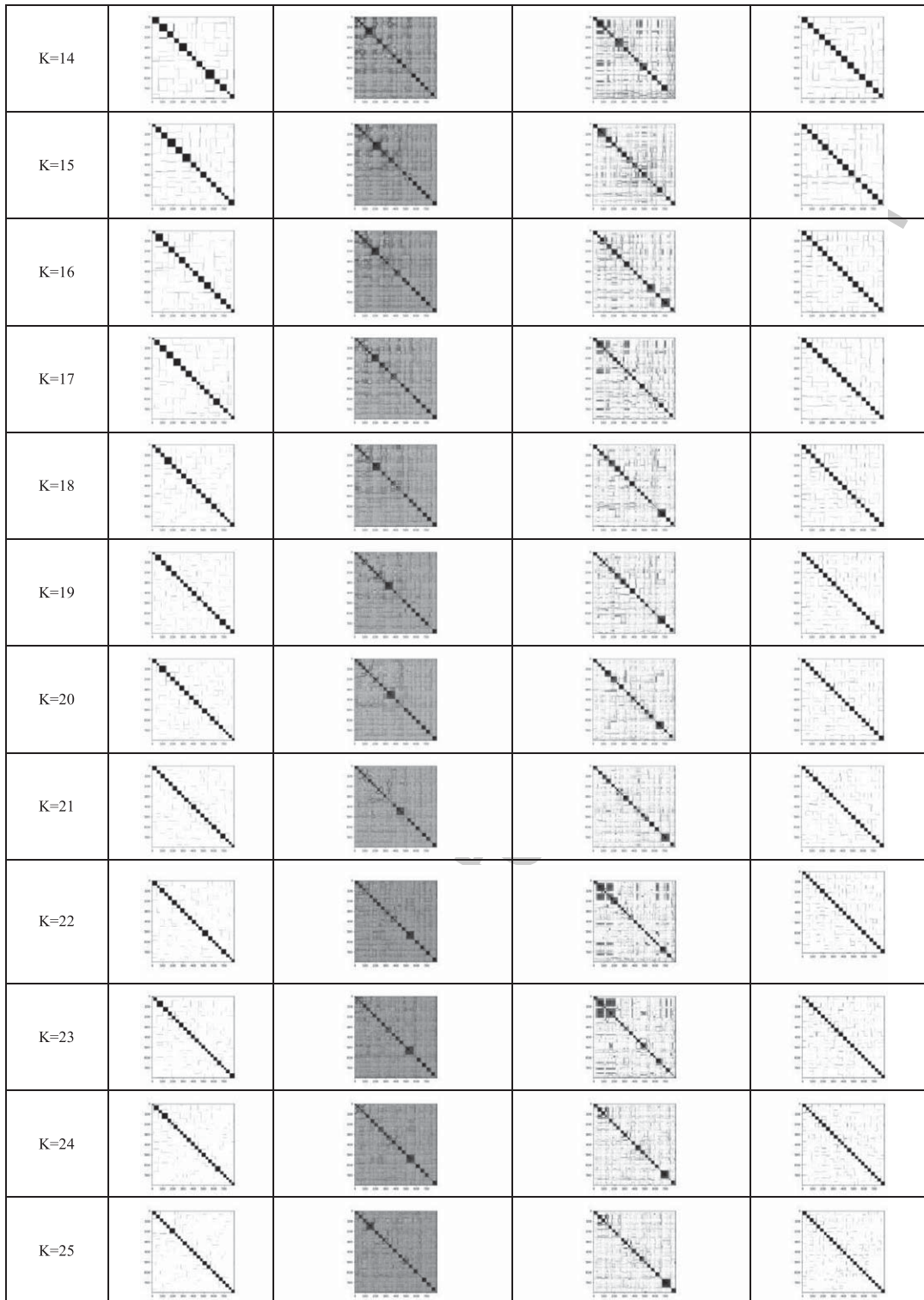


Fig. 3. Assessment of the number of topics with VAT images using hybrid topic modeling techniques.

