The Journal of Supercomputing (2022) 78:9619–9641
https://doi.org/10.1007/s11227-021-04300-7

A novel sampling-based visual topic models with computational intelligence for big social health data clustering

K. Narasimhulu[1] · K. T. Meena Abarna[1] · B. Siva Kumar[1,2] · T. Suresh[1]

Accepted: 28 December 2021 / Published online: 19 January 2022

Abstract

Twitter is a popular social network for people to share views or opinions on various topics. Many people search for health topics through Twitter; thus, obtaining a vast amount of social health data from Twitter is possible. Topic models are widely used for social health-care data clustering. These models require prior knowledge about the clustering tendency. Determining the number of clusters of given social health data is known as the health cluster tendency. Visual techniques, including visual assessment of the cluster tendency, cosine-based, and multiviewpoint-based cosine similarity features VAT (MVCS-VAT), are used to identify social health cluster tendencies. The recent MVCS-VAT technique is superior to others; however, it is the most expensive technique for big social health data cluster assessment. Thus, this paper aims to enhance the work of the MVCS-VAT using a sampling technique to address the big social health data assessment problem. Experimental is conducted on different health datasets for demonstrating an efficiency of proposed work. Accuracy of social health data clustering is improved at a rate of 5 to 10% in the proposed S-MVCS-VAT when compared to MVCS-VAT. From obtained results, it also proved that the proposed S-MVCS-VAT is a faster and memory efficient for discovering social health data clusters.

Keywords Big social data · Health cluster tendency · Visual techniques · Topic models · Tweet data

✉ K. Narasimhulu
narsimhulu.kolla@gmail.com

Extended author information available on the last page of the article

1 Introdcution

Twitter is one of the platforms commonly used by social users to share opinions or trusts across the world. People can share experiences or opinions through tweets. Health care data are an emerging need for society, and it is necessary to automate tweet health data to identify major health problems in society. Usually, health-care tweet data are extensive, and tweet data need to be assessed to find knowledge about significant health problems (or health clusters). This is the crucial motivation for addressing the health cluster tendency problem. Visual techniques, such as VAT [1], cVAT [2], and MVCS-VAT [3], can be used to access information about several clusters of tweet health data (or social health data). Popular topic models, including nonmatrix factorization (NMF) [4], latent semantic indexing (LSI) [5], probabilistic LSI (PLSI) [6], and latent Dirichlet allocation (LDA) [7], are used to extract the topic features of tweet data. The topic-tweet document matrix is created using the topic models for the set of tweet documents. TF-IDF is another alternative matrix for describing tweet document features based on term analysis, and the matrix usually known as the TF-IDF matrix [8]. Tweet document analysis using topics is more practical than using the TF-IDF matrix because data sparsity occurs in the TF-IDF matrix.

The topic-document matrix (TDM) is the most recommended approach in text clustering applications [9] [25]. Dissimilarity features are derived using a Euclidean distance measurement in a VAT. In a cVAT, the dissimilarity features are derived using the cosine distance metric. In the majority of text clustering applications [10] [23][26], the authors proved that cosine-based cluster assessment is more informative than a standard Euclidean distance formula. In a cVAT, the cosine-based similarity is measured using a single reference viewpoint, i.e., the origin. An extended version of the cVAT is the MVCS-VAT [3]. In MVCS-VAT, the cosine-based similarity values are derived using multiple viewpoints. Deriving the similarity using multiple viewpoints is a more accurate mechanism than a single viewpoint approach in the cVAT. Justifying the cluster assessment using the multiviewpoint cosine-based similarity values is more appropriate than the justification of a single viewpoint. The recent MVCS-VAT methods conducts the cluster assessment of health data in an excellent manner [27][31]. Each cluster represents a health cluster, which clusters the tweets; and those tweets belong to the same health topic are discussed. The tweets are categorized into health clusters based on the similarities among tweet documents. The problem of the MVCS-VAT is that it takes more computational time and memory space due to the assessment of health clusters using multiple viewpoints. For example, finding the similarity between two tweets documents t1 and t2 among the n documents is performed using n-2 viewpoints. Every tweet among the n tweets is taken as a

viewpoint except t1 and t2; hence, there are 'n-2' viewpoints. The cosine similarity is computed between two tweet documents for n-2 viewpoints. Finally, similarity computation is applied for n(n-1)/2 cases concerning n-2 viewpoints. Thus, the total computation time is n(n-1)(n-2)/2. Therefore, the MVCS-VAT is a more expensive cluster assessment model for a large number of tweet documents. The proposed work uses an effective sampling procedure to further extend the MVCS-VAT[28]. The existing study proposes using a constant number of sample viewpoints instead of taking the n-2 multiple viewpoints in the proposed sampling-based MVCS-VAT (S-MVCS-VAT) algorithm. The algorithm and experimental details are demonstrated in the next sections.

The key contributions of the paper are summarized as follows:

1. Health clusters from big social data are assessed.
2. A sampling-based visual technique for determining the health clusters in a visual form is proposed.
3. Crisp partitions are derived from the visual images from the proposed S-MVCS-VAT.
4. Significant social health data cluster results are derived.
5. The performance of visual techniques for social and benchmark health data is empirically demonstrated.

The remaining sections are summarized as follows: Sect. 2 presents the literature on visual techniques for precluster assessment; Sect. 3 introduces the proposed sampling-based MVCS-VAT; Sect. 4 illustrates the experimental study; and, finally, Sect. 5 provides the conclusion and future scope of the work.

2 Literature of visual techniques for precluster assessment

Top clustering methods, such as k-means [11] and hierarchical clustering, are widely used in clustering-related applications [12]. The data clustering process depends on two crucial steps: finding the knowledge about the number of clusters and making a data partition of the data. Determining the number of clusters is known as the cluster tendency problem. Social health data are the opinions or views of social users on Twitter. Social health data are tweeted health data. Finding the categories of clusters of social health data based on health topics is known as finding the health cluster tendency [29]. The preassessment of several health topics in social data is a challenging problem. With this motivation, many visual techniques are surveyed for the precluster assessment of social health data. Bezdek et al. [1] proposed a basic model, namely the visual assessment of (cluster) tendency (VAT), for determining the number of clusters of numerical data. It works for numerical data. Its algorithmic is shown in the following.

```
Algorithm: Visual Assessment of Tendency [1]
Input: ObjdissM[][]

Output  : Number of clusters (or cluster tendency k)

Step 1:
        Obj_order= { };
        Obj_int_order={0,1,.....n-1}
        Determine  max of ObjdissM[ ] [ ], and its   index value is stored into (i,j)
        OrderP(0)=i;
       Obj_order ={i};
       Obj_int_order = Obj_int_order -{ Obj_order;

Step 2:
        for (s=1;s<n;s++)
        {
          Find(i,j) from min {ObjdissM[i][j], where
                                  i€ IV, j e{JV}}
          Obj_order = { Obj_order } U{j};
          Obj_int_order={Obj_int_order}-{ Obj_order};
          OrderP(s)=j;
        }

Step 3:
        /*Reordered Dissimilarity Matrix Computation*/
       for(i=0;i<n;i++)
            for(j=0;j<n;j++)
              RDM=ObjdissM(OrderP[i],OrderP[j]);

Step 4:
         Display Image (RDM)
```

Thus, social data are initially preprocessed into the topic-document matrix using various topic models [13]. This is a better representation of social data than the TF-IDF matrix. Four topic models, latent Dirichlet allocation, latent semantic indexing (LSI), probabilistic latent semantic indexing (PLSI), and nonnegative matrix factorization (NMF), are the recommended topic models in text clustering-related applications. These models are used to convert the social data into a numeric topic-document matrix. With this matrix, social health data are denoted in the form of a numeric representation. In a VAT [14], the social health topic-document matrix is used to find the dissimilarity features using the Euclidean distance matrix. The reordered dissimilarity matrix (RDM) [15] is derived according to the given steps of the VAT and then displays the image of the RDM. The number of health clusters (or health cluster tendency) is derived from the count of the number of square-shaped dark colored blocks in the RDM image (also known as the VAT image). A cosine metric uses vectors' magnitude and distance to find the similarity features between two data objects whereas a Euclidean distance metric only uses the distance. Therefore, in a text clustering application, cosine-based cluster assessment succeeds more than Euclidean distance assessment. Following a cosine metric, another visual technique, i.e., the cosine-based VAT (cVAT), was developed in [12] for the precluster assessment problem.

