An effective assessment of cluster tendency through sampling based multi-viewpoints visual method

K. Rajendra Prasad¹ · B. Eswara Reddy² · Moulana Mohammed³

Received: 15 September 2020 / Accepted: 18 November 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Social networks are the rich sources to people for sharing the knowledge on health-related issues. Nowadays, Twitter is one of the great significant social platforms to the people for a discussion on topics. Analyzing the clusters for the tweets concerning terms is a complex process due to the sparsity problem. Topic models are useful or avoiding this problem with derivations of topic clusters. Finding pre-cluster tendency is the major problem in many clustering methods. Existing methods, such as visual access tendency (VAT), cosine-based VAT (cVAT), multi viewpoints-based cosine similarity VAT (MVS-VAT) majorly used to access the prior information about clusters tendency problem. Solution of cluster tendency indicates the tractable number of clusters. The MVS-VAT enables the cluster tendency for the tweet documents effectively than other visual methods. However, it takes a higher number of viewpoints, thus requiring more computational time for the clustering of tweets data. Therefore, sampling-based visual methods are proposed to overcome the computational problem. Several standard health keywords are used for the extraction of health tweets to illustrate the effectiveness of proposed work in the experimental study.

Keywords Cluster tendency · Tweets data clustering · VAT · cVAT · MVS-VAT · Sampling

1 Introduction

Essential techniques of data clustering problem widely used for organizing the data into meaningful partitions based on similarity features of data objects. Relevant tweet data clustering problem is applied among the tweets (or data objects), in which the labels of tweets are initially unknown. Well-known topic clustering techniques (Xu and Wunsch 2005); (Yi et al. 2020); (Jose and Babu 2019) describes the

modelling of tweets and generation of tweets features in the form of bag-of-words concerning topics rather than terms (Pessiot et al. 2010). Similarity features among the tweets are measured using distance measures, like Euclidean and cosine (Shirkhorshidi et al. 2015). The VAT (Bezdek and Hathaway 2002) and cVAT (Prasad et al. 2019b, a) use the Euclidean and cosine measures while evaluating the clustering tendency of tweets data. Cosine accurately measures the similarity features for the text documents (Oghbaie and Zanjireh 2018) than Euclidean. Thus, for the assessment of clusters, cVAT performs better than VAT. In cVAT, cosine similarity is computed among the tweets with a single viewpoint. MVS-VAT (Prasad et al. 2019b, a) is a pre-cluster assessment technique, which computes the cosine similarity among the tweets using multi viewpoints.

It is a more potential technique in the assessment of clusters than VAT and cVAT (Therese and Lingam 2017). The MVS-VAT computes the similarity features among two tweet documents for the set of N tweet documents is the average similarity concerning (N-2) viewpoints. Thus, it is computationally expensive than other visual methods, for reducing the computational requirements, the sampling technique proposed for the selection of a few viewpoints rather than the

✉ Moulana Mohammed
moulanaaphd@gmail.com

K. Rajendra Prasad
krprgm@gmail.com

B. Eswara Reddy
eswar.cse@jntua.ac.in

¹ Department of CSE, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Andhra Pradesh, India

² Department of CSE, JNTUA College of Engineering, Anantapur, Andhra Pradesh, India

³ Department of CSE, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India

selection of $(N-2)$ viewpoints. The sample viewpoints are selected from inter clusters of data objects, i.e., cosine similarity between two tweets (t_1, t_2) is measured with respect to sample viewpoints, and the selection of viewpoints do not belong to clusters of either t_1 or t_2 , and it should be other than these clusters. This selection depends on the inter-cluster viewpoints of data objects; therefore, it is known as sampling inter-cluster viewpoints (SICV). Its extensive sampling MVS-VAT is known as SICV-MVS-VAT. The processing steps of proposed SICV-MVS-VAT are illustrated in Fig. 1. After pre-processing the tweets documents, cosine similarity features for every pair of tweets document are computed with the selection of sample viewpoints.

Highlights of the contributions are described in following:

1. Develop the sampling technique for the selection of inter-cluster viewpoints of data objects
2. Estimate the clustering tendency for the tweet's dataset with the selection of sample viewpoints
3. Effective visualization for cluster images of tweets documents
4. Develop the SICV-MVS-VAT for the best visual assessment of cluster of the social data (or tweets)
5. Derive the crisp parts of the visual cluster for finding the cluster labels of tweets objects in order to find the generation of social data clusters

Fig. 1 Processing Steps of Proposed SICV-MVS-VAT

Other sections of the paper are described as follows: Sect. 2 presents the related visual methods Sect. 3 describes the proposed sampling multi-viewpoint-based visual method, Sect. 4 discusses the experimental study and performance discussion, and Sect. 5 presents the conclusion and scope.

2 Related visual methods

For the tweet data clustering (Sechelea et al. 2016); (Rehioui and Idrissi 2019), topic models (Amelio and Pizzuti 2015; Xu et al. 2019; Hu et al. 2012; Ismail et al. 2018) are required for modelling the tweets with respect to topics instead of terms for avoiding the problem of data sparsity. Topic models, i.e. non-matrix factorization (NMF) (Lee and Seung 2001), Latent Dirichlet Allocation (LDA) (Blei et al. 2003), Latent Semantic Indexing (LSI) (Deerwester et al. 1990), and Probabilistic Latent Semantic Indexing (PLSI) (Hofmann 1999) are widely used for the modelling and extraction of tweets features in bag-of-words (Wallach 2006) format. Each tweet document features are considered as a single data object. The data objects i.e. tweet feature are used for the clustering tweet documents. Clustering of text documents (or tweets documents) is performed through TF-IDF analysis (Neogi et al. 2020). However, such analysis is suffered with data sparsity problem. Topic models are well suited for performing of text clustering based on topics derivations instead of deriving the TF-IDF. It less sensitive to data sparsity problem when compared to TF-IDF based clustering approaches. Topic models are recommended for big text data clustering problems (Sukhija et al. 2016). The LDA is the probabilistic topic model; which enables the text data clustering results in an unsupervised approaches for large set of text documents. The LDA is a generative topic model in which each document represents the mixture of

word probabilities of certain topics (Ramage et al. 2009). NMF model is optimistic model when compared to other topic models. It has the capable of extracting the documents information in terms of valid topics without knowing any prior information (Suri and Roy 2017). Key objective of NMF is to decompose the input matrix (document-term matrix) into two matrices, namely, document-topic matrix and topic term matrix. So that it perfectly extracts the topics information for the set of documents.