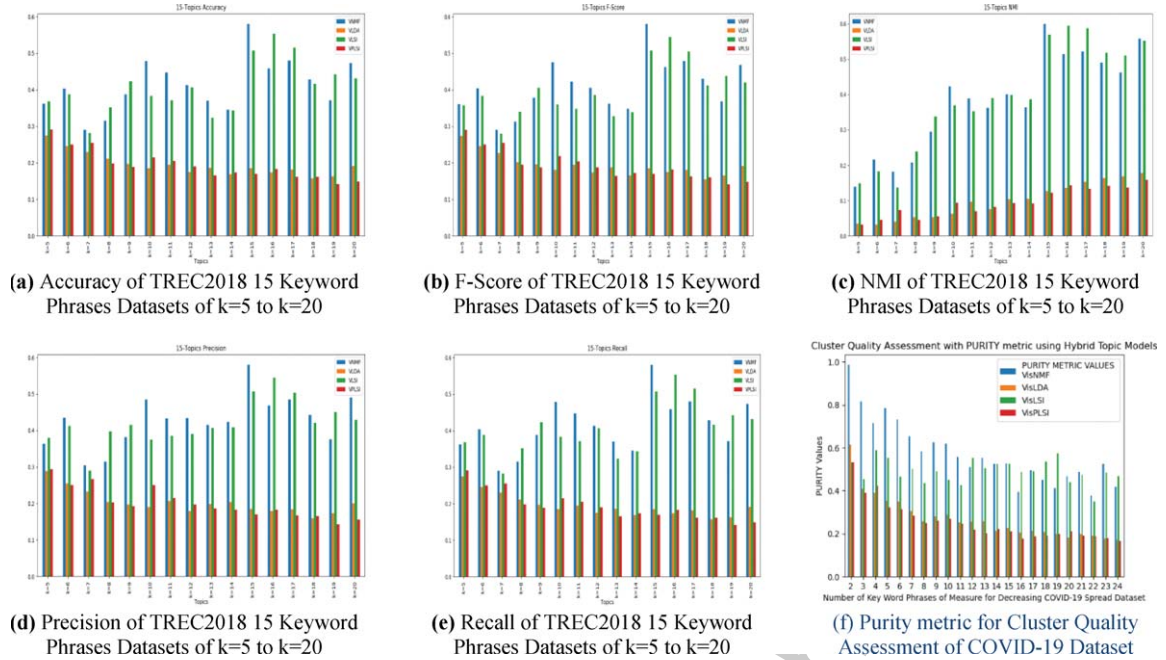


Fig. 4. Performance measure of TREC2018 fifteen keyword phrases datasets and purity metric of COVID-19 dataset.

sometimes anti-correlated. Topic coherence [9] is a relative measure to find the number of topics produced by a particular topic model and to measure the quality of the topic. It measures by considering the semantic similarity of topics. It is not suitable to measure semantically unrelated topic words. The use of cluster analysis is one of the tools for analyzing the stability of topics and in finding the optimal number of topics. For cluster validity, external and internal validity indices are widely applied to a large set of document clustering. Recent advances in biotechnology have generated a massive amount of biological and medical data for disease diagnosis, pathogen identification, identification of patterns and structures among a large set of samples, and pulseNet merged by the Center for Disease Control (CDC). Here we used document-based validity indices to measure quality and to find the number of topics viz. accuracy [20], normalized mutual information [1], F-score, recall, precision [17, 30], Adjusted Rand Index (ARI), Adjusted Mutual information (ADJ), Fowlkes-Mallows Index (FMI), Silhouette Coefficient Index (SIL), Calinski-Harabasz Index (CHI), and Davis-Bouldin Index (DBI). A sample confusion matrix [29] generated for TREC2018 twenty keyword phrases datasets collected as described in Table 1, when  $k = 20$  for four models generated results are shown in Table 2, from these true positive (TP),

true negative (TN) and false-negative (FN) values are calculated, and with these values, accuracy, NMI, F-Score, recall and precision values for each topic varying values from  $k=2$  to  $k=25$  are calculated and tabulated. Results are represented in the form of a sample table and sample line graphs from which the optimal number of topics is estimated based on these values of indices. In Table 3 and Table 4 performance measure values of accuracy, normalized mutual information, F-Score, recall, and precision values of TREC2018 seventeen keyword phrases same dataset for different values of  $k$  ranging from  $k=5$  to  $k=20$  are tabulated as a sample. On observation of these values, it is found that at  $k=17$  all validity indices values are higher for each of these five validity indices which are highlighted by bolded, and among these four hybrid topic models, Visual NMF and Visual LSI performed well. Visual NMF is probabilistic, whereas Visual LSI is not a probabilistic topic model. Similarly, all values for all topics for different values of  $k$  are tabulated. On overall observations, the Visual NMF topic model performs well in visual representation and in finding the optimal number of topics based on values of validity indices measures. In Fig. 4(a) to Fig. 4(e), a performance measure of TREC2018 fifteen keyword phrases tweets datasets collected from Twitter on four different topics models Visual NMF (VNM), Visual LDA (VLDA), Visual



Table 2  
Confusion Matrices of TREC2018 Twenty Keyword Phrases for  $k=20$

Confusion Matrix of Visual NMF	Confusion Matrix of Visual LDA
[[220000000080000000000000]]	[[432004310320011101113]]
[741220111200000000000000]	[04240114113301100301]
[061410000000000000000000]	[01506010022313204000]
[000241600000000000000000]	[00061205311640043202]
[000014160000000000000000]	[2022432414240000000000]
[0200613241270100000002]	[41124711204401111500]
[00000122028430600011]	[01014150700010311113]
[050000819500001000020]	[30201135422131043212]
[01200011001100010000401]	[00421014641260003006]
[101432230350001200213]	[23202616391422351260]
[0010015007230100001011]	[03243412216411212272]
[0100005012225110000003]	[00423536022700316204]
[0030001010011912600007]	[41430100312790111660]
[003000000301141400032]	[11010112132237252033]
[0000022101500018000101]	[03053132014201610620]
[0020100110000011416013]	[032311003111251112003]
[00000000000000001921000]	[31170011123302017403]
[000000030000000532900]	[22412102111401100773]
[003001230610021104160]	[32344112021312202340]
[001002030510302100166]	[21022023200072132425]
Confusion Matrix of Visual LSI	Confusion Matrix of Visual VPLSI
[[2180000000000000000001]]	[[40315410001120001133]]
[127000101000000000000000]	[07110043012101301311]
[08120000000040000000006]	[15311201001121116201]
[00121600000000000005007]	[2073223254015000020]
[00799000000100003001]	[02145015010400110050]
[025091900000200000012]	[10142311225132135012]
[0000081100000000000110]	[21000273232210010301]
[6000160241002000000000]	[16120008341401020250]
[21002212270200000010]	[10321323812302021510]
[0000000063420000000000]	[111244413105015225333]
[0000203016181800000020]	[04043312056341040415]
[517552203001800000011]	[122122220044741421045]
[213010000002371000021]	[61212561112471032212]
[0000100200003340000000]	[12411043043316120202]
[0000100800010030000000]	[30122030126013431242]
[001020000000000290080]	[10210333135011344203]
[0100411020001000118020]	[23021310213213518011]
[003000100000100123000]	[11213013201102554710]
[310021212001040021722]	[15122033005214031151]
[0100010000000000000029]]	[10013445332101010128]]

