Hybrid Topic Cluster Models for Social Healthcare Data

K Rajendra Prasad[1]
Dept. of CSE
Institute of Aeronautical Engineering
Hyderabad, India

Moulana Mohammed[2]
Dept. of CSE
Koneru Lakshmaiah University
Guntur, India

R M Noorullah[3]
Dept. of CSE
Koneru Lakshmaiah University
Guntur, India

Abstract—**Social media and in particular, microblogs are becoming an important data source for disease surveillance, behavioral medicine, and public healthcare. Topic Models are widely used in microblog analytics for analyzing and integrating the textual data within a corpus. This paper uses health tweets as microblogs and attempts the health data clustering by topic models. The traditional topic models, such as Latent Semantic Indexing (LSI), Probabilistic Latent Schematic Indexing (PLSI), Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and integer Joint NMF(intJNMF) methods are used for health data clustering; however, they are intractable to assess the number of health topic clusters. Proper visualizations are essential to extract the information from and identifying trends of data, as they may include thousands of documents and millions of words. For visualization of topic clouds and health tendency in the document collection, we present hybrid topic models by integrating traditional topic models with VAT. Proposed hybrid topic models viz., Visual Non-negative Matrix Factorization (VNMF), Visual Latent Dirichlet Allocation (VLDA), Visual Probabilistic Latent Schematic Indexing (VPLSI) and Visual Latent Schematic Indexing (VLSI) are promising methods for accessing the health tendency and visualization of topic clusters from benchmarked and Twitter datasets. Evaluation and comparison of hybrid topic models are presented in the experimental section for demonstrating the efficiency with different distance measures, include, Euclidean distance, cosine distance, and multi-viewpoint cosine similarity.**

Keywords—Multi-viewpoint based metric; traditional topic models; hybrid topic models; topic visualization; health tendency

I. Introduction

Twitter, Facebook, and microblogs [21], [22], [23] reveals the opinions of public and assessment of this social data [13], [14] is an emerging need in the applications like topics detection [1], [4], product promotion in business [4], political predictions [6], and health recommendations [2], [6]. Rapid urbanization is posing the number of public health-related problems, including accidents and injuries, healthcare disparities, and increasing disease burdens due to changes in lifestyle and nutrition, as well as increased environmental pollution. Twitter data contain sufficient health-related information, which dynamically updates millions of user posts for various health topics with all polarities of information [18]. The rate of flow of information increased because of tweets and retweets. In this paper, relationships between tweets as the interactions based on users (mentions), actions (reply and retweet), and contentment similarity are defined. A massive amount of data is generated, which makes it more difficult to extract topics [3] from these text documents manually. Topic modeling is an approach, which extracts implicit topics from document sets. There are several independent topics with different topic probabilities, and topics are consisting of words with different word probability. People frequently search for health-related topics in social media for envisaging solutions towards healthcare. Document clustering has become an increasingly important technique for unsupervised document organization [5], automatic topic extraction, and fast information retrieval [35], [38], or filtering. Clustering techniques [8], [17], used based on topic classification, namely, LSI [27], PLSI [29], LDA [26], supervised LDA [66], multi-class sLDA [67], medLDA [68], NMF [19], [36] and intJNMF [37], [40], sNMF [7] are specifically used for text mining. Thes methods are most suitable for post clustering and cannot able to determine the pre-cluster tendency in such cases; there is possible to get poor health clustering results. Bezdek et al. proposed [9], [20], [15] visual access tendency (VAT) for assessment of health data, which imposes good health topics (or clusters) visually. VAT use the Prim's logic for re-ordering the dissimilarity features for the set of data objects and saves this data in the matrix of reordered dissimilarity matrix (RDM) [15] and shows the visual clusters while displaying the image of RDM. The final image of RDM is known as VAT Image that gives the clarity of visualized square-shaped dark blocks along the diagonal for representing the clusters. We aim to investigate which public health issues are discussed in social media and in particular Twitter, and we use both VAT and traditional topic models in the proposed hybrid framework to overcome the problem of health cluster tendency, these hybrid topic models are VNMF, VLDA, VLSI, and VPLSI. Euclidean, cosine based, and multi-viewpoint cosine based metrics are used for finding similarity features between tweets documents in proposed models. Text documents similarity features are very sparse and high dimensional. Cosine based metric computes the distance between objects concerning magnitude and direction of document vectors. Hence, it strongly supports topics clustering than Euclidean. In traditional similarity/dissimilarity uses only a single viewpoint, which is the origin. In a multi-viewpoint cosine similarity based metric, we used many different viewpoints; objects assumed not to be in the same cluster. Using this more accurate assessment of how close or distant a pair of points if we look at them from many different viewpoints and average of similarities measured relatively from the views of all other documents outside that cluster. The overall similarity is determined by taking an average over all the viewpoints not belonging to the cluster. Multi-viewpoints cosine similarity offers a more informative assessment of similarity than the single-origin point-based similarity measure. Our work is useful in the development of smart healthcare applications for recognition of health problems [31], [32] based on symptoms, and produces social recommend solutions for each specific health problem. The key contribution of our work is summarized as follows:

1) Pre-processing of tweets is initiated for removing of unwanted symbols, URLs, and stemming on words for maintaining uniformity in text analysis.

2) Estimation of health topics tendency is done.

3) Visual clusters are developed for these health topics.

4) Hybrid topic cluster models are proposed for topics tendency and health cluster analytics.

5) Proposed techniques give a solution for health classification and recommended solutions for health problems.

The remaining part of the paper is organized as follows: Section 2 presents the related work of topic models; Section 3 describes the proposed hybrid topic models; Section 4 discusses the experimental study; Visual cluster analysis and discussion are presented in Section 5. Finally, Section 6 presents the conclusion and future scope of the work.

II. Related Work of Topic Models

With minimal human intervention, topic models can discover prominent topics in the text without specific knowledge of the dataset and used to derive clusters. Topic models have a wide range of applications especially in the fields of text mining and information retrieval. Probabilistic models describe the topics distributions of terms and subjects, whereas non-probabilistic models used for describing the topics from the importance of terms.