In the cVAT, the similarity (or dissimilarity) features between two data objects are derived using a single viewpoint, i.e., the origin. Computing similarity features using a single viewpoint cannot provide a more informative assessment. Thus, multiple viewpoints are used in the later development of visual techniques, such as the multiviewpoint-based cosine similarity features VAT (MVCS-VAT) [3]. The MVCS-VAT is the most recommended visual technique to acquire accurate similarity features using a multiple viewpoint strategy instead of just a single viewpoint. For n tweet documents, as per the MVCS-VAT, n-2 viewpoint computations are needed to find the cosine-based similarity features among any two tweet documents. Finally, average n-2 similarity features concerning n-2 viewpoints are taken as the similarity features between the two tweet documents. This method is most accurate for visualizing the number of clusters for the set of n tweet documents [30]. The approach for the similarity feature computation between any two documents for the set of five tweet documents is shown in Fig. 1.

The key limitation of the MVCS-VAT is that it demands more computational time and memory allocation for finding the social data clustering results from a large set of tweet documents. The proposed methods present the best sampling-based MVCS-VAT for the scalable computation of social data health clustering results.

Further work must find the similarity features between the tweet documents for sample viewpoints instead of n-2 viewpoints. Social data are enormous big data; thus, this proposed base sampling idea optimizes the time and memory requirements in finding health cluster tendencies. This optimized approach to find the health cluster tendency from social data is derived in the next section.

3 Proposed sampling-based Mvcs-Vat (S-Mvcs-Vat)

The clustering of social data (tweet health data) depends on the similarity features of data objects. The cosine-based similarity features are very successful in text data clustering applications. The similarity features concerning a single origin or a single reference viewpoint are derived. The MVCS-VAT uses multiple viewpoints to find accurate similarity features among the tweet documents compared to a single

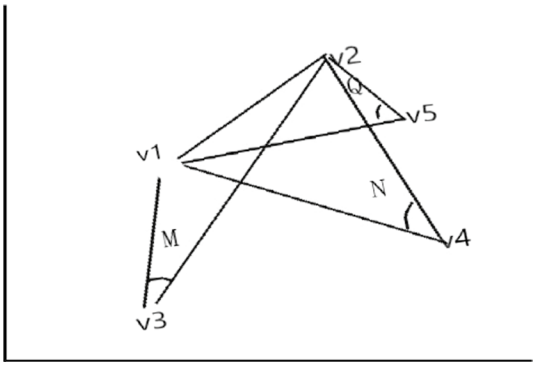

Fig. 1 Sampling viewpoints using cosine similarity

reference viewpoint. Due to the expensiveness of the MVCS-VAT, our proposed work takes the sample viewpoints to determine the quality of social health data clustering results. Algorithm 1 illustrates the procedural steps of the proposed work.

Algorithm 1: S-MVCS-VAT

Input: 'N' - Number of health documents

s- Sample size in percentage

Output: 'k' - Health cluster tendency

'C' – Health clusters

Method:

Step 1: Extract the features of health tweets.

Use topic models to model the health tweets in order to extract the features of health tweets $\{TF_1, TF_2, \ldots TF_N\}$.

Step 2: Find the initial centroid for the starting cluster.

Randomly select any tweet document numbered 'r' among {1, 2, …, N}. Compute the distances TF_r to $\{TF_1, TF_2 \ldots TF_N\}$, save the distances Dist_Start and save the index regarding the maximum distance-maintained tweet document number with respect to the selected tweet document number 'r'. The index and maximum distance are computed as follows:

$$index = argmax_{I \in \{1,2,\ldots N\}} \{ distance(TF_r, TF_I) \}$$

$$Dist\_Start = distance(TF_r, TF_I))$$

Step 3: Determine other approximated centroids.

```
//Update the distance value
For i = 1 to N
  Dist_I=min(Dist_Start,distance(TF_maxindex, TF_i))

//Form the remaining centroids
The next centroid index is determined as follows:
index=argmax_{I ∈ {1,2,... N}} {Dist_I}.
Update the index and Dist_Start as per Step 2. Repeat Step 3 until the expected centroids are obtained.
```

Step 4: Find the clusters of tweets with the nearest centroids $\{C_1, C_2..C_k\}$.

```
Sample_VP={ }

For i=1 to N
  For j=1 to N

//Apply simple random sampling without replacement (SRSWR) for the selected sample viewpoints
For m= 1 to k
 If ((TFi is not in C_m) or (TFj is not in C_m))
Use SRSWR[ ] to select of the sample viewpoint from C_m and save in 'Sample.'
Sample_VP= Sample_VP ∪ {Sample}
LS=size(Sample_VP)
```

$$S\_MVCS = \frac{1}{LS} \sum_{vp \in Sample\_VP} cos(TFi, TFj)$$

$$Dissimilarity\_Matrix(DM(i,j)) = 1 - S\_MVCS$$

$$NormS = (Normalize(S\_MVCS(TF1, TF2), 0, 1)$$

Step 7: Find the reordered dissimilarity matrix for NormS using the VAT [1].

Step 8: Visualize the RDM image and count the detected square-shaped dark colored blocks. The count value defines the clustering tendency of health datasets.

Step 9: Determine the aligned crisp partitions based on the square-shaped dark colored blocks that appear. The crisp partitions given the predicted cluster labels for the health tweet documents are used to discover the health data clusters.

The proposed algorithm uses topic models, such as LDA, LSI, PLSI, and NMF, to extract the features of health tweets in topic-document matrix form. The proposed algorithm reduces the sparsity problem of tweet data. The topic-document matrix was then converted into a bag-of-features representation of tweet data. The features of tweets are denoted in the vector representation $\{TF_1,TF_2,\ldots TF_N\}$.

Randomly select the r^{th} tweet document feature, find the distances between TF_r and $\{TF_1,TF_2,\ldots TF_N\}$ and save the distances into 'Dist_Start.' The maximum distance-maintained tweet data object is determined using the argmax function, and the corresponding tweet document number is saved into the variable 'index.' These are in Step 1 and Step 2. Next, the distance array $Dist_I$ is updated according to explored tweet documents, and this is in Step 3. Again, the tweet document with the largest deviation is selected by applying the argmax function to $Dist_I$. The corresponding index found by the argmax is another centroid of tweet datasets. The same procedural steps are repeated to find the remaining expected number of centroids of the clusters. After selecting the centroids, the remaining tweet documents are moved into the nearest centroids based on the distances measured in Step 4. The distances are measured using the cosine distance metric of the sample viewpoints. The size of the sample viewpoints is measured based on a percentage of s. The mentioned percentage of samples is equally sampled from every cluster (except clusters TF1 and TF2). These steps are clearly illustrated, similarity features concerning the sample viewpoints are computed, and the C_MVCS computational statement is shown. Dissimilarity values are stored in DM, and normalized matrix values are stored in NormS.