LSI (VLSI), and Visual PLSI (VPLSI) of accuracy, normalized mutual information, F-Score, precision, and recall are represented as bar graphs for the same dataset for different values of  $k=5$  to  $k=20$ . On observation of these experimental results of all validity indices Visual NMF (VNMF) values are high at  $k=15$ , in the case of Visual LSI values are high in between  $k=15$  to  $k=18$ . The rest of the two models' results are not at a considerable level. From these results, the optimal number of topics determined is Fifteen, which also proved from the visual representation of VAT images. In Fig. 4(f) the cluster quality of Measures to decrease COVID-19 spread dataset

ensured with purity metric calculated using TP, TN, FP, and FN values.

$$\text{Purity} = \frac{\text{TP}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

On observation of the comparative bar graph, we can infer that purity of the cluster decreases as the number of keywords increases, but consistency is maintained with all four hybrid topic models. Among these four models, VisNMF model values are higher and maintain a good quality of cluster than other models.

Table 3  
Performance measures of TREC2018 seventeen keyword phrases dataset for values of  $k=5$  to  $k=20$

k Value	Accuracy				Normalize Mutual Information				F-Score			
	Vis NMF	Vis LDA	Vis LSI	Vis PLSI	Vis NMF	Vis LDA	Vis LSI	Vis PLSI	Vis NMF	Vis LDA	Vis LSI	Vis PLSI
k=5	0.3897	0.2852	0.5058	0.2867	0.2122	0.0320	0.2205	0.0322	0.3907	0.2909	0.5136	0.2893
k=6	0.4102	0.2514	0.4294	0.2867	0.2446	0.0374	0.2437	0.0618	0.4187	0.2546	0.4302	0.2876
k=7	0.4735	0.2176	0.4485	0.2235	0.3681	0.0414	0.2666	0.0422	0.4773	0.2186	0.4413	0.2287
k=8	0.3838	0.2117	0.3647	0.1897	0.2876	0.0499	0.2675	0.0366	0.4026	0.2081	0.3560	0.1845
k=9	0.3411	0.2205	0.4470	0.1823	0.2684	0.0695	0.3562	0.0462	0.3423	0.2225	0.4361	0.1817
k=10	0.3102	0.1838	0.4426	0.1823	0.2954	0.0759	0.3563	0.0591	0.3077	0.1859	0.4420	0.1833
k=11	0.3720	0.1852	0.3794	0.1897	0.3251	0.0707	0.3367	0.0798	0.3687	0.1838	0.3928	0.1879
k=12	0.3661	0.1897	0.4044	0.1764	0.3772	0.1048	0.3577	0.0783	0.3675	0.1900	0.4029	0.1774
k=13	0.4205	0.1647	0.3705	0.1720	0.4159	0.0852	0.3728	0.0833	0.4164	0.1620	0.3862	0.1714
k=14	0.4147	0.1676	0.3676	0.1573	0.4033	0.1084	0.3908	0.0926	0.4012	0.1652	0.3615	0.1580
k=15	0.4073	0.1779	0.3397	0.1764	0.4351	0.1473	0.3807	0.1282	0.3886	0.1695	0.3297	0.1717
k=16	0.3705	0.1617	0.3632	0.1676	0.4567	0.1289	0.4209	0.1160	0.3610	0.1527	0.3609	0.1650
<b>k=17</b>	<b>0.4955</b>	0.1897	<b>0.5117</b>	0.1705	<b>0.5727</b>	0.1554	<b>0.5618</b>	0.1365	<b>0.4955</b>	0.1897	<b>0.5117</b>	0.1705
k=18	0.4720	0.1617	0.4544	0.1720	0.5023	0.1554	0.5396	0.1392	0.4702	0.1622	0.4468	0.1723
k=19	0.4000	0.1588	0.3941	0.1647	0.4748	0.1599	0.5031	0.1363	0.3976	0.1579	0.3914	0.1642
k=20	0.3764	0.1544	0.3970	0.1455	0.4737	0.1626	0.4830	0.1480	0.3825	0.1545	0.3991	0.1479