Latent semantic indexing (LSI) [27] useful for defining of term-document matrix with latent semantic factors that use the singular value decomposition (SVD) [28], [29]. Post cluster tendency methods [16], [17] are usually considered as lazy topic learners. To overcome this problem, pre-cluster tendency methods preferred in cluster tendency assessment. Normalized cut (Ncut) described as Ncut-NMF [24], Xiaohui Yan *et al.* employed alternating non-negative least squares (ANLS) [25] algorithm for Ncut-NMF. Latent Dirichlet Allocation (LDA) [26], a Bayesian three-level hierarchical statistical model of PLSA implements posterior distribution using Bayesian inferences. Topics derived from interactions between tweets and re-tweets in the intJNMF method [49]. For visual assessment of clusters VAT and cosine VAT (cVAT) [18] extracts the dissimilarity features of documents concerning topic features and stored in dissimilarity matrix DM. The crisp partition matrix [17] is a broad approach of visual techniques, and it discovers document clustering results to topics, proposed as visual topic models in this paper. In this paper, we used a broad approach of VNMF, VLDA, VLSI, and VPLSI for the automation of tweets clustering of health tweets with prior knowledge of cluster health tendency and is presented in the proposed framework.

A. Similarity and Clustering Documents

Similarity metrics play an essential role in the success or failure of a clustering method. The effectiveness of the clustering algorithm depends on the appropriateness of the similarity measure to the data available. Documents comparison concerning topics features is performed with either distance or similarity metrics. Smaller the distance between two objects, the more similar they are to each of the feature vectors. Euclidean distance is the most common metric for computing similarity features between documents, used in most of the traditional topic modeling algorithms computed as in "(1)".

$$\text{D(Twn, Twm)}=\sqrt{(x1-y1)^2 \pm - - \pm (xz-yz)^2} \qquad (1)$$

Cosine similarity is the best suitable when the data is in high-dimensional and sparse. It is a big similarity score in text mining and information retrieval. Cosine measure is used in spherical k-means algorithm [45], min-max cut graph-based spectral method [46], average weight [47], normalized cut [48] and document clustering using pairwise similarity score [49], [50]. In [51], Strehl et al. compared four measures Euclidean, cosine, Pearson correlation, and extended Jaccard, and concluded cosine and Jaccard are the best ones on web documents. Lenco et al. [52] introduced a similar context-based distance learning method for categorical data. Lakkaraju et al. [53] employed a conceptual tree-similarity measure to identify similar documents. Chim and Deng [54] proposed a phrase-based document similarity. The cosine metric uses both the magnitude and direction of the vectors. Similarity features between two tweet documents computed using "(2)".

$$\text{Cosine (Twn, Twm)} =\frac{Twn.Twm}{\|Twn\|\|Twm\|} \qquad (2)$$

Multi-viewpoint based cosine similarity: We have a more accurate assessment of how close or distant a pair of points is if we look at them from multi-viewpoints computed and this similarity measure is called as multi-viewpoint based cosine similarity, and it is shown in "(3)".

$$\text{Sim}\left(d_i, d_j\right)_{d_i,d_j \,\epsilon S_r} = \frac{1}{n-n_r}\, A \qquad (3)$$

Where A = $\sum_{d_h \epsilon S \backslash S_r} Sim(d_i - d_h, A = \pi r^2 - d_h)$

The compatible visual hybrid topic models with Euclidean, cosine based, and multi-viewpoints cosine based metrics for assessment of twitter health topics is described as follows:

Step1: The term-document matrix is transformed into a topic-document matrix with a size of n x m, in which n and m describe the number of topics and documents, respectively.

Step2: Dissimilarity matrix (DM) is computed using distance metrics, and it is shown in the following matrix.

$$\text{DM} = \begin{pmatrix} \text{T11} & \dots & \text{T1m} \\ \dots & \dots & \dots \\ \text{Tm1} & \dots & \text{Tmm} \end{pmatrix}$$

Step3: Reorder the dissimilarity features of tweet documents using the procedure of VAT [9], which resulting matrix is RDM.

Step4: The RDM matrix is normalized within the scale of 0 to 1, then display the RDM grayscale image for visualizing topic clusters, in which the topic clusters are shown as square-shaped dark colored blocks.

The visual images were displayed with Euclidean, cosine based distance, and multi-viewpoint based distance RDMs, more clarity of visual topics are shown in multi-viewpoints cosine based RDM. The proposed framework uses the topic derivation techniques along with visual models for deriving clusters estimations health tweets and explore the health topics based classification results. Respective proposed models are illustrated in the following section.

III. Proposed Hybrid Topic Models

For clusters estimations, visual models show an impressive result for unlabeled datasets. Traditional models include PLSI, NMF, LDA, and intJNMF are popular in the determination of topics clustering results. It is required to determine the prior classification of social health data clusters information; which informative assessment produces the quality of topics clusters. In such cases, users guessing is not required for topics clustering models; thus, there is a broad scope to improving the efficiency of topics cluster models. A combination of visual model (i.e., VAT) and topic models are proposed, by nature, these are known as hybrid models, which are Visual LSI (VLSI), Visual PLSI (VPLSI), Visual NMF (VNMF), and Visual LDA (VLDA). Proposed hybrid models compute the low-rank matrices for active learning of topics of tweets documents. The low-rank matrices are less sparse that can be used for representing the related topics and tweets documents rather than the relationship between terms and tweet documents. The proposed VNMF, VLDA, VPLSI, and VLSI exploit the relationships between documents and ignores retweet interactions. The VLSI uses the singular value decomposition (SVD) for finding the document scores for topics. The dissimilarity features between documents are computed based on document scores.

Algorithm 1: VLSI

```
Input: X, term-document matrix in w-d form
Output: Number of Topics (or Clusters)
Method:
  1. Compute the document scores ‘V’ for given X using
     SVD [ ] approach.
  2. Compute the Dissimilarity Matrix (DM) of V
  3. Find the Reordered Dissimilarity Matrix (RDM) of V
     using [15] and display image(RDM).
  4. Assess a good number of topics from the image of RDM
```

The term-document matrix 'X' is constructed for the set of tweet documents, and it is taken as input in algorithm 2 (VPLSI). Step 1 computes the prior probability P(d, w) with the probabilities of P(d) and P(w|d), whereas P(d) and P(w|d) describes the initial probability, and conditional independence probability respectively. The values of P(w|z), P(z), P(d|z) are re-calculated by EM procedure [30], which support for extraction of certain topics distributions from the set of tweet documents. Probabilities of topic-document are stored in the variable of V. Re-estimations of probabilities are described in Step 3. The topic-document probabilities are stored into matrix V. Step 4 computes the dissimilarity matrix 'DM' using distance metrics (Either Euclidean or Cosine). Step 4 uses the Prim's logic for dissimilarities of data objects and finds the result in another matrix, known as re-ordered dissimilarity matrix (RDM). Step 6 finds the image of RDM, which shows the clusters in the shape of square dark colored blocks. Estimations of clusters are performed by the assessment of square-shaped dark colored blocks from the image of RDM. All these steps are described in the proposed VPLSI, which shown as follows.