Cluster tendency is one the crucial problem in data clustering (Tang et al. 2011; Suleman Basha et al. 2019). (Pushpalatha et al. 2020; Devisetty et al. 2019) of text documents. Many pre-cluster tendency methods are surveyed (Mahallati et al. 2019; Hu and Hathaway 2008; Park et al. 2016; Kumar et al. 2015; Kumar and Bezdek 2020; Varish et al. 2020) for finding the cluster tendency. It is recognized that VAT and iVAT (Havens and Bezdek 2011) are the best choices for finding the assessment of cluster tendency, whereas VAT and iVAT are suited for normal and path-shaped datasets respectively. They are initially measuring the dissimilarity features using distance metrics (Blei et al. 2003) and re-order the indices of data objects (or tweets) in order to find their visual images as shown in Fig. 2.

It illustrates the clusters as visual dark black coloured blocks in the images (Srinivas 2018) of corresponding VAT. Cluster tendency (or clusters prior information) determined with the count of dark black coloured blocks obtained in the diagonal of visual images. Based on spectral features, the affinity values are computed for dissimilarity features of data objects).

Both VAT and cVAT are procedurally same, however, they compute the dissimilarity features with Euclidean and cosine respectively. Cosine takes the distance among the data objects based on magnitude and directions, whereas in an Euclidean takes distances based on differences among two data object vectors. Thus, cVAT estimates the good

$$DM = \begin{bmatrix} 0.00 & 0.73 & 0.19 & 0.71 & 0.16 \\ 0.73 & 0.00 & 0.59 & 0.12 & 0.78 \\ 0.19 & 0.59 & 0.00 & 0.55 & 0.19 \\ 0.71 & 0.12 & 0.55 & 0.00 & 0.74 \\ 0.16 & 0.78 & 0.19 & 0.74 & 0.00 \end{bmatrix} \quad I = \begin{bmatrix} \text{Irregular Structure of Image} \end{bmatrix}$$

(a) Prior to Apply VAT- Distance (or Called as Dissimilarity) Matrix – DM and its Irregular Structure of Image

$$RDM = \begin{bmatrix} 0.00 & 0.12 & 0.59 & 0.73 & 0.78 \\ 0.12 & 0.00 & 0.59 & 0.71 & 0.74 \\ 0.59 & 0.55 & 0.00 & 0.19 & 0.19 \\ 0.73 & 0.71 & 0.19 & 0.00 & 0.16 \\ 0.78 & 0.74 & 0.19 & 0.16 & 0.00 \end{bmatrix} \quad I = \begin{bmatrix} \text{Regularized Clustering Image Structure} \end{bmatrix}$$

(b) After Applying VAT, Re-ordered Distance Matrix (RDM), and its Regularized Clustering Image Structure

Fig. 2 Visual Method—VAT- Visual Images

assessment of clusters with cosine based dissimilarities when compared to VAT.

Spectral features-based VAT is known as SpecVAT (Wang et al. 2008). SpecVAT finds the affinity matrix for the set of data objects and then Laplacian matrix is derived for determining the spectral (or Eigen features) of data objects. Cluster tendency is accessed from the VAT image of spectral features. It also shows the clusters effectively in the visual image for the set of data objects. Cosine measure is more effective than Euclidean for measuring the distance data objects in text data clustering (Nguyen et al. 2010). It computes the cosine similarity among the objects with reference of a single viewpoint. cVAT use the cosine measure for the assessment of cluster tendency. It justifies the cluster tendency using a single viewpoint only. Recently, multi viewpoints-based cosine similarity is developed and it is MVS-VAT (Reddy and Prasad 2016). It justifies the cluster tendency using more than two viewpoints unlike cVAT, it also produces the efficient data clustering results for the tweet datasets. MVS-VAT effectively utilized for the different topics of tweet datasets in (Prasad et al. 2019b, a) for the experimental of health data clustering results. It suffers with the problem of high computational time

for the reason of spent much time in the measuring of similarity among the tweets with respect to (N-2) viewpoints, here N is the total number of viewpoints. Based on the observation of state-of-the-art of visual methods, it needs be development of sampling technique for the selection of few viewpoints rather than selection of (N-2) viewpoints. Thus, sampling based visual method needs to be developed to overcome the problem of high computational time. The proposed work is discussed in the following section, which explains the sample procedure for the selection of viewpoints.

3 Proposed sampling based multi viewpoints visual method

Cluster tendency is the major issue for the tweets data clustering. This problem is effectively handled by visual methods, VAT, cVAT, and MVS-VAT. In this paper, sampling strategy is developed with the selection of best samples of inter clusters viewpoints (SICV) in the extended work of MVS-VAT, say as SICV-MVS-VAT method. Algorithm 1 shows the procedural steps of SICV-MVS-VAT.

Algorithm 1 : SICV-MVS-VAT

Input : N number of pre-processed tweets

Output : Sample viewpoints, Cluster Tendency ('k'), tweets data clustering results

Methodology :

1. Model the tweets using topic models with bag-of-words for representing the tweets with respect to topics features, topics features of tweets denoted by $\{T_1, T_2, \dots, T_N\}$
2. Select the random number 'r' from $\{1, 2, \dots, N\}$
3. Find the distances between T_r to $T_1, T_2, \dots, T_n$, and update the values into $D_1, D_2, \dots, D_n$
4. $\max_index = \operatorname{argmax}_i \{D_1, D_2, \dots, D_n\}$, where $\max_index$ denotes the index of first cluster centroid
5. Compute the distance between $T_{\max_index}$ and $T_1, T_2, \dots, T_n$, update the new distances into $ND_1, ND_2, \dots, ND_n$.
6. For $i = 1$ to n
 $D_i = \min \{D_i, ND_i\}$
 Repeat the steps from 4 to 6 until to get the centroids of assumed number of topics.
7. $MVS(D_i, D_j) = \operatorname{avg}(\sum_{vpno=1, vp}^{np\text{sample}} \operatorname{Sim}(Dp1, Dp2))$

$$\operatorname{Sim}(Dp1, Dp2) = \operatorname{Cos}(D_i - vp, D_j - vp)$$

Here vp refers to the sample viewpoints

inter clusters of D_i, D_j , np sample refers to sample size of each inter-cluster.