Table 4  
Performance measures of TREC2018 seventeen keyword phrases dataset for values of  $k=5$  to  $k=20$

k Value	Recall				Precision			
	Vis NMF	Vis LDA	Vis LSI	Vis PLSI	Vis NMF	Vis LDA	Vis LSI	Vis PLSI
k=5	0.389706	0.285294	0.505882	0.286765	0.430997	0.311134	0.558298	0.332057
k=6	0.410294	0.251471	0.429412	0.286765	0.522565	0.295172	0.596572	0.304951
k=7	0.473529	0.217647	0.448529	0.223529	0.523545	0.236385	0.466879	0.250301
k=8	0.383824	0.211765	0.364706	0.189706	0.463195	0.225968	0.372013	0.209321
k=9	0.341176	0.220588	0.447059	0.182353	0.3635	0.235639	0.45542	0.197593
k=10	0.310294	0.183824	0.442647	0.182353	0.32619	0.199748	0.461301	0.187542
k=11	0.372059	0.185294	0.379412	0.189706	0.381439	0.191255	0.423079	0.192218
k=12	0.366176	0.189706	0.404412	0.176471	0.394472	0.219083	0.419595	0.182765
k=13	0.420588	0.164706	0.370588	0.172059	0.455007	0.167989	0.448979	0.17259
k=14	0.414706	0.167647	0.367647	0.157353	0.415973	0.172749	0.393156	0.176142
k=15	0.407353	0.177941	0.339706	0.176471	0.415463	0.179652	0.38936	0.202642
k=16	0.370588	0.161765	0.363235	0.167647	0.446776	0.170922	0.418926	0.177841
<b>k=17</b>	<b>0.495588</b>	0.189706	<b>0.511765</b>	0.170588	<b>0.495588</b>	0.189706	<b>0.511765</b>	0.170588
k=18	0.472059	0.161765	0.454412	0.172059	0.470431	0.163419	0.442384	0.173398
k=19	0.400000	0.158824	0.394118	0.164706	0.398442	0.159865	0.391518	0.164986
k=20	0.376471	0.154412	0.397059	0.145588	0.405054	0.157726	0.421563	0.156772

In Fig. 5(a) to Fig. 5(f), a performance measure of Measures to decrease COVID-19 spread tweets dataset collected from Twitter on four different topics models Visual NMF (VNMF), Visual LDA (VLDA), Visual LSI (VLSD) and Visual PLSI (VPLSI) of ADJ, ARI, FMI, HI, CHI, and DBI metrics are represented as bar graphs for the same dataset on different values of  $k=5$  to  $k=24$ . On observation of these experimental results of all validity indices Visual NMF (VNMF) values are high at  $k=10$ , and low for the DBI validity index. From these results, the optimal number of topics determined is ten, which also proved from the visual representation of VAT images. In Fig. 6(a) to Fig. 6(d), performance measures results of accuracy, NMI, F-Score, Recall and Precision of TREC2018 Twenty keyword phrases tweets datasets collected

from Twitter and processed on four different topics model Visual NMF (VNMF), Visual LDA (VLDA), Visual LSI (VLSD) and Visual PLSI (VPLSI) are represented as line graphs for the same dataset for different values of  $k=5$  to  $k=20$ . On observation of these experimental results of all validity indices Visual NMF (VNMF) values are high at  $k=20$ , in the case of Visual LSI values are higher in between  $k=18$  to  $k=20$ . The line graph of Visual NMF (VNMF) at  $k=20$  is steep and well defined, whereas in the case of Visual LSI (VLSD) line is blunt and lies between 18 and 22. The rest of the two models' results are not at a considerable level. From these results the optimal number of topics determined as  $k=20$ , which also proved from the visual representation of VAT images. Similarly, the optimal number of topics is