Algorithm 2: VPLSI

```
Input: X, term-document matrix in w-d form
Output: Number of Topics (or Clusters)
Method:
  1. Compute the prior probability $\mathrm{P(d, w) = P(d)P(w|d)}$
     with
  2. Conditional probability independence $\mathrm{P(w|d) =}$
     $\sum_{z \epsilon Z} \mathrm{P(w|z)P(z|d)}$
     Re-estimate the parameters using EM algorithm, and
     Update the value of P(d,w). where$\mathrm{P(d, w) =}$
     $\sum_{z \epsilon Z} \mathrm{P(w|z)P(z)P(d|z)}$
  3. Store the values of P(d|z) into V
  4. Compute Dissimilarity Matrix (DM) of V
  5. Reordered Dissimilarity Matrix (RDM) of DM.
  6. Assess a good number of topics from the image of RDM
```

The proposed VPLSI uses the term-documents [10] and related parameters of EM for determining the topic-document matrix. The terms-correlation features are not considered in the estimation of topic parameters. These features are essential in the determination of topics of tweet documents. For this reason, another visual topic model VNMF is proposed; it initially finds the topic-document matrix with consideration of the term-correlation matrix 'S.' It derives the useful conceptual (or hidden) topics before deriving the term-topic matrix ‘U’. The steps of VNMF are shown in Algorithm 3.

```
Algorithm 3: VNMF

Input: X, Term-Document Matrix
S, Term-Correlation Matrix
Output: Number of Topics (or Clusters)
Method:
  1. Term-topic matrix U minimizes the function L(U), which satisfies L(U)= $\|S - UU^T\| >= 0$
  2. Apply convergence for finding $\mathrm{U} = \mathrm{SU}(\mathrm{U^TU})^{-1}$
  3. Apply convergence for finding topic-document matrix V using $V = (\mathrm{U^TU})^{-1}\mathrm{U^TX}$
  4. Compute Dissimilarity Matrix (DM) of V
  5. Compute Reordered Dissimilarity Matrix (RDM) of V using [15]
  6. Assess the suitable number of topics from the image of RDM.
```

Term-document X shows the association between tweet documents and their respective terms. Correlations between the terms are also taken as input (S) of VNMF. Step 1 shows the term-topic matrix, which satisfies the objective function L(U). Step 2 and 3 show the optimal term-topic matrix 'U' derivations with convergence and derives the optimal topic-document matrix V. In V; the documents are denoted as document vectors. Step 4 computes the distances between document vectors using distance metrics and stored in dissimilarity matrix (DM). In Step5, the values of DM are reordered according to Prim's logic, and final re-ordered values are stored in RDM and procedure given in [11]. In step 6, the RDM image is displayed that shows the visual clusters through dark-colored square-shaped blocks. Each square-shaped dark colored block shows an individual crisp partition for a similar set of objects. With the crisp partitions of square-shaped blocks, the cluster labels of objects are predicted which gives the complete topics clustering results.

```
Algorithm 4: VLDA

Input: X, Term-Document Matrix
Output: V, Topic- Document Matrix, Number of Topics (or Clusters)
Method :
  1. Generate multiple topics P(W) with the likelihood
  2. Estimate word distribution and topic convergence distribution using Dirichlet distribution by enforcing the sparse conditions.
  3. Find an optimal solution using the EM algorithm
      3.1 E-Step: Calculate Posterior probability using Bayesian inferences
      3.2 M-Step: Re-estimate probability of document covering topics and update the values into 'V'
4. Compute Dissimilarity Matrix (DM) of V.
5. Reordered Dissimilarity Matrix (RDM) of V using [15].
6. Assess the good number of topics from the image of RDM.
```

Another topic model, say, LDA finds the Dirichlet coefficients [39] for various topics while performing topics clusters for the set of tweet documents. It uses EM concept to update the probability topics of documents and saves the values into V. Dissimilarity values of V and RDM are computed for further assessment of a number of clusters and clustering objects for the topics of tweet documents. All these steps are described in Algorithm 4.

IV. Experimental Study

Baseline topic models [12] are enhanced as hybrid topic cluster models to separate topics from tweet documents visually. Proposed models derive topics or clusters with pre-clustering techniques and visually represented. Here we combine traditional topic models with VAT. We come up with hybrid topic models with Euclidean, cosine based and multi-viewpoint cosine similarity measures for extracting and visualizing topic clouds and health tendency of different health topics from the benchmark and Twitter datasets.

A. *The Architecture of Hybrid Topic Models*

Fig. 1 shows the architecture of proposed hybrid topic modeling. It supports the entire knowledge discovery procedure, including analysis, inference, evaluation, and applications for health data clustering. In the data layer, read the health-related data of 2-topics to 20-topics, TREC2014, and TREC2015 health-related keyword phrases are considered in the extraction of health data from Twitter. In the processing layer, it provides an implementation of VNMF, VLDA, VintJNMF, VPLSA, and VPLA topic models. Besides that, this architecture presents three aspects of how to evaluate the performance of proposed algorithms by using internal validity indexes, external validity indexes, and computational complexity. We used internal indexes DB, CHI, SI, XI, PC, PEI, and SM to measure given clustering structures, and external validity indexes CA, NMI, Precision, Recall, and F-Score to measure the fitness of data and expected structure. The total number of iterations for convergence, time for execution (in seconds), and allocated memory (in Kbs) are taken as measures in the evaluation of topic models.

B. *Datasets Description*

The experiments are executed with i7 processor, 16 GB RAM, under MATLAB 2019. Implementation details are as follows: Tweets documents and benchmark datasets are extracted with keyword phrases, preprocessed with the NLP tool. VAT used for assessment of the number of topics from tweet documents. Derived topic-document matrix is treated as input to partition matrix and derived VAT images for topics clustering of social data.