8. Find the dissimilarity matrix (M) = 1-(normalize (MVS),0,1)
9. Apply VAT on M and reorder the indices of tweet documents and obtain RDM
10. Display - visual image of RDM, and access the cluster tendency 'k' with count of dark- blocks from displayed image
11. Compute the crisp partitions of visual clusters of Step 10 in order to determine the tweets cluster labels for discovering of social tweets data clusters.

Samples of data objects are selected from the inter-cluster data objects. Inter clusters are identified with the selection of most dissimilar data objects (or the centroids of inter clusters); Next, nearest data objects of inter-clusters are identified with the least dissimilar data objects. These inter clusters are generated for the selection of best samples in our proposed work.

Feature extraction of tweets is performed with the derivations of bag-of-features. Topic model aims to derive the tweets feature with respect to topics rather than terms. Thus, problem complexity of data sparsity is handled and simplified with the topics-based feature extraction for wide range of tweets- documents, which process denoted in Step 1, and the features of N tweets are derived as $\{T_1, T_2, \dots, T_N\}$. Pre-estimations of inter clusters are derived using Step 2 to Step 6. These steps describe the most deviated data object based distances, in which random object is selected in Step 2 and most deviated data objects (it is considered as centroid of one of assumed cluster) is selected by finding the data object who maintains the maximum distance and it shown in Step 3 and Step 4. Updates the corresponding distances of data objects with centroid in Step 5. The centroids are excluded for further processing steps and considered the new centroids from remaining data objects and it is performed

by Step 6. Step 7 shows the multi viewpoints-based cosine similarity (MVS) computation procedure with respect to selection of few viewpoints as per the sample size of 'np sample'. MVS is applied for every pair of tweet documents (or data objects), which is (D_i, D_j) , the viewpoints vp should be selected from inter clusters of D_i and D_j ; whereas the information of inter clusters are obtained from the earlier steps. The MVS of data pairs shown the similarity value between respective pair and the value is normalized between 0 and 1; find the dissimilarity matrix (M) with subtraction of normalized similarity value from 1, which shown in Step 8. In Step 9, visual method VAT is applied on M for re-ordered dissimilarity matrix (RDM), which represents that re-alignment of data based on ordering of dissimilarity features. Step 10 displays the image of RDM for visualizing the cluster tendency (' k '). Finally, the cluster labels of tweet documents are computed with the crisp partition matrix of visual clusters shown in Step 11. The crisp partitions are derived with finding of squareness properties of appeared black coloured blocks along the diagonal of VAT image. The squareness property of dark coloured blocks are derived from the difference of pixel values appeared between dark coloured blocks and non-dark coloured blocks. The significance of the proposed work is to assess the excellent value of

Table 1 Description of Tweet Datasets for the Topics 2-Topics to 15-Topics

SNo	Health Dataset in terms of N-T (where as T denotes Topics)	Tweets topics during extraction of datasets	Size of the Data
1	2-T	Aids, blood_pressure	40 Documents
2	3-T	Aids, blood_pressure, bone_density	80 Documents
3	4-T	Aids, blood_pressure, bone_density, common_cold	120 Documents
4	5-T	Aids, blood_pressure, bone_density, common_cold, constipation	160 Documents
5	6-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona	200 Documents
6	7-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue	240 Documents
7	8-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue, diarrhea	280 Documents
8	9-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue, diarrhea, headache	320 Documents
9	10-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue, diarrhea, headache, heart_disease	360 Documents
10	11-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue, diarrhea, headache, heart_disease, jaundice	400 Documents
11	12-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue, diarrhea, headache, heart_disease, jaundice, kidney_stones	440 Documents
12	13-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue, diarrhea, headache, heart_disease, jaundice, kidney_stones, obesity	480 Documents
13	14-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue, diarrhea, headache, heart_disease, jaundice, kidney_stones, obesity, rabies	520 Documents
14	15-T	Aids, blood_pressure, bone_density, common_cold, constipation, corona, dengue, diarrhea, headache, heart_disease, jaundice, kidney_stones, obesity, rabies, sars	560 Documents
15	TREC2015	Paradox side effects, Lyme disease study	80 Documents
16	TREC2017	Cancer and depression, Crohns Disease treatment, H1N2 Bird Flu	120 Documents

Fig. 3 Cluster Tendency Assessment for Visual Methods (TREC2015-Two Topics)

a) VAT Image b) MVS-VAT Image

(c) SICV-MVS-VAT Image

cluster tendency and generate efficient clustering results with the technique of sampling inter-cluster viewpoints (SICV). The efficiency of the proposed methodology (SICV-MVS-VAT) for the tweets data clustering is demonstrated experimentally in the following section.

4 Experimental study and performance discussion

Existing visual methods, VAT, cVAT, MVS-VAT and proposed method SICV-MVS-VAT are experimented on tweet documents, which are extracted based on standard health keywords, TREC 2015 and TREC2017. Data sets

descriptions and performance evaluation of visual methods are presented in this section for demonstrating the efficiency of proposed method in assessment of cluster tendency and its data partitioning results.

4.1 Tweet datasets description

Table 1 shows the tweets description in terms of total number of topics, keywords or health topics. In the experimental, various subsets of tweets are merged as per mentioned keywords. The standard health keywords are used in the experimental, which are trendy or most discussed tweets. Benchmarked TREC2015 and TREC2017 keywords are used for tweets extraction.

(a) VAT Image

(b) MVS-VAT Image

(c) SICV-MVS-VAT Image

(d) Crisp Partitions of SICV-MVS-VAT Image

Fig. 4 Cluster Tendency Assessment for Visual Methods (TREC2017-Three Topics)

4.2 Topic models

Popular topic models, NMF, LDA, LSI, and PLSI, are used to model extracted tweets. During the modelling, bag-of-words of the tweets are extracted to the purpose of feature extraction. The dissimilarity matrix and re-ordered

dissimilarity matrix are computed in the existing and proposed visual methods after the extraction of feature extraction. Exploration of tweets data clustering results are demonstrated in the following sub-section for a comparison study of visual methods for the four topic models.