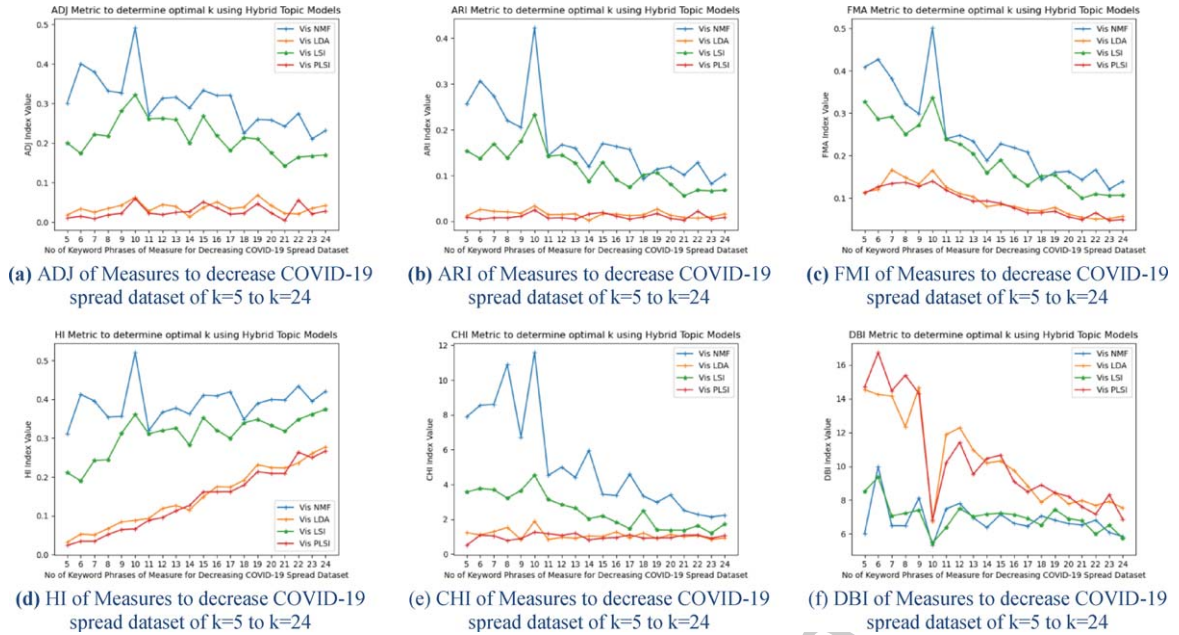


Fig. 5. Performance measure of measures to decrease COVID-19 spread dataset with ten keyword phrases.

determined and results are recorded for all datasets used in our experiments. In Fig. 6(e) to Fig. 6(h), performance measures results of accuracy, NMI, F-Score, Recall and Precision of 7 Topics of health tweets datasets collected from Twitter and processed on four different topics model Visual NMF (VNMF), Visual LDA (VLDA), Visual LSI (VLSI) and Visual PLSI (VPLSI) are represented as line graphs for the same dataset for different values of  $k = 2$  to  $k = 10$ . On observation of these experimental results of all validity indices, Visual NMF (VNMF) values are high at  $k = 7$  than other models. The line graph of Visual NMF (VNMF) at  $k = 7$  is steep and well defined, whereas in the case of Visual LSI (VLSI) line graph is steep but values are less than VNMF results. From these results the optimal number of topics determined as  $k = 7$ , which also proved from the visual representation of VAT images. Similarly, the optimal number of topics is determined and results are recorded for all health-related datasets used in our experiments. In Fig. 6(i) to Fig. 6(l), performance measures results of accuracy, NMI, F-Score, Recall and Precision of TREC2018 Twenty Four keyword phrases tweets datasets collected from Twitter and processed on four different topics models Visual NMF (VNMF), Visual LDA (VLDA), Visual LSI (VLSI) and Visual PLSI (VPLSI) are represented as line graphs for the same dataset for different values of  $k = 5$  to  $k = 20$ . On observation of these experimental results of all

validity indices Visual NMF (VNMF) values are high at  $k = 24$ , in the case of Visual LSI values are higher at  $k = 24$  but lesser than Visual NMF (VNMF) values. The line graph of Visual NMF (VNMF) and Visual LSI (VLSI) at  $k = 24$  is steep and well defined, but values vary. From these results the optimal number of topics determined as  $k = 24$ , which also proved from the visual representation of VAT images. Similarly, the optimal number of topics is determined and results are recorded for all datasets used in our experiments.

In Fig. 7(a) to Fig. 7(d), performance measures results of accuracy, NMI, F-Score, Recall, Precision, ARI, ADJ, FMI of Measures to decrease COVID-19 spread tweets datasets collected from Twitter and processed on four different topics model Visual NMF (VNMF), Visual LDA (VLDA), Visual LSI (VLSI) and Visual PLSI (VPLSI) are represented as line graphs for the same dataset for different values of  $k = 5$  to  $k = 24$ . On observation of these experimental results of all validity indices, Visual NMF (VNMF) values are high at  $k = 10$  than other models. The line graph of Visual NMF (VNMF) at  $k = 10$  is steep and well defined, whereas in the case of Visual LSI (VLSI) line graph is steep but values are less than VNMF results. From these results the optimal number of topics determined as  $k = 10$ , which also proved from the visual representation of VAT images. In Fig. 7(e) and Fig. 7(f), performance measures results