Subsets of labeled social health-related data sets from 2-topics to 20-topics are extracted from Twitter. Collected documents consist of tweets, and retweets related to healthcare are preprocessed by standard NLP tools for removing irrelevant data in tweet documents. Datasets used in our study are described in [61]. The standard bench-marked health key phrases of TREC2015 [33] and TREC2014 [34] are used for the extraction of relevant health tweets phrases are mentioned in [61].

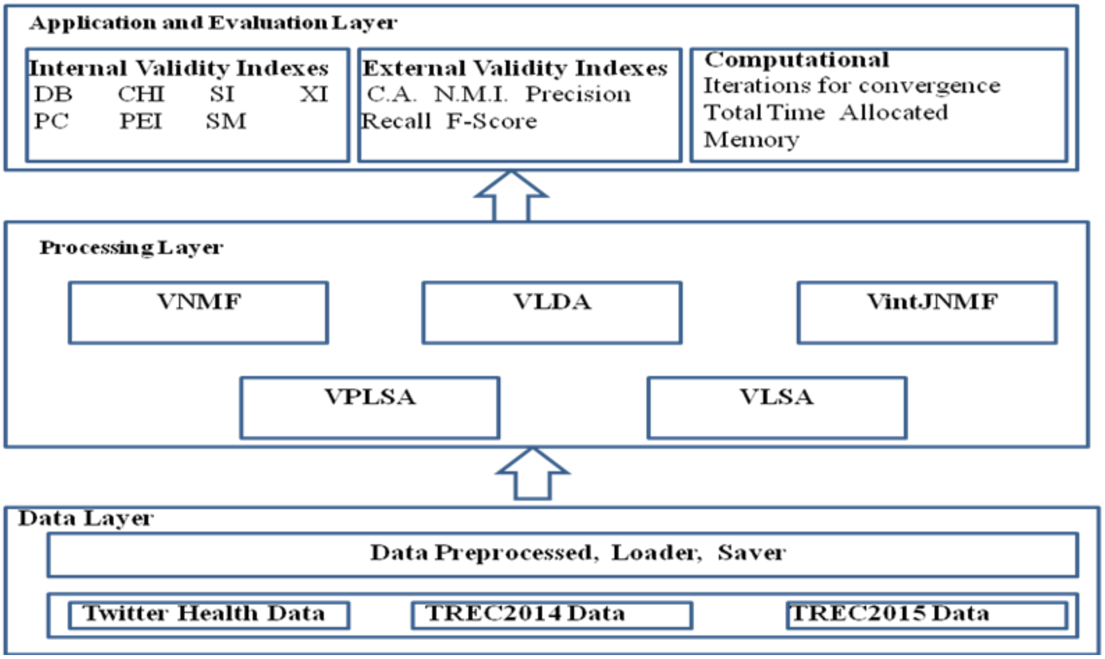


Fig. 1. Architecture of Hybrid Topic Models.

C. *Features of Hybrid Topic Algorithms Comparison*

Hybrid topic cluster modeling algorithms are compared for different datasets with three factors i.e., data set size, the number of clusters, and data set type. Datasets of [61] shows the details of these features of datasets used in our experiments and the respective result analysis shown in the following sections. For each feature, four tests are made, one for each algorithm. Results are represented in the form of tables and graphs in the following sections.

V. Visual Cluster Analysis And Discussion

A. *Topics Clouds Description*

Word cloud is a tool to describe results and trending keywords with the visual representation of word content commonly used to represent user-generated content in the respective health-related topic. It depicts the most frequently utilized words, the most important word is the biggest in size, and likewise, the size of the word decreases with each word's frequency correlated with font size. Fig. 2 shows the sample experimental word clouds formation of health data sets from Twitter, TREC2014, and TREC2015 keyword phrases data sets with following generalized steps.

1) Text is extracted from collected tweet documents.

2) Bag of words or bag of topics is extracted from preprocessed tweets.

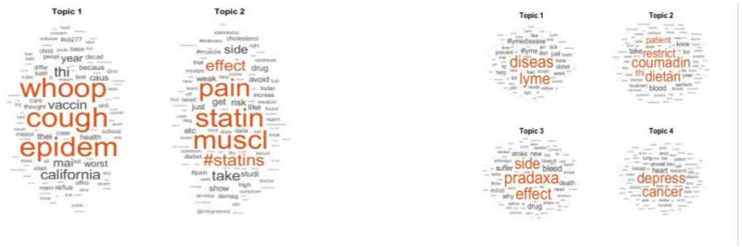
3) Infrequent words are removed from these bag of words or bag of topics.

4) By using hybrid topic models, relevant words are found and represented as a word cloud based on the number of health-related topics.

5) Words, along with the size depicting intensity, are colorfully-plotted.