4.3 Performance discussion

The comparative visualizations for the cluster tendency shown in Figs. 3, 4, 5, and 6 for TREC2015 (two-topic), TREC2017 (three-topic), standard health tweets (10-topics), and standard health tweets (15-topics), respectively. Resulting visual images of three existing methods and proposed SICV-MVS-VAT. The following figures having three variations of visual images, which are derived from VAT, MVS-VAT, and SICV-MVS-VAT methods respectively. Key variation of these images is that visual appearance and clarity of square-shaped black coloured dark blocks along the diagonal for the three methods in Figs. 3 and 4. Based on the experimental of existing VAT and MVS-VAT and SICV-MVS-VAT, it is observed that proposed SICV-MVS-VAT effectively determines the similarity features among the data objects with inter-cluster viewpoints.

Experimental is conducted under different topic models with finding of similarity features using cosine and Euclidean distance metrics. Cosine is based text data clusters assessment proven as the best when compared to Euclidean for any topic model. Experimental results of VAT based topic models, MVS-VAT based topic models, and SICV-MVS-VAT based topic models under Euclidean and cosine metrics are illustrated in Figs. 5 and 6; which results states that cosine based clusters assessment are the best when compared to Euclidean under any topic model and proven that our SICV-MVS-VAT visualizes the best clusters assessment compared to others.

With the sampling of inter-cluster viewpoints, the similarity among the tweets documents is computed. Cluster tendency generates with the count of dark black coloured blocks of visual method images. More clarity of visual clusters indicates proper and accurate social data clusters; for the text tweets, cosine shown the quality of clusters than Euclidean.

This assessment of cluster tendency is performed with a single viewpoint and it is with origin—the proposed methodology used multiple viewpoints during the assessment of cluster tendency. More informative assessment is extracted with multi viewpoints rather than a single viewpoint.

Therefore, SICV-MVS-VAT has shown the best quality of clusters than other visual methods, and the same is observed in Figs. 3 to 6. With the derivations of crisp partitions, the

Fig. 5 Cluster Tendency Assessment for Visual Methods (Standard Health Tweets—Ten Topics)

complete clustering results are obtained. Table 2 shows the sample tweets data. Table 3 shows the performance evaluation of visual methods using the performance measures mentioned in (Datta et al. 2018); (Vinh et al. 2010); (Bhatnagar et al. 2018) and with these five measures, study of performance for the visual methods are evaluated and presented in Table 4.

For the evaluation of performance, five performance measures are used, cluster accuracy (CA) (Pattanodom et al. 2016), normalized mutual information (NMI) (Amelio and Pizzuti 2015), precision (P), Recall (R), and F-Measure (F) (Xu et al. 2019; Li et al. 2016).

More visual clarity of clusters are appeared with the black coloured blocks of visual images using cosine similarity than Euclidean. The similarity features are computed with

Fig. 6 Cluster Tendency Assessment for Visual Methods (Standard Health Tweets—Fifteen Topics)

reference of a single viewpoint (i.e. origin) in cosine based visual method.

Recent visual method, MVS-VAT uses $(N-2)$ viewpoints for measure the cosine similarity between any two documents among the ‘ N ’ documents. Hence, it showed the more clarity of visual clusters than cosine metric. However, it taken high computation time for the similarity computation of N documents with respect to $(N-2)$ documents. Our proposed method (SICV-MVS-VAT) also uses the multiple

viewpoints for the good assessment of visual clusters. This method use the sample viewpoints of pre-generated inter clusters instead of $(N-2)$ viewpoints. Thus, it takes the fewer amounts of computational time and memory allocations than MVS-VAT.

Figures 7 and 8 visualizes the performance comparison with the measures, CA and NMI respectively. It states that proposed visual method SICV-MVS-VAT-NMF shown an

Table 2 Sample Tweets (2-topics to 4-Topics)

2-Topics Tweets	If you're too big to use XL mobility aids, then maybe your doctor telling you to lose weight has nothing to do with "FatPHo-biA" but because you're too fucking fat and it's a detriment to your life. People who consume a diet rich in flavanol-heavy food such as berries, tea and apples—tend to have lower blood pressure, according to a new study
3-Topics Tweets	If you're too big to use XL mobility aids, then maybe your doctor telling you to lose weight has nothing to do with "FatPHo-biA" but because you're too fucking fat and it's a detriment to your life. People who consume a diet rich in flavanol-heavy food such as berries, tea and apples—tend to have lower blood pressure, according to a new study. Next injection after this one will be in May. Then he will get a bone density test in August to see if there is an improvement
4-Topics Tweets	If you're too big to use XL mobility aids, then maybe your doctor telling you to lose weight has nothing to do with "FatPHo-biA" but because you're too fucking fat and it's a detriment to your life. People who consume a diet rich in flavanol-heavy food such as berries, tea and apples—tend to have lower blood pressure, according to a new study. Next injection after this one will be in May. Then he will get a bone density test in August to see if there is an improvement. This is about the same length of antibody resistance we get from a common cold. So the notion of herd immunity is blown out of the water