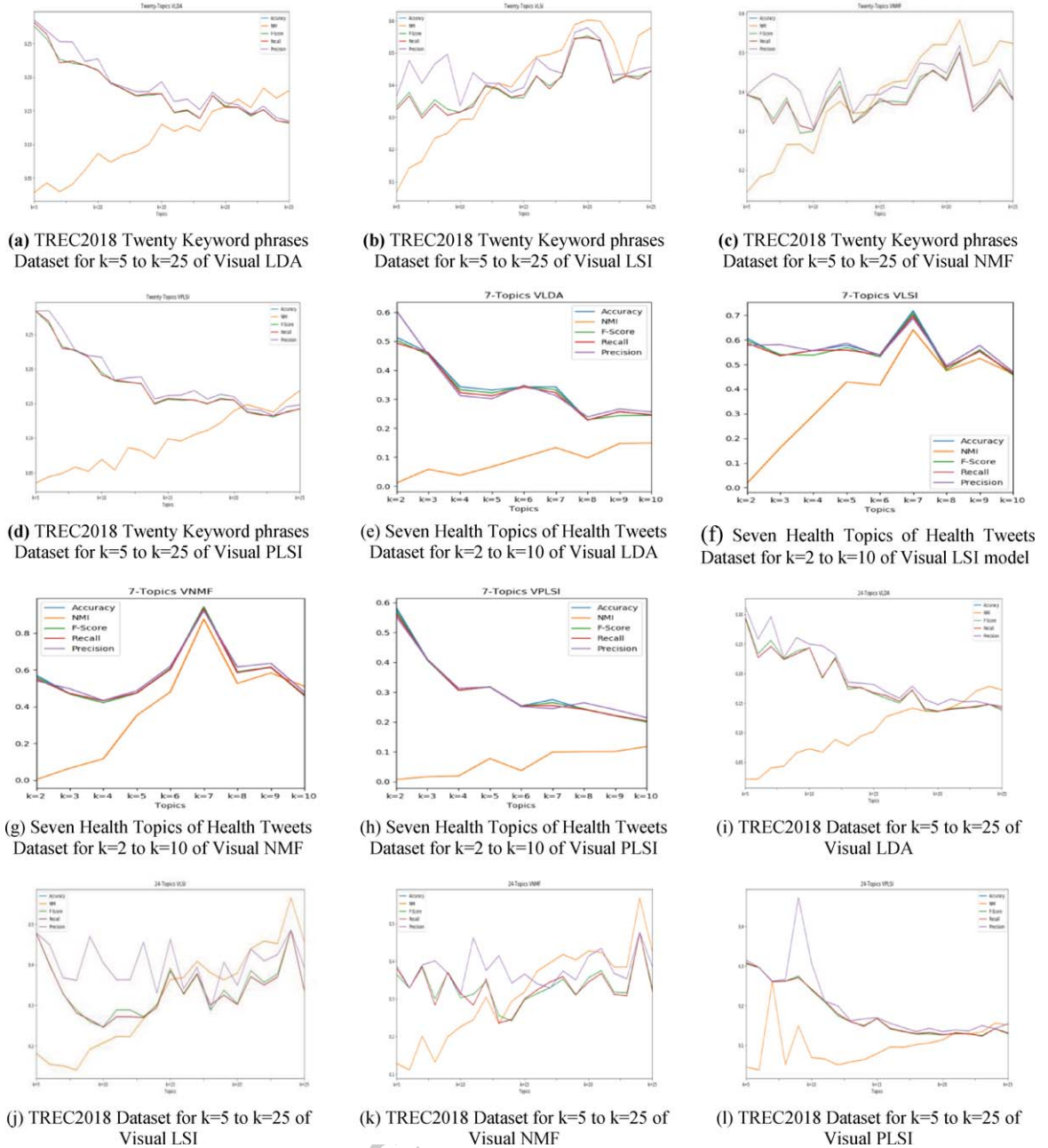


Fig. 6. Performance comparison of TREC2018 and health topics datasets with four hybrid topic models.

of SIL, CHI, and DBI internal indices of Measures to decrease COVID-19 spread tweets datasets collected from Twitter and processed on four different topics model Visual NMF (VNMF), Visual LDA (VLDA), Visual LSI (VLSI) and Visual PLSI (VPLSI) are represented as line graphs for the same dataset for different values of  $k = 5$  to  $k = 24$ . On observation of these experimental results of validity indices SIL, and

CHI values of Visual NMF (VNMF) values are high and DBI values are low at  $k = 10$  than other models. The line graph of Visual NMF (VNMF) at  $k = 10$  is steep and well defined, whereas in the case of Visual LSI (VLSI) line graph is steep but values are less than VNMF results. From these results the optimal number of topics determined as  $k = 10$ , which also proved from the visual representation of VAT images.