B. *Assessment of Health Tendency*

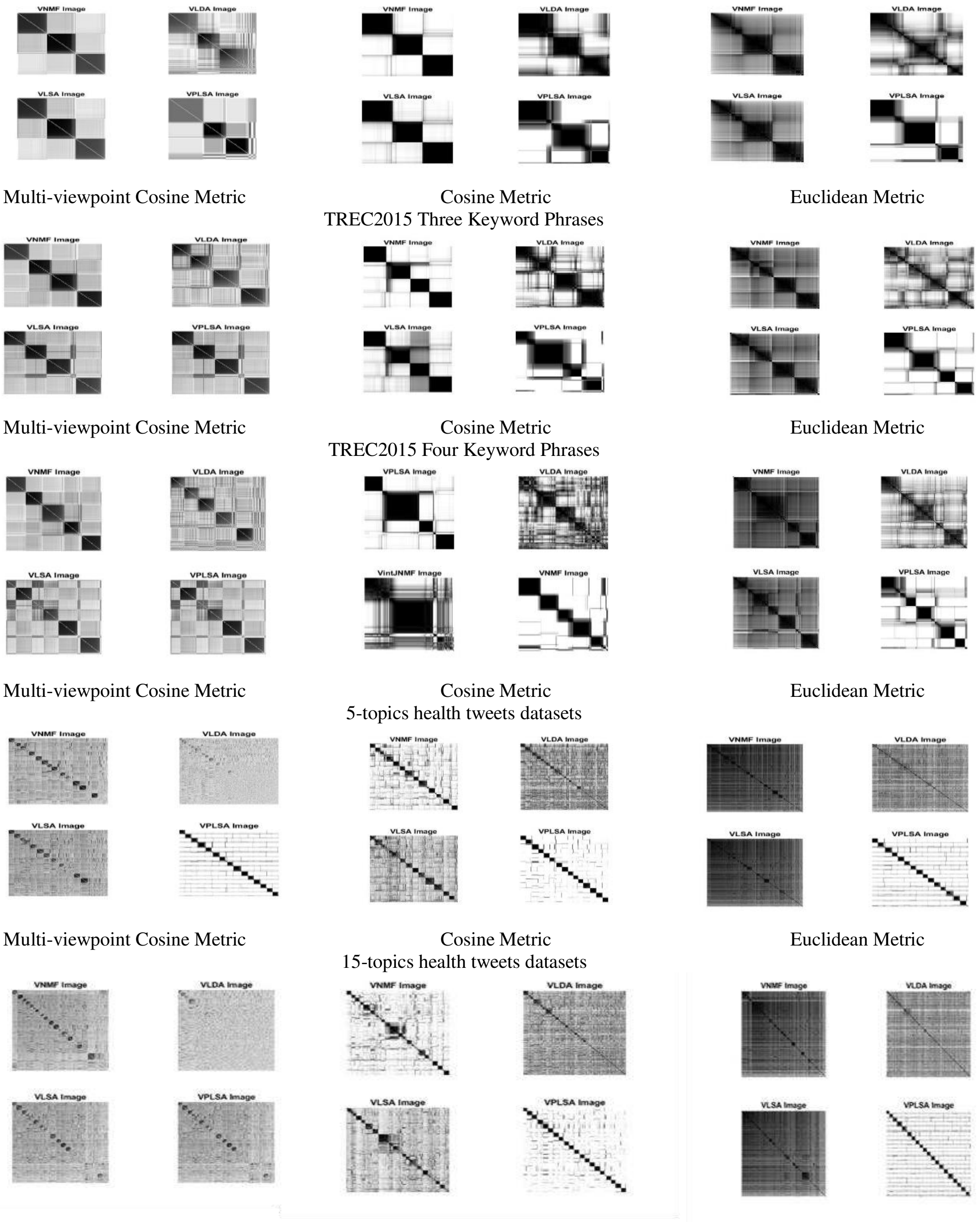
The visual evidence of topics clusters i.e., VAT images for the proposed models is shown in Fig. 3 for health tweets of TREC2015-3 keyword, TREC2015-4 keywords, 5-topics, 15-topics, and 20-topics respectively. The VAT images show the visual assessment of topics clusters for VNMF, VLDA, VLSA, and VPLSA using Euclidean, cosine based and multi-viewpoint cosine similarity measures, in which, every individual topic represented as dark square-shaped blocks along the diagonal. The quality of clusters is recognized with more clarity of square-shaped dark-colored in VAT Images. From the visual evidence of Fig. 3, it is observed that more clarity of visual topics is found in most of the health-related datasets using a multi-viewpoint cosine metric than Euclidean and cosine based metrics. Several topics increase in some cases; more clarity is observed under cosine based metrics. Significance of these visual results stated that VNMF, VLSA, and VPLSA efficiently performed for detection of health topics cluster tendency in healthcare applications and observed that VLDA shows the less clarity of visual results when compared to other models.



Word cloud of 2-topics data

Word cloud of 4 keyword phrases

Fig. 2. Sample Word Clouds of Health Data Sets.



Multi-viewpoint Cosine Metric

Cosine Metric

Euclidean Metric

TREC2015 Three Keyword Phrases

Multi-viewpoint Cosine Metric

Cosine Metric

Euclidean Metric

TREC2015 Four Keyword Phrases

Multi-viewpoint Cosine Metric

Cosine Metric

Euclidean Metric

5-topics health tweets datasets

Multi-viewpoint Cosine Metric

Cosine Metric

Euclidean Metric

15-topics health tweets datasets

Multi-viewpoint Cosine Metric

Cosine Metric

Euclidean Metric

20-topics health tweets datasets

Fig. 3. Visual Clusters of Health Tweets Datasets.

C. Performance Measures Evaluation

Our proposed mechanisms use both traditional topic models and visual technique i.e., VAT to find cluster tendency and to represent visually. A combination of VAT and topic models are used in our proposed framework (hence, they called as VNMF, VLDA, VLSI, and VPLSI) and experimented with Euclidean, cosine based and multi-viewpoint cosine based metrics. Evaluation of proposed techniques are measured with five external validity indexes, namely, clustering accuracy (CA) [41], normalized mutual information (NMI) [42], precision (P), recall (R) and F-Score (F) [43], [44] and seven internal validity indexes viz., Davies-Bouldin index (DB) [55], [56], [60], Calinski-Harabasz Index (CHI) [55], [56], Silhouette Index (SI) [55], [56], Xie-Beni Index (XI) [57], Partition Coefficient (PC) [59], Partition Entropy Index (PEI) [57], [58], and Separation Measure (SM) [60]. Health tweets are assigned to topic clusters which are maintained the highest similarity with the topic clusters to improve the value of CA. Proposed visual topic models properly compute the similarity computations. NMI [42] computes the cluster accuracy by computation of mutual information I(W;C) divided by the clusters W and classes C. Here mutual information I(W;C) statistically computed by Eq. (4).

$$I(W;C) = \sum_k \sum_j P(w_k \cap c_j) log \frac{P(w_k \cap c_j)}{P(w_k)P(c_j)} \tag{4}$$

Variables k and j denote the number of clusters in W and C. w_k represents the cluster at index k from the topic clusters W and CJ represent the specific topic cluster with index k measured in evaluated clusters by proposed topic models. Precision (P) computes the fraction of relevant topics objects among extracted topics objects; Recall (R) computes the fraction of relevant topics objects among total topics objects. Equations of precision and recall are shown in"(5)" and "(6)" respectively.

$$P = \frac{TP}{TP+FP} \tag{5}$$

$$R = \frac{TP}{TP+FN} \tag{6}$$

Whereas TP refers to the true positives, TN refers to the true negative, FP refers to the false positive, and FN refers to the false negative. F-measure (F) computes both precisions and recall harmonic means, and it is shown in "(6)"

$$F = 2\,x \frac{PxR}{P+R} \tag{7}$$

External validity indexes, i.e., CA, NMI, P, R and F-Score of 5-topics, 10-topics, 15-topics, and 20-topics health data tweets of proposed visual topic models are shown in Table I. VLSA and VPLSA under multi-viewpoint cosine distance similarity measure shows better performance than other hybrid models in case of CA and NMI external validity indexes. VNMF under multi-viewpoints performs better in precision (P), recall (R), and F-Score (F). As number topics are increased, external validity index values are decreased; however, except VLDA other three models under multi-viewpoint cosine metric maintains good clustering accuracy values.