Table 3 Performance Evaluation of Visual Methods Using CA and NMI

Number of topics of TWEET Data	VN	cVN	MVN	SN	VL	cVL	MVL	SL	VLS	cVLS	MVS	SLS	VPL	cVPL	MPL	SPL
Clustering Accuracy (CA)																
2	1	1	0.975	1.000	0.525	0.525	0.5	0.526	0.800	0.750	0.500	0.801	0.575	0.575	0.55	0.585
3	0.767	0.767	0.683	0.769	0.408	0.408	0.392	0.425	0.675	0.450	0.733	0.742	0.425	0.425	0.492	0.493
4	0.988	0.988	0.975	0.988	0.325	0.350	0.325	0.351	0.594	0.481	0.681	0.594	0.325	0.325	0.294	0.331
5	0.965	0.965	0.955	0.967	0.310	0.300	0.295	0.310	0.565	0.445	0.495	0.568	0.320	0.305	0.280	0.328
6	0.937	0.937	0.941	0.945	0.262	0.262	0.270	0.279	0.533	0.479	0.495	0.534	0.262	0.270	0.254	0.272
7	0.921	0.921	0.954	0.954	0.250	0.254	0.254	0.264	0.454	0.400	0.454	0.458	0.261	0.275	0.282	0.291
8	0.675	0.675	0.653	0.680	0.213	0.200	0.219	0.225	0.434	0.388	0.541	0.349	0.228	0.222	0.241	0.245
9	0.647	0.647	0.683	0.684	0.214	0.228	0.206	0.229	0.408	0.308	0.419	0.410	0.206	0.208	0.183	0.210
10	0.598	0.598	0.568	0.600	0.215	0.190	0.210	0.217	0.463	0.303	0.443	0.471	0.213	0.208	0.198	0.219
11	0.502	0.502	0.497	0.511	0.198	0.182	0.191	0.198	0.325	0.255	0.386	0.414	0.186	0.184	0.186	0.190
12	0.540	0.540	0.463	0.542	0.200	0.175	0.190	0.205	0.398	0.258	0.425	0.429	0.188	0.181	0.185	0.194
13	0.479	0.479	0.438	0.481	0.196	0.179	0.185	0.198	0.387	0.225	0.387	0.391	0.185	0.188	0.179	0.194
14	0.461	0.461	0.493	0.495	0.193	0.171	0.184	0.195	0.345	0.202	0.329	0.348	0.161	0.171	0.170	0.178
15	0.418	0.418	0.435	0.437	0.172	0.158	0.145	0.175	0.345	0.187	0.322	0.348	0.157	0.163	0.168	0.172
Normalized mutual information (NMI)																
2	1	1	0.831	1.000	0.001	0.001	0	0.005	0.278	0.188	0	0.282	0.016	0.016	0.007	0.019
3	0.607	0.607	0.580	0.610	0.029	0.029	0.011	0.030	0.267	0.047	0.351	0.359	0.030	0.030	0.102	0.109
4	0.958	0.958	0.916	0.958	0.043	0.035	0.027	0.048	0.368	0.221	0.386	0.395	0.027	0.027	0.015	0.030
5	0.903	0.903	0.874	0.906	0.078	0.046	0.064	0.080	0.360	0.197	0.291	0.385	0.068	0.069	0.024	0.072
6	0.859	0.859	0.861	0.882	0.058	0.052	0.054	0.067	0.371	0.303	0.281	0.379	0.049	0.064	0.043	0.068
7	0.851	0.851	0.906	0.908	0.071	0.060	0.061	0.076	0.290	0.228	0.322	0.328	0.052	0.073	0.069	0.078
8	0.626	0.626	0.632	0.642	0.056	0.047	0.069	0.072	0.339	0.243	0.399	0.420	0.071	0.066	0.078	0.081
9	0.585	0.585	0.590	0.598	0.077	0.075	0.059	0.080	0.332	0.193	0.308	0.341	0.073	0.073	0.055	0.079
10	0.564	0.564	0.565	0.570	0.125	0.077	0.095	0.135	0.384	0.200	0.375	0.389	0.092	0.079	0.067	0.098
11	0.454	0.454	0.438	0.471	0.089	0.074	0.080	0.095	0.288	0.180	0.315	0.334	0.089	0.085	0.077	0.091
12	0.514	0.514	0.449	0.520	0.103	0.084	0.084	0.110	0.336	0.202	0.364	0.370	0.094	0.094	0.096	0.099
13	0.447	0.447	0.453	0.455	0.118	0.109	0.102	0.125	0.335	0.189	0.297	0.340	0.102	0.106	0.106	0.109
14	0.434	0.434	0.464	0.474	0.126	0.097	0.114	0.10	0.304	0.183	0.302	0.308	0.097	0.104	0.089	0.110
15	0.407	0.407	0.439	0.442	0.115	0.099	0.086	0.105	0.312	0.160	0.282	0.315	0.108	0.110	0.104	0.115

VN:VAT-NMF, cVN: cVAT-NMF, MVN:MVS-VAT-NMF, SN:SICV-MVS-VAT-NMF, VL:VAT-LDA, cVL: cVAT-LDA, MVL:MVS-VAT-LDA, SL:SICV-MVS-VAT-LDA, VLS:VAT-LSI,cVLS:cVAT-LSI, MVS:MVS-VAT-LSI, SLS:SICV-MVS-VAT-LSI, VPL:VAT-PLSI, cVPL:cVAT-PLSI, MPL:MVS-VAT-PLSI, SPL:SICV-MVS-VAT-PLSI

Bold indicates that high performance value

Table 4 Performance Evaluation of Visual Methods Using Precision, Recall and F-Measure