The reordered dissimilarity matrix is computed by applying the visual assessment tendency (VAT) to NormS, as shown in Step 7. The RDM image is visualized to assess the number of visual clusters by counting the squared shaped dark colored blocks that appear along the diagonal. The crisp partitions of the RDM image show the predicted cluster labels of health tweets, which discover the health data clustering results; and these steps are clearly illustrated in Step 8 and Step 9.

For the proposed algorithm, the similarity features for the pair of tweet documents are derived using every viewpoint; and finally, the average of the obtained similarity values is used in the computation of tweet document similarity features. The similarity feature computation is less expensive due to taking sample viewpoints instead of a large number of all viewpoints. This provides a considerable improvement for finding the social data clustering results compared to the state-of-the-art visual topic models.

In the recent MVCS-VAT technique, effective social health data clustering results are derived using all given viewpoints. For small datasets, the MVCS-VAT is very impressive at determining the clustering tendency and individual social data clustering results. However, the amount of social data is massive; therefore, the MVCS-VAT uses many viewpoints to find the social health data clustering results. Ultimately, the method demands large computational and spatial costs. The MVCS-VAT is always suitable for finding social data clustering results, and it is expensive for big social data. Our proposed S-MVSC-VAT uses the sampling schema to perform scalable computations for big social data clustering. The experimental demonstrations are presented in the following section.

4 Experimental study

Tweet data [2] are collected on different health topics to assess health data clustering results. Each subset of data is created with specific health topics. Table 1 presents the details of the social health data in terms of a number of health topics [18], names of health diseases, and the size of the datasets.

Benchmarked health datasets are retrieved based on the health keywords provided by TREC [16] [17], which are mentioned in the same table.

After extracting the tweet features in the form of a bag-of-features, various big social data visual clustering methods are tested in the experimental study. Three traditional visual methods, the VAT, cVAT, MVS-VAT, and the proposed S-MVCS-VAT are applied to the provided big social data. Visual images with excellent clarity are provided by both the S-MVCS-VAT and MVCS-VAT compared to other visual methods. The notable improvement of the proposed method is that it can derive faster health data clustering results than the MVCS-VAT.

The crisp partitions are derived based on the diagonal and nondiagonal pixel intensity values. The cluster labels of data objects are derived based on these cluster partitions, and the results are shown in Fig. 5b for three data topics.

Tweet document features are extracted through the four different topic models: LDA, LSI, PLSI, and NMF. Figure 2, Fig. 3, Fig. 4, and Fig. 5a show the results of visual health data clustering for these topic models. From the illustration of the visual health data clustering results, S-MVCS-VAT shows the visual clusters.

in the form of diagonal square-shaped dark colored blocks with outstanding clarity under all four topic models.

The clarity of the proposed work with sampling viewpoints is the best. With sampling viewpoints and without sampling approaches showed almost the same clarity of visual clusters.

Crisp partitions and consequent quality clustering results depend on the clarity of visual image clusters. The S-MVCS-VAT has the ability to obtain social health data clustering results with optimized time and space values. All four proposed variants are developed with the four specified topic models. These are the LDA-S-MVCS-VAT, LSI-S-MVCS-VAT, PLSI-S-MVCS-VAT, and NMF-S-MVCS-VAT. All the comparative analyses of time values (taking the speed parameter) of four variants of existing and proposed models are shown in Figs. 6, 7 and 8. These figures compare the same models using the memory space parameter and time comparison parameter. Empirical analysis of the speed, memory, and time and space costs shows that the proposed S-MVCS-VAT is a more scalable visual health data clustering model in speed and memory efficiency. This leads to the S-MVCS-VAT being faster and more memory efficient than other visual health data clustering models.

The performance or quality of the visual data clustering models is evaluated using four parameters: the cluster accuracy (CA) [19], normalized mutual

Table 1 Social health datasets topics description

S. No.	Number of topics (e.g., 2-T refers 2 Topics dataset)	Health disease topics	Amount of social health data
1	2-T	blood_pressure, bone_density	2.24 MB
2	3-T	blood_pressure, bone_density, common_cold	2.65 MB
3	4-T	blood_pressure, bone_density, common_cold, AIDS	3.10 MB
4	5-T	blood_pressure, bone_density, common_cold, AIDS, Dengue	3.57 MB
5	6-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea	4.07 MB
6	7-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache	4.61 MB
7	8-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache, jaundice	5.20 MB
8	9-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache, jaundice, kidney_stones	5.81 MB
9	10-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache, jaundice, kidney_stones, obesity	6.48 MB
10	11-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache, jaundice, kidney_stones, obesity, stroke	7.16 MB
11	12-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache, jaundice, kidney_stones, obesity, stroke, thyroid_cancer	7.87 MB
12	13-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache, jaundice, kidney_stones, obesity, stroke, thyroid_cancer, SARS	8.62 MB
13	14-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache, jaundice, kidney_stones, obesity, stroke, thyroid_cancer, SARS, rabies	9.04 MB
14	15-T	blood_pressure, bone_density, common_cold, AIDS, Dengue, diarrhea, headache, jaundice, kidney_stones, obesity, stroke, thyroid_cancer, SARS, rabies, corona	9.84 MB
TREC DATA—2018			
1	2-T	Liposarcoma, Meningioma	2.45 MB
2	3-T	Liposarcoma, Meningioma, Breast cancer,	3.01 MB
3	4-T	Liposarcoma, Meningioma, Breast cancer, Melanoma	2.35 MB
4	5-T	Liposarcoma, Meningioma, Breast cancer, Melanoma, Ampullary carcinoma	2.98 MB

information (NMI) [20], precision [21], and recall [21]. These values are given in Tables 2, 3,4, and 5, respectively.

From the crisp partitions, the data object labels are predicted, and the performance of visual health cluster models is evaluated based on the matching the predicted cluster labels and ground truth labels using CA, NMI, precision, and recall.