Internal validity indexes, i.e., DB, CHI, SI, XI, PC, PEI, and SM of 5-topics, 10-topics, 15-topics, and 20-topics of proposed visual topic models are shown in Table I. It is observed that in case of DB internal validity index VLSI, VPLSI under multi-viewpoint cosine shows better performance, as the number of topics increases VNMF under multi-viewpoint shows better results. CHI and SI internal validity indexes perform well under multi-viewpoints in VLSA and VPLSA models, whereas XI, PEI, and SM under Euclidean distance perform well for VNMF and PC results are scattered under Euclidean distance. It is proved that the multi-viewpoint cosine metric is greatly succeeded in text clustering as per the overall observations of both external and internal validity indexes performance measures.

External validity indexes, i.e., CA, NMI, P, R and F-Score and internal validity indexes DB, CHI, SI, XI, PC, PEI and SM of TRC2014 and TRC2015 Health keyword phrases datasets of proposed visual topic models are shown in Table II and Table III respectively. It is observed that for both.

TREC2014 and TREC2015 datasets all external indexes perform well under multi-viewpoint cosine similarity for VNMF and VLSA models and all internal indexes except XI and CHI perform well under cosine metric similarity for VNMF.

Fig. 4 shows the comparative results with a bar graph of external validity indexes for 4-topics, 8-topics, 12-topics, and 20-topics health tweets, TREC2014 4-keyword, 3-keyword, and TREC2015 4-keyword phrases, 5-keyword phrases. Fig. 5 shows comparative results of internal indexes with line graphs of 18-topics, 20-topics, and TREC2014 and TREC2015 3-keyword and 4-keyword phrases. It shows the performance of hybrid visual topic models with Euclidean, cosine based and multi-viewpoint cosine based comparative metric analysis of VNMF, VLDA, VLSA, and VPLSA models. From this analysis, it is confirmed that VLSA, VPLSA and VNMF is capable of clustering health tweets with a better accuracy rate under multi-viewpoint cosine based similarity metrics. Comparative analysis shows that better performance values are obtained in VLSA, VPLSA, and VNMF visual topic models are more suitable topic models for accessing topics and for discovering complete clustering results of health datasets.

TABLE. I. EXTERNAL AND INTERNAL INDEXES OF HEALTH TWEETS DATASETS

Tweets Dataset C.A.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.5700	0.3550	**0.9760**	**0.9760**	0.8650	0.4100	0.9000	0.3100	0.5700	0.3550	0.9750	0.9750
10-Topics	0.6600	0.2825	**0.6700**	**0.6700**	0.6000	0.3075	0.6225	0.2550	0.6600	0.2825	0.6600	0.6600
15-Topics	**0.5250**	0.2000	0.4917	0.2100	0.5217	0.2233	0.5150	0.1917	**0.5250**	0.2000	0.4917	0.2100
20-Topics	0.4325	0.1725	**0.4800**	**0.4800**	0.4775	0.2475	0.5550	0.2013	0.4325	0.1725	**0.4800**	**0.4800**
N.M.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.5052	0.1080	**0.9318**	**0.9318**	0.7836	0.2077	0.9000	0.0949	0.5560	0.2009	0.8154	0.0941
10-Topics	0.5834	0.1501	**0.6098**	**0.6098**	0.5534	0.2321	0.6004	0.1630	0.4949	0.1891	0.4722	0.1602
15-Topics	**0.4977**	0.1487	0.4813	0.1552	0.4871	0.1641	0.4846	0.1443	0.3824	0.1743	0.3294	0.1457
20-Topics	0.4907	0.1545	**0.5211**	**0.5211**	0.5181	0.2452	0.6063	0.1976	0.3813	0.2401	0.3548	0.2107
Precision (P)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.8825	0.4679	**0.9750**	0.3180	0.8825	0.4679	1.0000	0.3259	0.8825	0.4679	0.9200	0.3259
10-Topics	**0.7655**	0.3827	0.6700	0.2410	0.6798	0.5083	0.7225	0.2842	**0.7655**	0.3827	0.5075	0.2564
15-Topics	**0.6459**	0.3156	0.4917	0.2183	**0.6459**	0.6359	0.5550	0.2078	**0.6459**	0.6359	0.3567	0.2078
20-Topics	**0.5850**	0.5650	0.4800	0.2381	0.6235	0.3117	0.5550	0.2089	0.5489	0.3591	0.3012	0.2170
Recall (R)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.8400	0.4850	0.9750	0.3200	0.8400	0.4850	1.0000	0.3250	0.8400	0.4850	0.9200	0.3250
10-Topics	**0.7700**	0.4000	0.6700	0.2450	0.7250	0.5000	0.7225	0.2850	**0.7700**	0,4000	0.5075	0.2625
15-Topics	**0.6733**	0.3100	0.4917	0.2233	0.6633	0.6533	0.5550	0.2100	**0.6733**	0.6233	0.3567	0.2100
20-Topics	**0.6125**	0.6025	0.4800	0.2463	0.6088	0.3387	0.5550	0.2162	**0.6487**	0.3725	0.3012	0.2225
F-Score (F)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.8297	0.4673	0.9750	0.3173	0.8297	0.4673	**1.0000**	0.3246	0.8297	0.4673	0.9200	0.3246
10-Topics	**0.7609**	0.3833	0.6700	0.2414	0.6867	0.4908	0.7225	0.2837	**0.7609**	0.3833	0.5075	0.2564
15-Topics	**0.6534**	0.3023	0.4917	0.2201	**0.6534**	**0.6534**	0.5550	0.2078	**0.6534**	**0.6534**	0.3567	0.2078
20-Topics	**0.6642**	0.4654	0.4800	0.2411	0.6542	0.3181	0.5550	0.2115	**0.6422**	0.3540	0.3012	0.2189
D.B.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	10.7540	64.4572	**3.4784**	**3.4784**	3.4804	11.0546	2.3716	14.4076	9.2761	21.0979	4.3503	14.3609
10-Topics	18,0747	48.0752	**14.8103**	**14.8103**	20.3867	27.7623	18.6712	17.8868	17.6975	37.4516	18.8138	23.0892
15-Topics	**26.4144**	83.4127	31.2064	31.6627	26.5696	31.1415	26.5796	27.8605	32.7241	55.3354	39.5790	35.3903
20-Topics	**33.8310**	82.9828	37.5860	37.5860	19.8077	36.7545	41.6406	34.7402	35.1299	63.0640	41.5584	39.1396
C.H.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	7.6755	0.5287	**189.750**	**189.750**	48.6197	2.8234	550.3889	2.4142	6.1710	2.8001	30.8947	3.0535
10-Topics	13.5893	1.6944	**20.5715**	**20.5715**	15.9857	2.0776	20.3201	3.4162	14.1817	1.5947	20.5714	3.0896
15-Topics	**8.5933**	1.4371	6.6995	2.1976	7.6664	1.4882	7.7097	2.1805	3,5418	1.3641	2.9697	1.9680
20-Topics	9.2787	1.0614	**10.5519**	**10.5519**	9.9539	2.0615	2.9526	2.6605	7.0669	1.1551	7.7888	2.0700
S.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	-0.0766	-0.0732	**0.7560**	0.4328	0.2717	-0.0940	-0.0912	-0.0872	-0.0883	-0.0852	-0.4355	0.3462
10-Topics	**0.0207**	-0.0966	0.0804	0.0804	-0.0090	-0.0928	-0.1038	-0.0954	-0.2759	-0.1170	-0.1012	-0.2012
15-Topics	**-0.0790**	-0.0959	-0.0979	-0.8642	-0.0859	-0.0986	-0.0997	-0.0874	-0.3685	-0.1811	-0.0929	-0.2432
20-Topics	-0.1246	-0.1932	**-0.1140**	-0.2243	-0.1976	-0.1162	-0.1174	-0.1246	-0.3277	-0.1814	-0.1898	-0.1124
X.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	17.7659	472.6124	274.7771	274.8100	21.0856	24.0764	32.1530	21.2387	0.0399	0.0426	**0.0373**	1.1370