Number of Topics of Tweet Datasets	VN	cVN	MVN	SN	VL	cVL	MVL	SL	VLS	cVLS	MVS	SLS	VPL	cVPL	MPL	SPL
Precision (P)																
2	1	1	0.981	1.000	0.525	0.528	0.506	0.535	0.808	0.747	0.522	0.811	0.572	0.572	0.556	0.578
3	0.759	0.761	0.684	0.781	0.411	0.409	0.397	0.421	0.671	0.454	0.736	0.739	0.426	0.428	0.486	0.492
4	0.985	0.985	0.975	0.997	0.324	0.345	0.321	0.349	0.591	0.481	0.682	0.691	0.327	0.329	0.297	0.337
5	0.961	0.966	0.953	0.967	0.305	0.307	0.295	0.317	0.563	0.446	0.493	0.572	0.325	0.312	0.288	0.331
6	0.937	0.939	0.941	0.945	0.261	0.261	0.268	0.271	0.537	0.479	0.494	0.545	0.262	0.273	0.257	0.276
7	0.922	0.924	0.954	0.959	0.258	0.255	0.255	0.262	0.451	0.457	0.456	0.461	0.268	0.264	0.281	0.284
8	0.671	0.677	0.654	0.682	0.217	0.213	0.215	0.224	0.432	0.384	0.547	0.549	0.234	0.235	0.246	0.251
9	0.645	0.646	0.682	0.689	0.214	0.224	0.219	0.231	0.417	0.317	0.412	0.423	0.203	0.207	0.188	0.214
10	0.605	0.594	0.572	0.615	0.213	0.207	0.215	0.218	0.461	0.312	0.441	0.465	0.217	0.214	0.197	0.222
11	0.507	0.507	0.498	0.511	0.198	0.186	0.197	0.203	0.327	0.254	0.387	0.405	0.198	0.184	0.187	0.205
12	0.541	0.544	0.471	0.555	0.205	0.173	0.198	0.211	0.385	0.265	0.422	0.427	0.194	0.189	0.189	0.204
13	0.485	0.482	0.442	0.489	0.198	0.162	0.184	0.204	0.384	0.227	0.384	0.398	0.184	0.187	0.175	0.194
14	0.465	0.464	0.494	0.502	0.194	0.178	0.183	0.199	0.322	0.208	0.325	0.338	0.166	0.177	0.174	0.182
15	0.422	0.419	0.432	0.439	0.177	0.168	0.144	0.181	0.342	0.191	0.327	0.351	0.153	0.169	0.167	0.178
Recall (R)																
2	1.000	1	0.978	1.000	0.526	0.527	0.501	0.531	0.811	0.745	0.521	0.814	0.575	0.571	0.552	0.582
3	0.769	0.767	0.685	0.775	0.41	0.407	0.39	0.415	0.676	0.451	0.734	0.736	0.421	0.425	0.482	0.485
4	0.989	0.988	0.977	0.995	0.327	0.342	0.326	0.348	0.595	0.48	0.68	0.685	0.323	0.325	0.292	0.335
5	0.964	0.965	0.957	0.968	0.311	0.305	0.29	0.315	0.567	0.442	0.497	0.57	0.321	0.308	0.282	0.327
6	0.939	0.937	0.943	0.948	0.264	0.266	0.271	0.279	0.534	0.475	0.495	0.54	0.265	0.272	0.251	0.272
7	0.924	0.921	0.958	0.961	0.252	0.257	0.257	0.264	0.455	0.454	0.455	0.458	0.263	0.269	0.282	0.285
8	0.674	0.675	0.651	0.687	0.215	0.217	0.215	0.221	0.435	0.389	0.542	0.545	0.231	0.235	0.245	0.245
9	0.648	0.647	0.68	0.682	0.216	0.229	0.204	0.234	0.41	0.315	0.418	0.421	0.208	0.208	0.184	0.215
10	0.601	0.598	0.571	0.611	0.218	0.201	0.215	0.218	0.465	0.306	0.44	0.469	0.215	0.211	0.195	0.219
11	0.505	0.507	0.497	0.509	0.197	0.184	0.193	0.202	0.322	0.257	0.384	0.401	0.191	0.186	0.183	0.198
12	0.542	0.548	0.468	0.551	0.201	0.177	0.199	0.206	0.389	0.261	0.42	0.425	0.192	0.188	0.181	0.199
13	0.481	0.479	0.44	0.484	0.195	0.169	0.187	0.191	0.388	0.228	0.388	0.391	0.187	0.189	0.178	0.195
14	0.463	0.461	0.497	0.501	0.19	0.175	0.189	0.195	0.329	0.205	0.325	0.335	0.164	0.172	0.175	0.181
15	0.42	0.418	0.438	0.442	0.174	0.161	0.147	0.166	0.34	0.187	0.321	0.347	0.159	0.165	0.163	0.172
F-Measure (F)																
2	1	1	0.975	1	0.525	0.528	0.5	0.535	0.8	0.75	0.5	0.812	0.575	0.576	0.55	0.585
3	0.767	0.768	0.683	0.78	0.408	0.409	0.392	0.425	0.675	0.45	0.733	0.742	0.425	0.427	0.492	0.498
4	0.988	0.989	0.975	0.989	0.325	0.351	0.325	0.365	0.594	0.481	0.681	0.602	0.325	0.327	0.294	0.331
5	0.965	0.968	0.955	0.972	0.31	0.311	0.295	0.315	0.565	0.445	0.495	0.571	0.32	0.305	0.28	0.325
6	0.937	0.939	0.941	0.952	0.262	0.264	0.272	0.279	0.533	0.479	0.495	0.541	0.262	0.27	0.254	0.268
7	0.921	0.925	0.954	0.957	0.25	0.254	0.254	0.264	0.454	0.4	0.454	0.458	0.261	0.275	0.282	0.289
8	0.675	0.677	0.653	0.681	0.213	0.214	0.219	0.225	0.434	0.388	0.541	0.451	0.228	0.222	0.241	0.245
9	0.647	0.649	0.683	0.685	0.214	0.228	0.206	0.235	0.408	0.308	0.419	0.425	0.206	0.208	0.183	0.211
10	0.598	0.601	0.568	0.61	0.215	0.199	0.211	0.22	0.463	0.303	0.443	0.471	0.213	0.208	0.198	0.215
11	0.502	0.503	0.498	0.512	0.198	0.182	0.191	0.201	0.325	0.255	0.386	0.414	0.186	0.184	0.187	0.189
12	0.54	0.542	0.463	0.548	0.2	0.175	0.198	0.205	0.398	0.258	0.425	0.431	0.188	0.181	0.185	0.194
13	0.479	0.481	0.438	0.485	0.196	0.179	0.185	0.199	0.387	0.225	0.387	0.391	0.185	0.188	0.179	0.194
14	0.461	0.463	0.493	0.498	0.193	0.171	0.184	0.204	0.345	0.202	0.329	0.348	0.161	0.171	0.17	0.178
15	0.418	0.421	0.435	0.44	0.172	0.158	0.145	0.179	0.345	0.187	0.322	0.348	0.157	0.163	0.168	0.174

MVS-VAT and SICV-MVS-VAT are good for accessing the cluster tendency as well as to discover the complete tweets data clustering results. However, existing MVS-VAT demands the more computational time. Our proposed method SICV-MVS-VAT overcomes the problem of time and space requirements. Fair assessment of clusters based on the criteria of detection of appeared square-shaped dark coloured blocks along the diagonal of visual images. Each square-shaped black coloured block defines the separate clusters. Tables 5 and 6 shown the computation time (in seconds) and space allocation for the visual methods