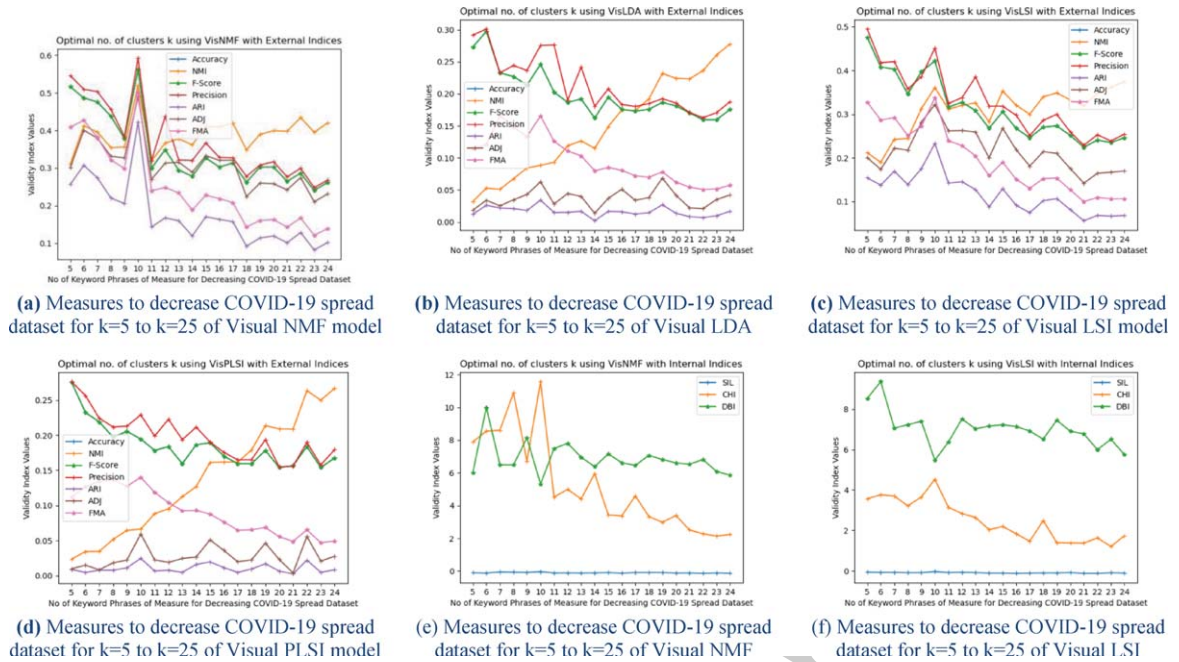


Fig. 7. Performance comparison of measures to decrease COVID-19 spread dataset with four hybrid topic models.

#### 4. Conclusion and future enhancement

Topic modeling is an effective means of data mining and cluster analysis in machine learning to build models from unstructured textual data, where samples are treated as documents. While applying topic modeling key challenge is the selection of the appropriate number of topics and choosing a quality metric in performance measurement. Most of the previous works in the determination of the number of topics are non-parametric and the quality of topics determined by using perplexity and coherence measures, to overcome limitations we proposed the parametric method, which is an extension of the traditional topic model with visual access tendency for visualization of the number of topics (clusters) to complement clustering and to choose the optimal number of topics based on results of cluster validity indices. To summarize, we have proposed a new measure for identifying the right number of topics in a given corpus by considering distributions generated from topic-word and document-topic matrix outputs of hybrid topic models, visual representation of the number of topics with VAT images, and performance measure with validity indices. Experimental evaluations on different datasets of health-related tweets, TREC2014, TREC2015, TREC2018, and Measures to decrease COVID-19 spread keyword phrases based datasets

with four proposed visual topic models proved that they are useful in interactive visualization of the optimal number of topics and for selecting the number of topics. Each of the proposed hybrid topic models' highest probable topic assignment, feature selection, feature extraction, visual representation, yielded the best clustering results of different types of datasets experimented and also in finding the different number of clusters among datasets. We examine the usefulness and stability of models with alternative models and applied cluster validity indices accuracy, NMI, F-Score, recall, precision, ARI, ADJ, FMI, HI, SIL, CHI, and DBI to measure quality and in the determination of the optimal number of topics by running the same datasets with different values of k and choose the optimal number of topics. This application of hybrid topic models approach to cluster analysis of large Twitter-based datasets and validity indices in finding an optimal number of clusters can greatly improve the accuracy and efficiency of subgroup identification and proposed hybrid models and techniques in tweets document clustering and optimal number identification provide a new approach for data mining of twitter-based datasets in biological and medical research. The results confirm that the Visual NMF (VNMF) model is stable, accurate, and effective for all numerical experiments presented and for various datasets used in visual representation



and in determining the optimal number of topics. The future work is to extend hybrid topic models as scalable approaches for reducing the computational cost and space complexity values.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Conflicts of interest

The authors declare no conflict of interest.