10-Topics	5.3797	0.0006	**0.0002**	0.0006	0.0462	0.7154	0.0007	0.1338	0.0334	0.0957	0.0088	2.3679
15-Topics	0.5134	0.9549	0.3745	5.1716	0.0211	0.5424	0.5649	0.5300	**0.0018**	0.0394	0.0022	1.2323
20-Topics	1.2912	0.2041	1.5867	0.5031	0.1529	1.1463	0.0203	0.2374	**0.0027**	0.0148	0.0057	3.6486
P.C.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.7349	0.4414	0.7511	0.7511	0.7688	0.3464	0.8345	0.9024	0.5802	0.3133	0.5133	**0.9084**
10-Topics	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000	**0.2895**	0.1033	0.2624	0.1000
15-Topics	0.0667	0.0667	0.0667	0.0667	0.0667	0.0667	0.0667	0.0667	0.1688	0.0846	**0.1713**	0.0667
20-Topics	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	**0.1622**	0.0592	0.1682	0.0500
P.E.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.5935	1.1376	**0.5757**	**0.5757**	0.4937	1.3136	0.3898	0.6504	0.8667	1.3735	1.0116	0.6023
10-Topics	2.3026	2.3026	2.3026	2.3026	2.3026	2.3026	2.3026	2.3026	**1.6654**	2.2841	1.7323	2.3026
15-Topics	2.7081	2.7081	2.7081	2.7081	2.7081	2.7081	2.7081	2.7081	**2.1304**	2.5716	2.1212	2.7081
20-Topics	2.9957	2.9957	2.9957	2.8857	2.9957	2.9957	2.9957	2.9957	2.3842	2,8926	**2.3037**	2.9957
S.M.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.3370	0.5471	0.0748	0.0748	0.1064	0.3627	0.0320	0.0196	0.8494	1.3215	0.4773	**0.0213**
10-Topics	0.6133	0.0443	0.5459	2.2830	0.0008	0.0000	0.0012	0.0014	0.0076	0.0230	**0.0007**	5.7261
15-Topics	0.6538	0.0619	0.0377	2.8245	0.0033	0.0002	0.0034	0.0044	**0.0001**	0.0010	0.0034	1.8791
20-Topics	1.2911	0.0850	3.0754	0.3444	0.0357	1.8886	0.5004	0.2595	**0.0002**	0.0054	0.0065	3.9345