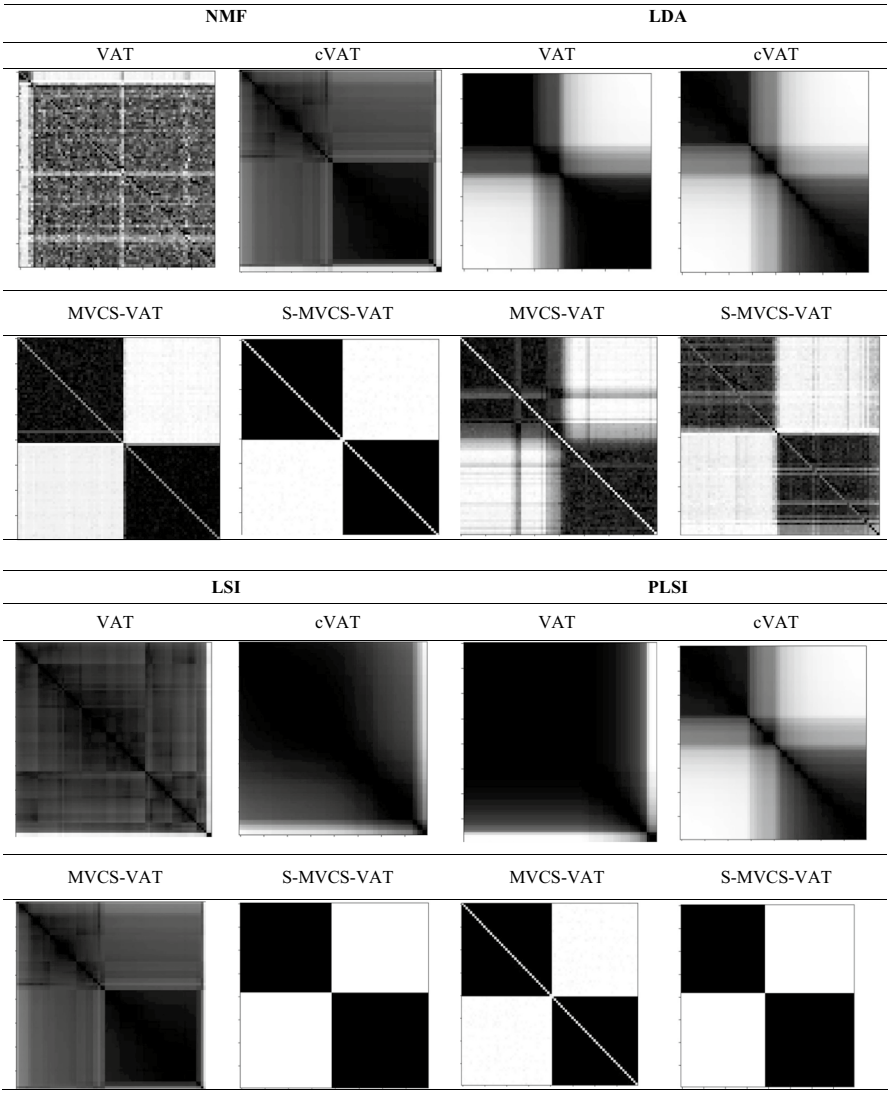

Fig. 2 Visual health data clustering results for big social health data (2 Topics)

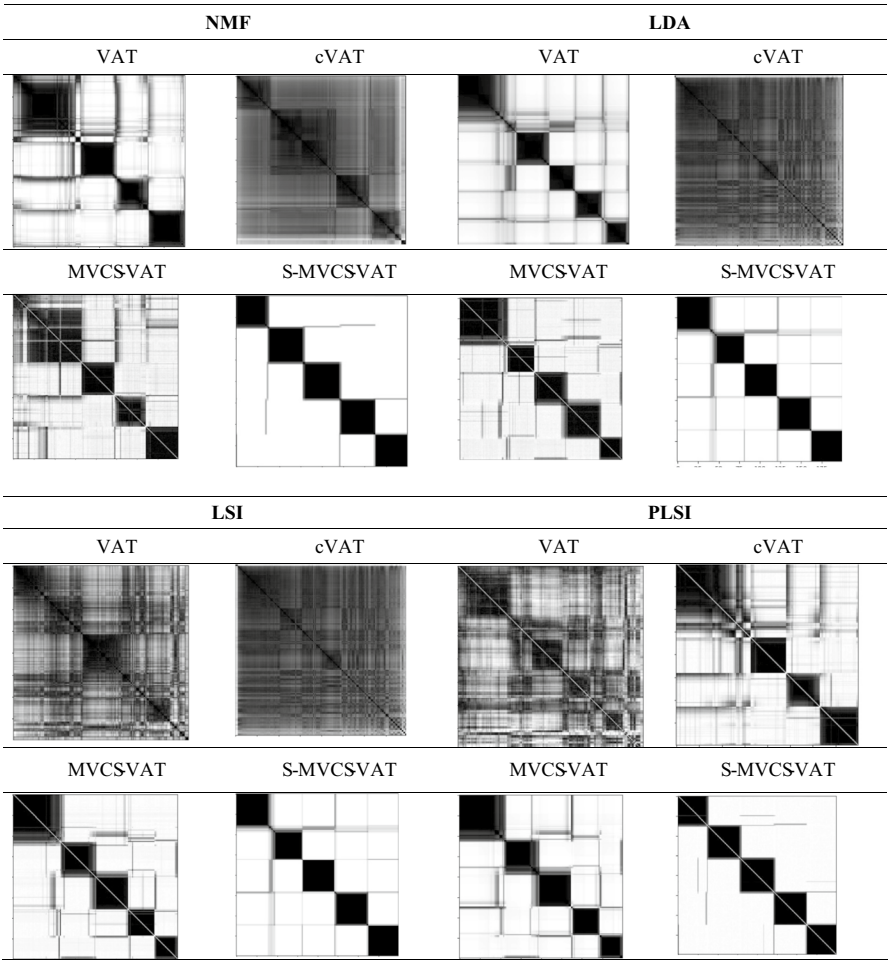

Fig. 3 Visual health data clustering results for big social health data (5 Topics)

4.1 Critical observations

The proposed method used the sample viewpoints only to assess the cluster tendency and data clustering results. Thus, the proposed method is faster method than the MCS-VAT. Crisp partition images with the best clarity and goodness-of-fit occur when using the proposed method. The proposed work is able to discover the quality of large social health data clustering results.

Table 6 presents the goodness-of-fit of the existing and proposed visual images and shows that S-MVCS-VAT scored higher than the other methods underlying the four topic models.

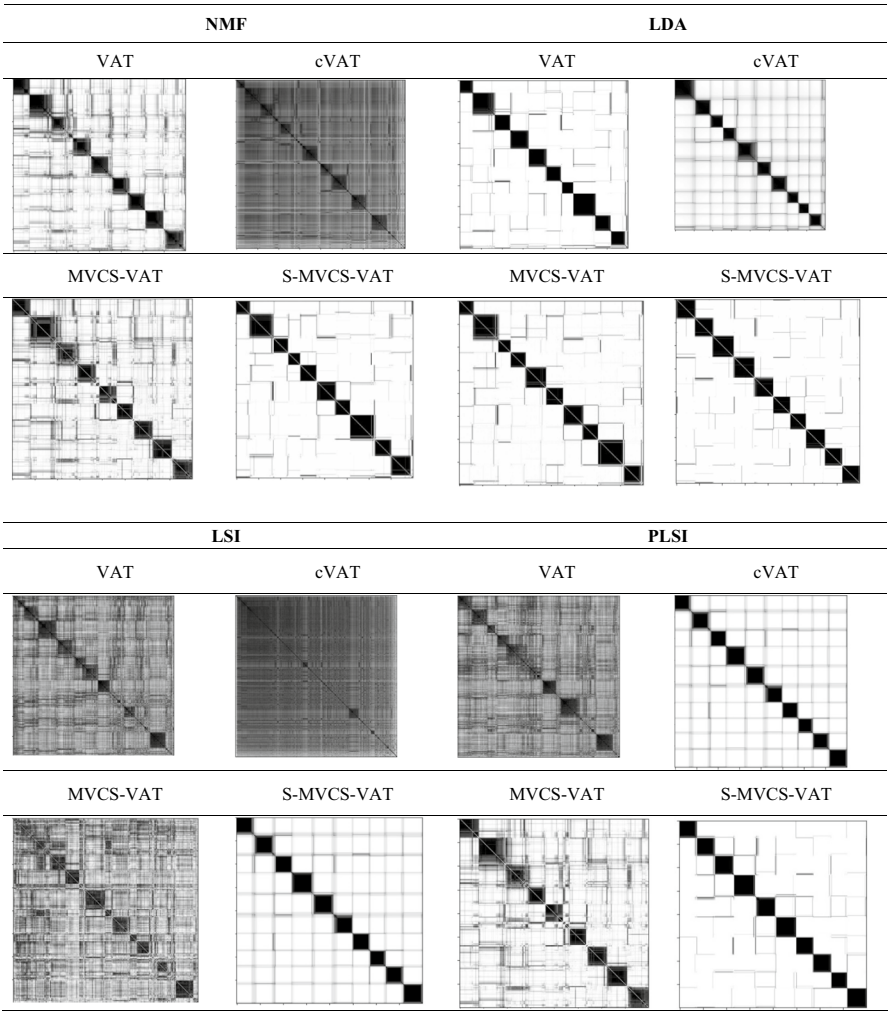

Fig. 4 Visual health data clustering results for big social health data. (10 Topics)