## References

- [1] Alessia Amelio & Clara Pizzuti (2016, February). Is Normalize Mutual Information a Fair Measure for Comparing Community Detection Methods? Conference paper in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France. DOI: 10.1145/2808797.2809344
- [2] D.M. Blei, A.Y. Ng and M.I. Jordan, Latent dirichlet allocation, *J. Mach. Learn* **3** (2003a), 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993
- [3] D.M. Blei, M.I. Jordan, T.L. Griffiths and J.B. Tenenbaum, (2003b, December). Hierarchical Topic Models and the Nested Chinese Restaurant Process. Conference paper in proceedings of the 16th International Conference on Neural Information Processing Systems, Whistler, BC, Canada.
- [4] D.M. Blei, T.L. Griffiths and M.I. Jordan, The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies, *J. ACM* **57** (2010), 1–7. <https://doi.org/10.1145/1667053.1667056>
- [5] Carina Jacobi, Wouter van Atteveldt and Kasper Welbers, Quantitative analysis of large amounts of journalistic texts using topic modeling. *Digital Journalism*. (2015). DOI: 10.1080/21670811.2015.1093271
- [6] Carson Sievert and Kenneth E. Shirley, LDAvis: A method for visualizing and interpreting topics. Conference paper in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA (2014, June). DOI:10.3115/v1/W14-3110.
- [7] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang and D.M. Blei, (2009, December). Reading Tea Leaves: How Humans Interpret Topic Models. Conference paper in proceedings of the 22<sup>nd</sup> International Conference on Neural Information Processing Systems, Vancouver, BC, Canada. DOI:10.1.1.154.992
- [8] J. Chuang, M.E. Roberts, B.M. Stewart, R. Weiss, D. Tingley, J. Grimmer and J. Heer, (2015, June). Topic Check: Interactive Alignment for Assessing Topic Model Stability. Conference paper in proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Denver, CO, USA.
- [9] C. Damir Korenci, Strahil Ristov and Jan Sneijder, Document-based topic coherence measures for news media text, *Expert systems with Applications* **114** (2018), 357–373. DOI: 10.1016/j.eswa.2018.07.063.
- [10] Dataset keyword phrases 2018 Precision Medicine Track TREC2018. Track web page <http://www.trec-cds.org/>
- [11] D. Greene, D. Callaghan and P. Cunningham, How many topics? stability analysis for topic models, *Machine Learning and Knowledge Discovery in Databases* **8724** (2014), 498–513. [https://doi.org/10.1007/978-3-662-44848-9\\_32](https://doi.org/10.1007/978-3-662-44848-9_32)
- [12] Georgios Balikas, Massih-Reza Amini, & Marianne Clausel (2016, July). On a Topic Model for Sentences. Conference paper in proceedings of 39<sup>th</sup> international aem sigir conference on research and development in information retrieval. <https://doi.org/10.1145/2911451.2914714>
- [13] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Skasko, & Haesun Park (2012, June). iVisClustering: An interactive Visual Document Clustering via Topic Modeling. Conference paper in proceedings of Eurographics Conference on Visualization (EuroVis). <https://doi.org/10.1111/j.1467-8659.2012.03108.x>
- [14] Y. Hu, J. Boyd-Graber, B. Satinoff and A. Smith, Interactive topic modeling, *Machine Learning* **95**(3) (2014), 423–469. <https://doi.org/10.1007/s10994-013-5413-0>
- [15] Jaegul Choo, Changhyun Lee, Chandan K. Reddy and Haesun Park, UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization, *IEEE Transaction on Visualization and Computer Graphics* **19** (2013), 1992–2001. DOI: 10.1109/TVCG.2013.212.
- [16] Laurens van der Maaten and Geoffrey Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research* **9** (2008), 2579–2605.
- [17] Z. Li, W. Shang and M. Yan, News text classification model based on the topic model, IEEE/ACIS 15<sup>th</sup> International Conference on Computer and Information Science (ICIS) (2016, June), DOI:10.1109/icis.2016.7550929.
- [18] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers and Padhraic Smyth (2014, July). The Author-Topic Model for Authors and Documents. Conference paper in proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI2004). DOI: arXiv:1207.44169[cs.LG]
- [19] D. Newman, J.H. Lau, K. Grieser and T. Baldwin, (2010, June). Automatic Evaluation of Topic Coherence. Conference paper in proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, USA.
- [20] M. Pattanodom, N. I am-On and T. Boongoen, (2016, January). Clustering data with the presence of missing values by an ensemble Approach. Conference paper in 2016 second Asian Conference on Defense Technology (ACDT), Thailand. DOI:10.1109/acdt.2016.7437660
- [21] K. Rajendra Prasad, Moulana Mohammed and R.M. Noorullah, Visual topic models for healthcare data clustering, *Evolutionary Intelligence* **1** (2019a), 1–17. <https://doi.org/10.1007/s12065-019-00300-y>
- [22] K. Rajendra Prasad, Moulana Mohammed and R.M. Noorullah, Hybrid topic cluster models for social healthcare data, *International Journal of Advanced Computer Science and Applications (IJACSA)* **10**(11) (2019b), 491–506. DOI: 10.14569/IJACSA.2019.0101168
- [23] Rubayyi Alghamdi and Khalid Alfalqi, A survey of topic modeling in text mining, *International Journal of Advanced Computer Science and Applications (IJACSA)* **6**(1) (2015), DOI: 10.14569/IJACSA.2015.060121.

- [24] S. Senthilkumar, M. Srivani and G.S. Mahalakshmi, (2017, June). Generation of word clouds using Document Topic Models. Conference paper in Second International Conference on Recent Trends and Challenges in Computational Models. DOI:10.1109/ICRTCCM.2017.60.
- [25] Sergey Karpovich, Alexander Smirnov, Nikolay Teslya and Andrei Grigorev (2017, October). Topic Model Visualization with (14) IPython. Conference paper in proceeding of the 20<sup>th</sup> conference of Fruct Association. DOI: 10.23919/FRUCT.2017.8071303.
- [26] ShaohuaLi and Tat-SengChua (2017). Document Visualization using Topic Clouds. Information retrieval, DOI: arXiv:1702.01520v1 [cs.IR]
- [27] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei (2004, June). Sharing Clusters Among Related Groups: Hierarchical Dirichlet Processes. Conference paper in proceedings of the 17<sup>th</sup> International Conference on Neural Information Processing Systems, Vancouver, BC, Canada
- [28] H.M. Wallach, I. Murray, R. Salakhutdinov and D. Mimno, (2009, June). Evaluation Methods for Topic Models. Conference paper in Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada. <https://doi.org/10.1145/1553374.1553515>
- [29] Wongkot Sriurai, Phayung Meesad and Choochart Haruechaiyasak, Hierarchical Web Page Classification Based on a Topic Model and Neighboring Pages Integration, *International Journal of Computer Science and Information Security (IJCSIS)* **7**(2) (2010), DOI:arXiv:1003.1510[cs.LG].
- [30] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang and H. Yao, Research on topic detection and tracking for online news texts, *IEEE Access* **7** (2019), 58407–58418. DOI:10.1109/access.2019.2914097.