Bold indicates that high performance value

Table 5 Computation Time (in Seconds) for SICV-MVS-VAT and MVS-VAT

Tweets data set	Computational time (in seconds)							
	MVS-VAT-NMF	MVS-VAT-LDA	MVS-VAT-VLSI	MVS-VAT-PLSI	SICV-MVS-VAT-NMF	SICV-MVS-VAT-LDA	SICV-MVS-VAT-VLSI	SICV-MVS-VAT-PLSI
2-Topics	26.34208	33.36562	16.19871	37.18909	4.042613	3.996161	2.079599	6.653691
3-Topics	60.58405	65.92725	38.15781	55.79712	10.12713	10.90094	13.03906	17.07504
4-Topics	86.76512	94.18409	102.684	88.7069	15.5009	23.47069	18.46501	28.36517
5-Topics	88.18191	135.4584	154.9304	121.3091	23.52997	26.22308	24.58498	35.25269
6-Topics	109.2492	128.2335	70.42687	129.1825	34.95217	38.08368	36.30715	42.00689
7-Topics	139.2638	156.7343	181.1331	141.2082	74.63157	86.87654	92.91105	95.558
8-Topics	229.8169	201.767	320.3579	192.3625	119.2987	129.2212	130.2145	103.7142
9-Topics	220.8313	208.1528	243.4769	215.9448	108.7435	103.2379	109.5956	120.1936
10-Topics	382.598	433.5549	381.0752	400.8983	226.0615	252.6879	239.7635	188.4831
11-Topics	582.7059	652.0813	531.357	506.8658	185.385	198.0097	196.9235	199.8446
12-Topics	501.5469	660.0621	571.2034	606.4286	231.3069	234.6661	224.2	413.3133
13-Topics	439.7768	530.0077	375.384	467.4771	297.954	315.1951	282.4218	332.3261
14-Topics	836.2546	1152.192	774.8114	938.6797	401.4186	422.6668	378.3016	425.2632
15-Topics	964.2074	863.0787	1208.834	4033.32	464.0149	482.0826	481.4223	529.5467

Table 6 Memory Allocation (in kbs) for SICV-MVS-VAT and MVS-VAT

Tweets data set	Memory allocation (in kbs)							
	MVS-VAT-NMF	MVS-VAT-LDA	MVS-VAT-VLSI	MVS-VAT-PLSI	SICV-MVS-VAT-NMF	SICV-MVS-VAT-LDA	SICV-MVS-VAT-VLSI	SICV-MVS-VAT-PLSI
2-Topics	19.18359	13.00391	15.09375	176.1016	17.98438	12.49609	12.78906	123.4648
3-Topics	19.89453	15.28125	14.16797	120.5469	18.35547	13.24219	14.26563	119.7461
4-Topics	20.66406	15.85156	14.73828	120.7109	19.38281	14.88672	14.41797	120.6602
5-Topics	19.90234	17.28516	18.32422	124.0664	19.80469	15.21875	15.43359	124.7148
6-Topics	20.33984	14.89844	15.51953	128.9922	18.81641	14.58594	15.66797	128.9609
7-Topics	21.18359	15.98438	16.24219	124.0859	20.875	15.62109	15.08594	115.8438
8-Topics	22.96484	17.10156	16.33984	130.6211	19.62891	15.6875	16.29297	129.0977
9-Topics	23.35547	19.73828	17.4375	126.832	20.55469	16.23828	18.71094	126.8242
10-Topics	22.63672	21.03125	18.87109	129.7695	19.69922	17.29688	20.33984	127.6406
11-Topics	25.44922	21.22656	57.2656	159.3125	22.59766	14.17969	20.15234	128.2578
12-Topics	25.24609	13.10547	11.84375	121.4883	23.53906	12.58984	22.89453	119.1406
13-Topics	28.25781	23.58203	22.50391	129.6563	22.75391	13.75391	10.74219	129.9414
14-Topics	23.66406	66.1719	19.97266	238.4453	22.3125	24.70703	23.28906	131.5
15-Topics	25.66797	55.0703	20.97266	218.4297	23.23438	26	11.78516	110.4648

excellent performance for the different sets of topic models i.e. 2-topics to 15-topics.

Topic model NMF greatly supports to proposed sampling based visual method compared to other topic models. The same performance values are illustrated experimentally and presented in the performance graph comparison charts.

Based on the observation of time and space values of SICV-MVS-VAT and MVS-VAT, we stated that proposed

SICV-MVS-VAT taken the less computational time and less memory allocation than MVS-VAT due to fact for the selected sample viewpoints instead considering (N-2) viewpoints, unlike MVS-VAT. Hence, proposed SICV-MVS-VAT has outperformed the other visual methods with respect to the quality of tweets data clustering results, time and space requirements.

Fig. 7 Performance Comparison of Visual Methods for the Topics Datasets (Using Clustering Accuracy)

Fig. 8 Performance Comparison of Visual Methods for the Topics Datasets (Using NMI)

5 Conclusion and scope

Visual methods are well suited for accessing the pre-clusters information (cluster tendency ‘k’) and generation of tweets data clustering results. It proved that cosine-based pre-clusters estimations are reasonable when compared to clusters assessment with Euclidean in text clustering. However, a single viewpoint is considered when considering similarity features among tweets in the tweet data clustering. Another visual method MVS-VAT access the cluster tendency with multi viewpoints rather than a single viewpoint. Thus, it produces more effective clustering results than cosine based visual methods. Due to overcome the time and space issues of MVS-VAT, proposed work uses the sampling of inter-cluster viewpoints. The in-depth experimental analysis proved that

proposed SICV-MVS-VAT is an efficient technique for the tweet data clustering.