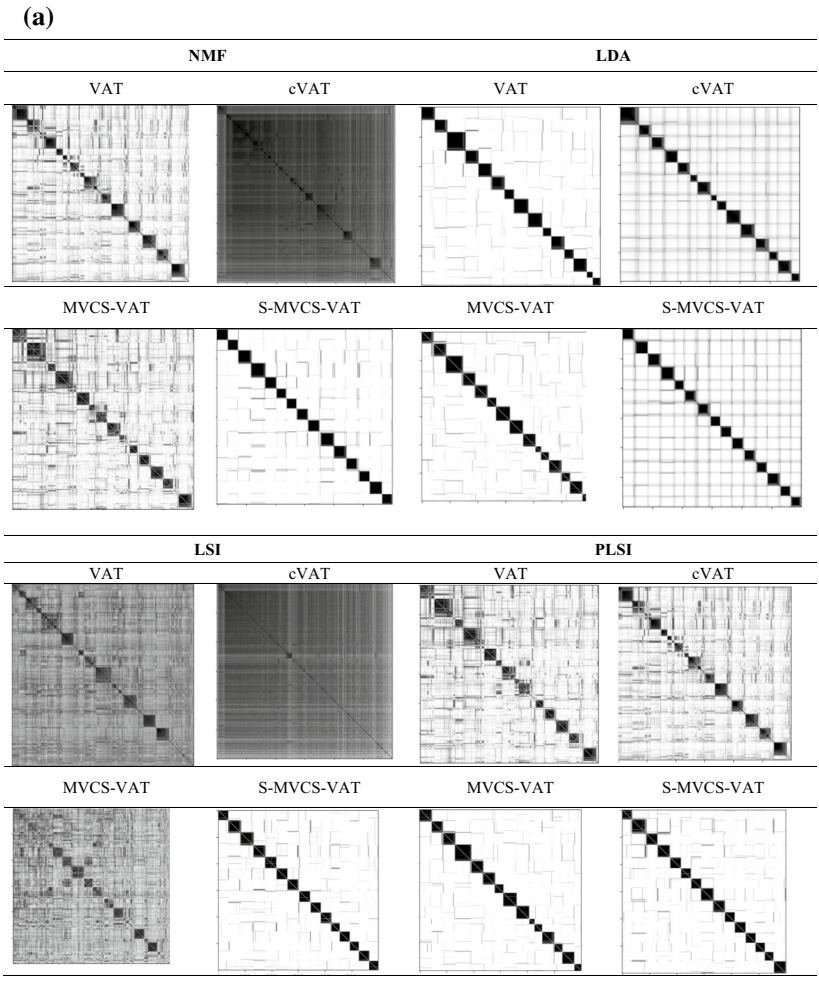

Fig. 5 **a** Visual health data clustering results for big social health data. (15 Topics) **b** Crisp partitions for three data topics

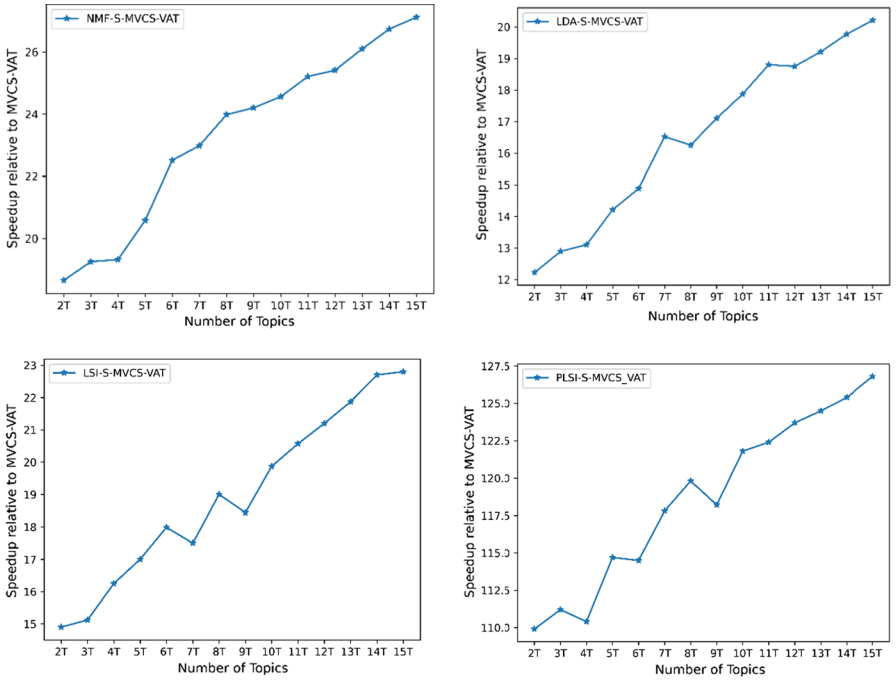

Fig. 6 Speed parameter analysis of visual social health data clustering models compared with the MVCS-VAT

The overall experimental analysis shows that the accuracy was improved at a rate of 5 to 10% in the proposed S-MVCS-VAT method underlying the four topic models NMF, LDA, LSI, and PLSI for big social health data.

5 Conclusion and future scope

Health data assessment is an emerging need in society. Twitter is one of the enriched social sources for people to exchange views or opinions on any topic. Big social data are extracted through Twitter using lakhs of tweets. For the lakhs of tweets, it is most expensive to find social health data clusters. The recent visual technique, the MVCS-VAT, effectively conducts social health data cluster assessment with n-2 multiple viewpoints. The proposed work uses an efficient sampling strategy and four topic models to enhance the MVCS-VAT. Experimental is carried out on 18 different

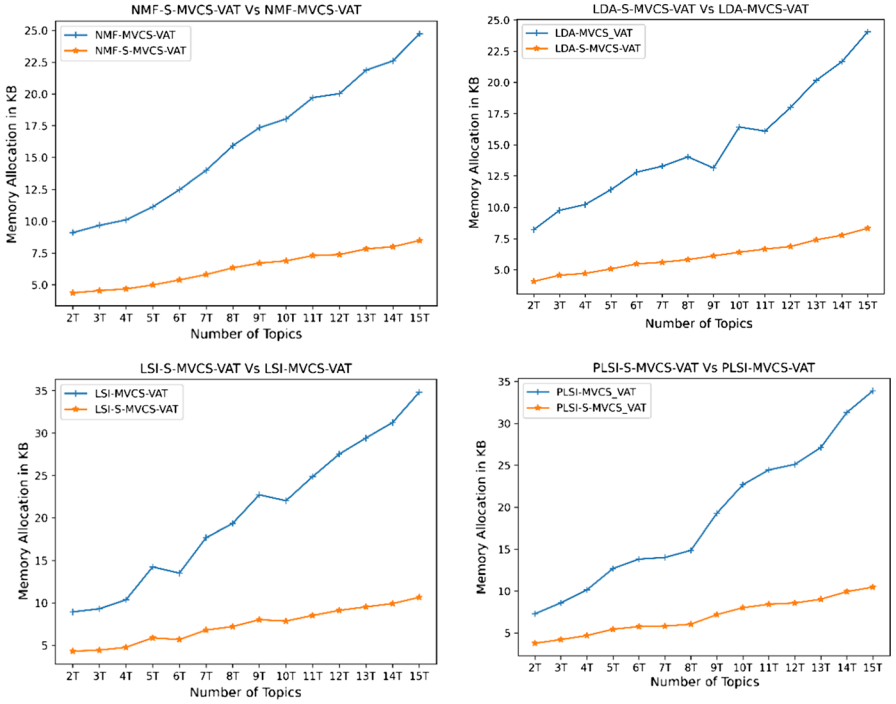

Fig. 7 Memory space analysis of visual social health data clustering models (S-MVCS-VAT vs. MVCS-VAT)

case studies, i.e., 18 different subsets of health datasets. Overall observation of these experimental states that proposed S-MVCS-VAT improves the quality of social health data clusters with significant rate of 5 to 10%. Goodness-of-fit images for the visual clusters are much improved in S-MVCS-VAT for all these datasets. Two scalable parameters, i.e., computational time and memory, are calculated for the proposed S-MVSC-VAT and existing MVCS-VAT underlying with different topic models for all 18 case studies (i.e., 2 topics to 15 topics; 2 topics to 5 topics in TREC 2018) carried in the experimental work. It proved that the proposed S-MVCS-VAT is more scalable with respect to computational time and memory allocation. Future work can be extended to develop scalable ailment visual techniques for health analysis and socially recommended solutions.

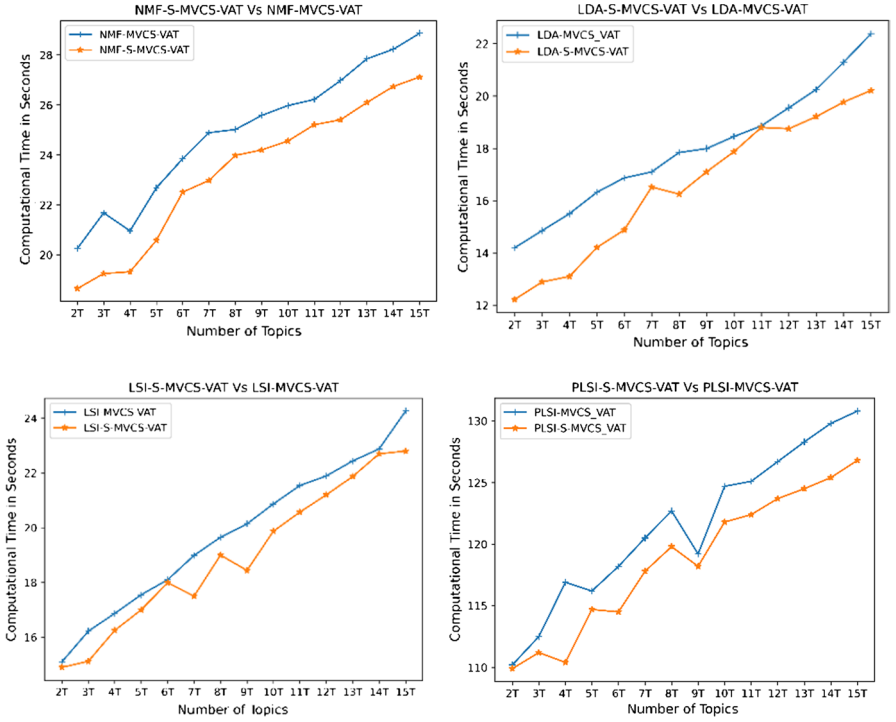

Fig. 8 Time analysis of visual social health data clustering models (S-MVCS-VAT vs. MVCS-VAT)

Table 2 Cluster Accuracy (CA) for the visual health data cluster models

Number of Topics for the Dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT
2	1.000	1.000	1.000	1.000	0.600	0.600	0.620	0.650	0.750	0.650	0.755	0.762	0.575	0.575	0.575	0.575
3	1.000	1.000	1.000	1.000	0.433	0.442	0.450	0.459	0.592	0.633	0.642	0.651	0.425	0.408	0.431	0.435
4	0.987	0.985	0.991	1.000	0.388	0.388	0.395	0.405	0.694	0.481	0.710	0.715	0.369	0.369	0.369	0.371
5	0.985	0.985	0.989	0.989	0.305	0.290	0.307	0.314	0.525	0.405	0.531	0.541	0.300	0.310	0.312	0.316
6	0.904	0.904	0.933	0.945	0.321	0.296	0.325	0.335	0.467	0.388	0.475	0.487	0.263	0.263	0.263	0.271
7	0.782	0.782	0.795	0.812	0.314	0.275	0.320	0.329	0.457	0.368	0.461	0.471	0.250	0.261	0.261	0.268
8	0.716	0.716	0.725	0.755	0.253	0.244	0.261	0.281	0.438	0.303	0.441	0.451	0.259	0.259	0.262	0.271
9	0.817	0.817	0.850	0.865	0.211	0.214	0.225	0.235	0.469	0.292	0.481	0.495	0.233	0.233	0.242	0.246
10	0.700	0.700	0.713	0.722	0.215	0.228	0.235	0.241	0.448	0.270	0.451	0.462	0.223	0.213	0.229	0.235
11	0.520	0.520	0.614	0.625	0.191	0.164	0.210	0.216	0.382	0.257	0.389	0.395	0.182	0.189	0.192	0.199
12	0.646	0.646	0.658	0.678	0.204	0.183	0.211	0.217	0.383	0.250	0.435	0.441	0.206	0.198	0.210	0.217
13	0.500	0.500	0.510	0.521	0.171	0.169	0.178	0.201	0.346	0.260	0.348	0.348	0.183	0.190	0.195	0.204
14	0.418	0.418	0.420	0.431	0.196	0.177	0.205	0.215	0.421	0.257	0.431	0.435	0.168	0.177	0.178	0.185
15	0.462	0.462	0.497	0.505	0.182	0.165	0.192	0.201	0.352	0.250	0.358	0.361	0.187	0.170	0.189	0.198
TREC-2	0.965	0.977	1.000	1.000	1.000	1.000	1.000	1.000	0.965	1.000	1.000	1.000	0.711	0.745	0.768	0.775
TREC-3	0.889	0.907	0.912	0.925	0.973	1.000	1.000	1.000	0.973	1.000	1.000	1.000	0.478	0.473	0.488	0.498
TREC-4	0.758	0.772	0.781	0.798	0.850	0.855	0.861	0.872	0.894	0.905	0.925	0.934	0.389	0.457	0.488	0.495
TREC-5	0.698	0.701	0.701	0.711	0.756	0.742	0.762	0.771	0.825	0.850	0.868	0.875	0.398	0.468	0.487	0.494

Table 3 Normalized mutual information (NMI) for the visual health data cluster models