References

- Amelio A, Pizzuti C (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, pp. 1584–1585.
- 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 (Cat. No. 02CH37290), pp 2225–2230.
- Bhatnagar V, Majhi R, Jena PR (2018) Comparative performance evaluation of clustering algorithms for grouping manufacturing firms. Arab J Sci Eng 43(8):4071–4083
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learning Res pp 993–1022.
- Datta S, Bhattacharjee S, Das S (2018) Clustering with missing features: a penalized dissimilarity measure based approach. Mach Learn 107(12):1987–2025
- Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inform Sci 41(6):391–407
- Devisetty SDP, Yarramneni MS, Akash VY, Pellakuri V (2019) Sentiment analysis of tweets using rapid miner tool. Int J Innovative Technol Exploring Eng 8(6):1410–1414
- Havens TC, Bezdek JC (2011) An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm. IEEE Trans Knowl Data Eng 24(5):813–822
- Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp 50–57.
- Hu Y, Hathaway RJ (2008) An algorithm for clustering tendency assessment. WSEAS Trans Math 7(7):441–450
- Hu Y, John A, Wang F, Kambhampati S (2012) ET-LDA: joint topic modeling for aligning events and their twitter feedback. In: AAAI, pp 59–65.
- Ismail M, Prakash KB, Rao MN (2018) Collaborative filtering-based recommendation of online social voting. Int J Eng Technol (UAE) 7(3):1504–1507
- Jose T, Babu SS (2019) Detecting spammers on social network through clustering technique. J Ambient Intell Human Comput, pp 1–15.
- Kumar D, Bezdek JC (2020) Visual approaches for exploratory data analysis: a survey of the visual assessment of clustering tendency (VAT) family of algorithms. IEEE Syst Man Cybernet Mag 6(2):10–48
- Kumar D, Bezdek JC, Palaniswami M, Rajasegarar S, Leckie C, Havens TC (2015) A hybrid approach to clustering in big data. IEEE Trans Cybernet 46(10):2372–2385
- Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562.
- Li Z, Shang W, Yan M (2016) News text classification model based on-the topic model. In: IEEE/ACIS 15th international conference on computer and information science (ICIS), pp 1–5.
- Mahallati S, Bezdek JC, Popovic MR, Valiente TA (2019) Cluster tendency assessment in neuronal spike data. PLoS ONE 14(11):1–29
- Neogi PPG, Das AK, Goswami S, Mustafi J (2020) Topic modeling for text classification. In: Emerging technology in modelling and graphics, pp 395–407.

- Nguyen DT, Chen L, Chan CK (2010) Multi-viewpoint based similarity measure and optimality criteria for document clustering. In: Asia information retrieval symposium, pp 49–60.
- Oghbaie M, Zanjireh MM (2018) Pairwise document similarity measure based on present term set. *J Big Data* 5(1):1–23
- Park LA, Bezdek JC, Leckie C, Kotagiri R, Bailey J, Palaniswami M (2016) Visual assessment of clustering tendency for incomplete data. *IEEE Trans Knowl Data Eng* 28(12):3409–3422
- Pattanodom M, Iam-On N, Boongoen T (2016) Clustering data with the presence of missing values by ensemble approach. In: 2016 second asian conference on defence technology (acdt), pp 151–156.
- Pessiot JF, Kim YM, Amini MR, Gallinari P (2010) Improving document clustering in a learned concept space. *Inf Process Manage* 46(2):180–192
- Prasad KR, Mohammed M, Noorullah RM (2019) Hybrid topic cluster models for social healthcare data. *Int J Adv Comput Sci Appl* 10(11):490–506
- Prasad KR, Mohammed M, Noorullah RM (2019) Visual topic models for healthcare data clustering. *Evolutionary Intelligence*, pp 1–18.
- Pushpalatha A, Harivarthini S, Akil S, Akshaya V, Amarendra K, Bhanu Prakash K (2020) Real time objects recognition and classification to audio conversion for visually impaired person. *Int J Adv Sci Technol* 29(3):8290–8297
- Ramage D, Hall D, Nallapati R, Manning CD (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing, pp. 248–256.
- Reddy BE, Prasad KR (2016) Improving the performance of visualized clustering method. *Int J Syst Assurance Eng Manag* 7(1):102–111
- Rehioui H, Idrissi A (2019) New clustering algorithms for twitter sentiment analysis. *IEEE Syst J* 14(1):530–537
- Sechelea A, Do Huu T, Zimos E, Deligiannis N (2016) Twitter data clustering and visualization. In: 2016 23rd international conference on telecommunications (ICT), pp. 1–5.
- Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS ONE* 10(12):1–20
- Srinivas K (2018) Test case prioritization with special emphasis on automation testing using hybrid framework. *J Theoretical Appl Inform Technol* 96(13):4180–4190
- Sukhija N, Tatineni M, Brown N, Van Moer M, Rodriguez P, Callicott S (2016) Topic modeling and visualization for big data in social sciences. In: IEEE Conferences on ubiquitous intelligence & computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, internet of people, and smart world congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp 1198–1205.
- Suleman Basha M, Mouleeswaran SK, Rajendra Prasad K (2019) Cluster tendency methods for visualizing the data partitions. *Int J Innovative Technol Exploring Eng* 8 (11): 2978–2982.
- Suri P, Roy NR (2017) Comparison between LDA & NMF for event-detection from large text stream data. In: International conference on computational intelligence & communication technology (CICT), pp. 1–5.
- Tang H, Chu S, Hasegawa-Johnson M, Huang T (2011) Partially supervised speaker clustering. *IEEE Trans Pattern Anal Mach Intell* 34(5):959–971
- Therese SS, Lingam C (2017) A linear visual assessment tendency based clustering with power normalized cepstral coefficients for audio signal recognition system. *J Ambient Intell Human Comput* pp. 1–14.
- Varish N, Pal AK, Hassan R, Hasan MK, Khan A, Parveen N, Memon I (2020) Image retrieval scheme using quantized bins of color image components and adaptive tetrolet transform. *IEEE Access*, pp 117639–117665.
- Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J Mach Learn Res* pp 2837–2854.
- Wallach HM (2006) Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning, pp 977–984.
- Wang L, Geng X, Bezdek J, Leckie C, Kotagiri R (2008) SpecVAT: Enhanced visual cluster analysis. In: 2008 eighth IEEE international conference on data mining, pp 638–647.
- Xu R, Wunsch D (2005) Survey of clustering algorithms. *IEEE Trans Neural Netw* 16(3):645–678
- Xu G, Meng Y, Chen Z, Qiu X, Wang C, Yao H (2019) Research on topic detection and tracking for online news texts. *IEEE Access*, pp 58407–58418.
- Yi F, Jiang B, Wu J (2020) Topic modeling for short texts via word embedding and document correlation. *IEEE Access*, pp 30692–30705.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.