Number of Topics for the Dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT
2	1.000	1.000	1.000	1.000	0.029	0.029	0.035	0.041	0.189	0.066	0.194	0.199	0.016	0.016	0.017	0.025
3	1.000	1.000	1.000	1.000	0.057	0.057	0.062	0.068	0.323	0.308	0.331	0.338	0.019	0.012	0.025	0.031
4	0.958	0.916	0.961	0.975	0.093	0.093	0.109	0.114	0.425	0.183	0.431	0.438	0.082	0.075	0.082	0.091
5	0.956	0.956	0.962	0.962	0.047	0.044	0.047	0.052	0.306	0.163	0.310	0.317	0.075	0.082	0.078	0.084
6	0.789	0.789	0.849	0.852	0.110	0.098	0.115	0.119	0.301	0.253	0.305	0.311	0.048	0.048	0.050	0.059
7	0.706	0.706	0.712	0.716	0.110	0.098	0.120	0.120	0.292	0.214	0.331	0.338	0.059	0.082	0.092	0.101
8	0.585	0.585	0.592	0.598	0.084	0.092	0.092	0.098	0.300	0.171	0.313	0.320	0.107	0.103	0.110	0.118
9	0.779	0.779	0.785	0.791	0.092	0.091	0.097	0.105	0.339	0.216	0.345	0.349	0.088	0.087	0.098	0.102
10	0.628	0.628	0.661	0.674	0.094	0.099	0.098	0.105	0.371	0.181	0.375	0.375	0.096	0.083	0.098	0.102
11	0.544	0.544	0.556	0.559	0.082	0.061	0.085	0.095	0.322	0.181	0.325	0.331	0.083	0.086	0.087	0.097
12	0.573	0.573	0.582	0.591	0.120	0.096	0.125	0.129	0.367	0.199	0.399	0.409	0.114	0.098	0.121	0.128
13	0.494	0.494	0.504	0.510	0.106	0.090	0.110	0.119	0.337	0.209	0.341	0.347	0.115	0.114	0.121	0.128
14	0.422	0.422	0.432	0.439	0.135	0.101	0.141	0.149	0.352	0.223	0.355	0.361	0.112	0.115	0.115	0.121
15	0.452	0.452	0.469	0.475	0.133	0.102	0.135	0.141	0.357	0.223	0.361	0.368	0.129	0.125	0.131	0.138
TREC-2	0.833	0.845	1.000	1.000	1.000	1.000	1.000	1.000	0.841	1.000	1.000	1.000	0.129	0.179	0.216	0.221
TREC-3	0.679	0.707	0.689	0.694	0.911	1.000	1.000	1.000	0.919	1.000	1.000	1.000	0.076	0.081	0.098	0.105
TREC-4	0.658	0.662	0.678	0.685	0.841	0.849	0.856	0.862	0.784	0.794	0.794	0.810	0.280	0.291	0.310	0.321
TREC-5	0.598	0.611	0.621	0.635	0.749	0.752	0.752	0.768	0.689	0.691	0.715	0.725	0.274	0.281	0.289	0.315

Table 4 Precision (P) for the visual health data cluster models

Number of Topics for the Dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT
2	0.803	0.805	0.810	0.816	0.575	0.579	0.581	0.587	0.652	0.653	0.682	0.688	0.521	0.532	0.558	0.561
3	1.000	1.000	1.000	1.000	0.579	0.821	0.531	0.537	0.721	0.701	0.728	0.735	0.42	0.358	0.378	0.385
4	1.000	1.000	1.000	1.000	0.412	0.421	0.426	0.431	0.621	0.614	0.635	0.641	0.25	0.347	0.389	0.394
5	0.765	0.769	0.821	0.835	0.28	0.285	0.295	0.304	0.602	0.558	0.610	0.618	0.31	0.324	0.368	0.375
6	0.821	0.821	0.832	0.841	0.31	0.299	0.315	0.320	0.555	0.458	0.568	0.574	0.214	0.234	0.287	0.295
7	0.598	0.598	0.745	0.745	0.252	0.261	0.268	0.271	0.441	0.448	0.452	0.458	0.25	0.261	0.289	0.995
8	0.856	0.854	0.918	0.925	0.225	0.226	0.238	0.248	0.451	0.462	0.475	0.482	0.21	0.224	0.249	0.254
9	0.735	0.736	0.741	0.747	0.228	0.221	0.235	0.241	0.442	0.448	0.468	0.474	0.175	0.185	0.212	0.219
10	0.648	0.651	0.678	0.681	0.215	0.21	0.221	0.228	0.418	0.425	0.457	0.461	0.214	0.221	0.245	0.251
11	0.671	0.678	0.689	0.694	0.205	0.214	0.224	0.231	0.525	0.529	0.558	0.564	0.215	0.221	0.251	0.251
12	0.653	0.658	0.668	0.672	0.205	0.204	0.212	0.219	0.507	0.512	0.538	0.541	0.198	0.21	0.242	0.253
13	0.558	0.559	0.568	0.571	0.198	0.181	0.191	0.199	0.432	0.438	0.452	0.462	0.178	0.185	0.214	0.219
14	0.498	0.499	0.521	0.527	0.165	0.171	0.176	0.184	0.415	0.428	0.448	0.453	0.207	0.214	0.248	0.255
15	0.458	0.459	0.471	0.476	0.164	0.162	0.169	0.178	0.508	0.512	0.539	0.547	0.168	0.174	0.182	0.182
TREC-2	0.743	0.745	0.758	0.768	1.000	1.000	1.000	1.000	0.952	1.000	1.000	1.000	0.132	0.141	0.158	0.168
TREC-3	0.659	0.647	0.668	0.672	0.952	0.958	1.000	1.000	0.959	1.000	1.000	1.000	0.052	0.065	0.081	0.098
TREC-4	0.589	0.591	0.599	0.615	0.852	0.861	0.872	0.872	0.812	0.816	0.818	0.818	0.260	0.265	0.275	0.295
TREC-5	0.574	0.579	0.598	0.610	0.810	0.819	0.829	0.829	0.789	0.795	0.810	0.818	0.247	0.257	0.268	0.275

Table 5 Recall (R) for the visual health data cluster models

Number of Topics for the Dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT
2	0.802	0.803	0.810	0.817	0.581	0.585	0.591	0.598	0.662	0.668	0.672	0.677	0.512	0.515	0.534	0.541
3	1.000	1.000	1.000	1.000	0.585	0.856	0.861	0.871	0.752	0.698	0.758	0.762	0.352	0.359	0.378	0.381
4	1.000	1.000	1.000	1.000	0.415	0.419	0.425	0.425	0.635	0.710	0.715	0.721	0.352	0.350	0.374	0.380
5	0.768	0.762	0.841	0.849	0.281	0.291	0.312	0.318	0.605	0.512	0.612	0.619	0.305	0.312	0.338	0.342
6	0.819	0.821	0.832	0.837	0.308	0.295	0.311	0.319	0.578	0.395	0.581	0.588	0.220	0.234	0.258	0.268
7	0.595	0.595	0.741	0.748	0.251	0.258	0.268	0.278	0.459	0.462	0.478	0.482	0.250	0.261	0.289	0.295
8	0.851	0.852	0.912	0.918	0.220	0.221	0.241	0.249	0.462	0.465	0.472	0.482	0.201	0.211	0.238	0.247
9	0.731	0.735	0.739	0.748	0.221	0.218	0.250	0.258	0.441	0.421	0.482	0.491	0.185	0.195	0.228	0.235
10	0.642	0.649	0.672	0.680	0.210	0.198	0.212	0.219	0.421	0.408	0.429	0.435	0.214	0.224	0.258	0.268
11	0.669	0.667	0.682	0.689	0.207	0.204	0.212	0.219	0.540	0.512	0.552	0.561	0.210	0.225	0.249	0.258
12	0.652	0.659	0.662	0.668	0.207	0.198	0.214	0.218	0.510	0.512	0.525	0.535	0.185	0.195	0.225	0.235
13	0.551	0.552	0.562	0.567	0.175	0.169	0.179	0.185	0.449	0.425	0.458	0.461	0.168	0.171	0.210	0.218
14	0.491	0.491	0.528	0.535	0.162	0.168	0.175	0.185	0.465	0.425	0.475	0.481	0.210	0.221	0.252	0.261
15	0.452	0.452	0.462	0.470	0.165	0.159	0.172	0.189	0.510	0.508	0.521	0.529	0.157	0.168	0.195	0.201
TREC-2	0.732	0.739	0.741	0.758	1.000	1.000	1.000	1.000	0.941	0.958	0.968	0.974	0.087	0.087	0.091	0.099
TREC-3	0.624	0.631	0.644	0.661	0.912	0.921	0.928	0.941	0.947	0.958	0.968	0.978	0.029	0.039	0.051	0.078
TREC-4	0.598	0.610	0.610	0.621	0.621	0.657	0.698	0.698	0.941	0.948	0.962	0.978	0.045	0.052	0.068	0.075
TREC-5	0.546	0.559	0.568	0.574	0.598	0.601	0.612	0.625	0.910	0.918	0.925	0.935	0.058	0.068	0.068	0.074

Table 6 Goodness-of-fit of the visual Images

Number of Topics for the Dataset	NMF				LDA				LSI				PLSI			
	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT	VAT	cVAT	MVCS-VAT	S-MVCS-VAT
2	0.21	0.245	0.451	0.471	0.412	0.432	0.438	0.478	0.202	0.198	0.289	0.525	0.217	0.198	0.487	0.545
5	0.31	0.335	0.342	0.425	0.31	0.347	0.349	0.41	0.411	0.415	0.452	0.513	0.415	0.446	0.487	0.525
10	0.51	0.515	0.535	0.575	0.558	0.564	0.578	0.654	0.61	0.624	0.658	0.724	0.625	0.639	0.61	0.689
15	0.605	0.525	0.61	0.715	0.685	0.695	0.71	0.725	0.525	0.541	0.51	0.714	0.31	0.315	0.425	0.555

References

1. Bezdek JC, Hathaway RJ (2002) VAT: a tool for visual assessment of (cluster) tendency. In: Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02, 2002, p 2225–2230
2. Vijeya Kaveri V, Maheswari V (2019) A framework for recommending health-related topics based on topic modeling in conversational data (Twitter). Cluster Computing.
3. Narasimhulu K, Meena AbarnaSivakumar KTB (2021) An enhanced cosine-based visual technique for the robust tweets data clustering. Int J Intell Comp Cybern. 14(2):170–184. https://doi.org/10.1108/IJICC-10-2020-0151
4. Lee D, Seung H (2000) Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing SYSTEMS 13, NIPS, Denver, CO, USA p 556–562
5. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407
6. Hofmann T (1999) Probabilistic latent semantic indexing. SIGIR. ACM, New York, pp 50–57
7. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
8. Wuhan (2018) TF-IDF based feature words extraction and topic modeling for short text. In: ICMSS2018
9. Wu X, Kumar V, Quinlan JR et al (2008) Top 10 algorithms in data mining, knowledge information system, vol 14. Springer, Heidelberg, pp 1–37
10. Suleman Basha M, Mouleeswaran SK, Rajendra Prasad K (2019) Cluster tendency methods for visualizing the data partitions. Int J Innovative Technol Explor Eng (IJTEE). https://doi.org/10.35940/ijitee.K2285.0981119
11. J. Wang and X. Su (2011) An improved K-Means clustering algorithm. In: IEEE 3rd International Conference on Communication Software and Networks, p. 44–46. https://doi.org/10.1109/ICCSN.2011.6014384.
12. Rajendra Prasad K, Mohammed M, Noorullah RM (2019) Hybrid topic cluster models for social healthcare data. Int J Adv Comput Sci Appl 10(11):490–506. https://doi.org/10.14569/IJACSA.2019.0101168
13. Suleman Basha M, Mouleeswaran SK, Prasad KR (2021) Sampling-based visual assessment computing techniques for an efficient social data clustering. J Supercomp. 77:8013–8037. https://doi.org/10.1007/s11227-021-03618-6
14. Kumar D, Bezdek JC, Palaniswami M, Rajasegarar S, Leckie C, Havens TC (2016) A hybrid approach to clustering in big data. IEEE Trans Cybern 46(10):2372–2385
15. Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) A comparison study on similarity and dissimilarity measures in clustering continuous data. PLoS 10(12):1–20
16. https://trec.nist.gov/data/web2014.html
17. https://trec.nist.gov/data/microblog2015.h
18. https://www.webmd.com/
19. Pattanodom et al. (2016) Clustering data with the presence of missing values by ensemble approach. In: Second Asian Conference on Defense Technology
20. Alessia Amelio, Clara Pizzuti (2015) Is normalized mutual information a fair measure for comparing community detection methods. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining
21. Bhatnagar V, Majhi R, Jena PR (2018) Comparative performance evaluation of clustering algorithms for grouping manufacturing frms. Arab J Sci Eng 43:4071–4083
22. Rajendra Prasad K, Mohammed M, Noorullah RM (2019) Visual topic models for healthcare data clustering. Evol Intel. https://doi.org/10.1007/s12065-019-00300-y
23. Basha S (2020) comparison of real datasets characteristics by using clustering approaches. J mech cont math sci. https://doi.org/10.26782/jmcms.2020.08.00061
24. Todd Gamblin, Bronis R.de Supinski, Martin Schulz, Rob Fowler, Danier A. Reed, (2010) Clustering performance data efficiently at massive scales. In: ICS '10 Proceedings of the 24th ACM International Conference on Supercomputing, p 243–252. https://doi.org/10.1145/1810085.1810119
25. Surya Bhupal Rao, S. Rahamat Basha, G. Ravi Kumar (2020) A comparative approach of text mining: classification, clustering and extraction techniques. J Mech Continua Math Sci. (5)120–131
26. Shafqat S, Kishwer S, Rasool RU et al (2020) Big data analytics enhanced healthcare systems: a review. J Supercomput 76:1754–1799. https://doi.org/10.1007/s11227-017-2222-4

27. Vidhya K, Shanmugalakshmi R (2020) Modified adaptive neuro-fuzzy inference system (M-ANFIS) based multi-disease analysis of healthcare Big Data. J Supercomput 76:8657–8678. https://doi.org/10.1007/s11227-019-03132-w
28. Hashimoto T, Shepard DL, Kuboyama T et al (2021) Analyzing temporal patterns of topic diversity using graph clustering. J Supercomput 77:4375–4388. https://doi.org/10.1007/s11227-020-03433-5
29. AlZubi AA (2020) Big data analytic diabetics using map reduce and classification techniques. J Supercomput 76:4328–4337. https://doi.org/10.1007/s11227-018-2362-1
30. Doghri W, Saddoud A, Chaari Fourati L (2021) Cyber-physical systems for structural health monitoring: sensing technologies and intelligent computing. J Supercomput. https://doi.org/10.1007/s11227-021-03875-5
31. Krishnaraj N, Bellam K (2020) Improved Distributed Frameworks to Incorporate Big Data through Deep Learning. Journal of Advanced Research in Dynamical & Control Systems 12:332–338

Authors and Affiliations

K. Narasimhulu[1] · K. T. Meena Abarna[1] · B. Siva Kumar[1,2] · T. Suresh[1]

K. T. Meena Abarna
abarnakt@yahoo.com

B. Siva Kumar
sivaapec@gmail.com

T. Suresh
suresh_raman06@yahoo.co.in

[1] Annamalai University, Chidambaram, Tamilnadu, India

[2] Department of CSE, Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhra Pradesh, India