TABLE. II. EXTERNAL AND INTERNAL INDEXES OF TREC2014 HEALTH KEYWORD PHRASES DATASETS

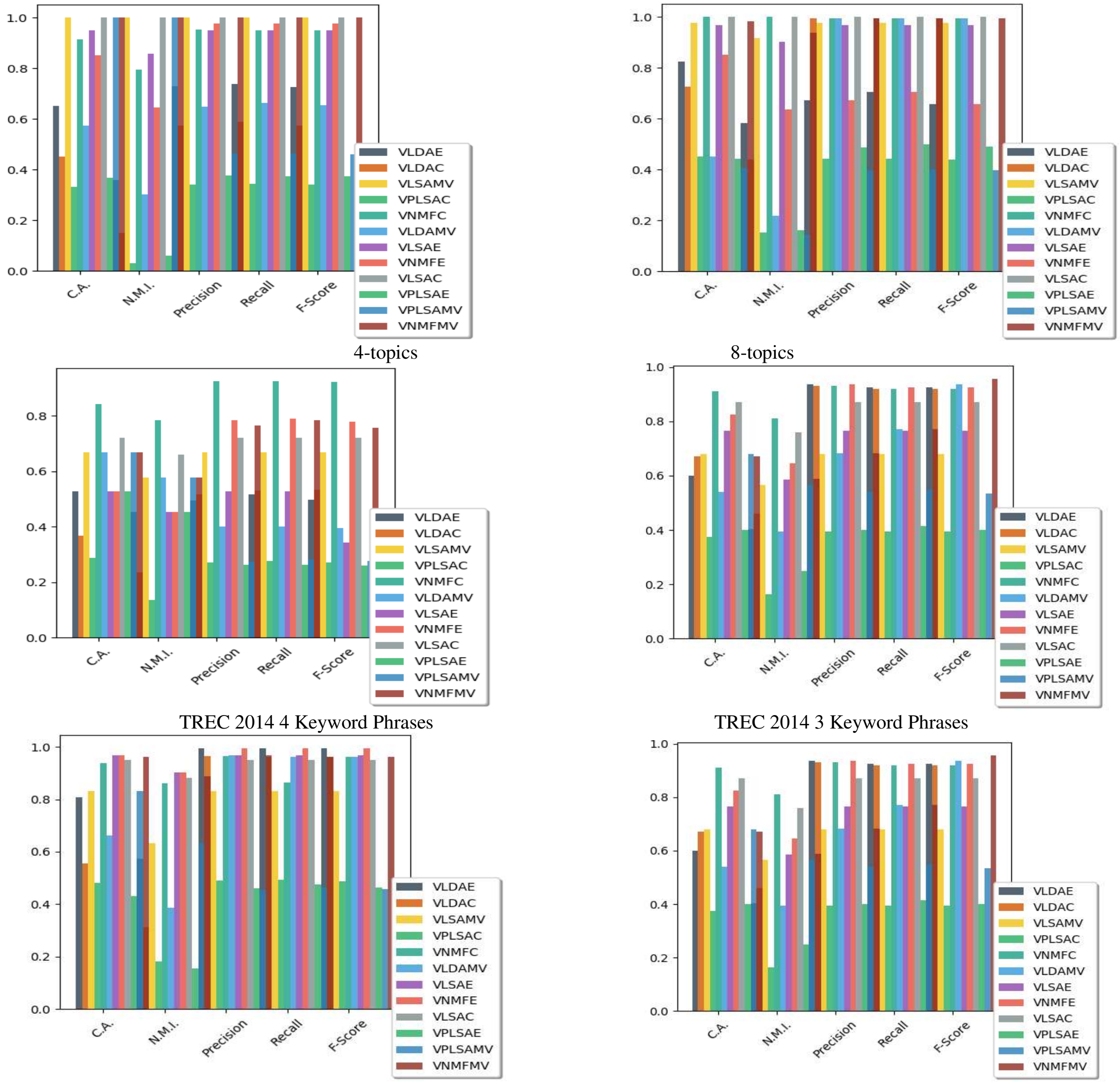
TREC2014 C.A.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.6250	**1.0000**	0.9750	**1.0000**	0.7500	**1.0000**	0.9750	0.9750	0.7000
3Keyword	**1.0000**	0.8750	**1.0000**	0.4083	**1.0000**	0.9083	**1.0000**	0.4833	0.9833	0.8917	0.9833	0.4833
4Keyword	0.9812	0.4500	0.9750	0.4063	**1.0000**	0.7250	**1.0000**	0.4500	0.8500	0.8250	0.9688	0.4437
N.M.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.0456	**1.0000**	0.8313	**1.0000**	0.1887	**1.0000**	0.8313	0.8313	0.1187
3Keyword	**1.0000**	0.6285	**1.0000**	0.0408	**1.0000**	0.7169	**1.0000**	0.0906	0.9291	0.6875	0.9291	0.0763
4Keyword	0.9368	0.2184	0.9157	0.1452	**1.0000**	0.4395	**1.0000**	0.1530	0.6362	0.5839	0.9011	0.1616
Precision (P)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.5833	**1.0000**	**1.0000**	**1.0000**	0.7941	**1.0000**	**1.0000**	**1.0000**	0.8148
3Keyword	**1.0000**	**1.0000**	**1.0000**	0.4047	**1.0000**	**1.0000**	**1.0000**	0.4602	**1.0000**	**1.0000**	0.9833	0.4602
4Keyword	0.9939	0.9939	0.9750	0.3982	0.9939	0.9939	**1.0000**	0.4413	0.6709	0.6709	0.9688	0.4865
Recall(R)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.5250	**1.0000**	**1.0000**	0.9750	0.6750	**1.0000**	**1.0000**	0.8750	0.5500
3Keyword	**1.0000**	**1.0000**	**1.0000**	0.4083	**1.0000**	**1.0000**	**1.0000**	0.4583	**1.0000**	**1.0000**	0.9833	0.4583
4Keyword	0.9939	0.9939	0.9750	0.4000	0.9938	0.9938	**1.0000**	0.4437	0.7063	0.7063	0.9688	0.5000
F-Score(F)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**1.0000**	**1.0000**	**1.0000**	0.5526	**1.0000**	**1.0000**	0.9873	0.7297	**1.0000**	**1.0000**	0.9333	0.6567
3Keyword	**1.0000**	**1.0000**	**1.0000**	0.4054	**1.0000**	**1.0000**	**1.0000**	0.4581	**1.0000**	**1.0000**	0.9833	0.4581
4Keyword	0.9937	0.9937	0.9750	0.3984	0.9937	0.9937	**1.0000**	0.4404	0.6561	0.6561	0.9688	0.4899
D.B.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.8600	0.8573	0.8712	3.3995	**0.6901**	0.7650	0.6906	1.2297	0.9290	2.8748	0.9324	2.0255
3Keyword	1.6057	2.8889	1.7053	24.8802	1.3063	1.5675	**1.3167**	4.1082	1.8455	2.1106	1.8785	5.9762
4Keyword	2.3388	19.3556	2.6151	7.9024	**1.8554**	3.8763	1.8750	6.1840	3.5706	3.9035	2.8485	5.4655

C.H.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	3.7917	1.4453	3.6289	0.0042	3.9685	0.0265	2.2770	0.0023	404.5580	378.7549	**405.7344**	8.5830
3Keyword	1.3269	0.0474	1.1660	0.0012	3.2002	0.0958	2.2716	0.0079	**195.2073**	72.2194	190.1012	6.4386
4Keyword	**346.6194**	3.4159	256.675	5.0847	1.7212	0.0184	1.2061	0.0060	34.3966	34.3527	104.3518	7.1884
S.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.9895	0.9661	0.0684	0.6542	**0.9989**	0.8004	0.8692	0.6542	0.8949	0.8591	0.0901	0.4326
3Keyword	0.9652	0.3941	-0.0605	0.7642	**0.9832**	0.5578	0.1658	0.1456	0.7648	0.4706	0.1531	0.5242
4Keyword	0.8520	-0.0693	-0.0743	0.6532	**0.9626**	0.1030	0.0658	0.0426	0.1637	0.2522	-0.0421	0.2436
X.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.3861	0.7182	0.3878	0.6867	**0.0386**	1.8074	0.0657	1.6163	1.9705	1.2353	3.6027	1.6387
3Keyword	115.3133	64.260	**2.1006**	5.3383	14.3863	20.4709	25.9446	17.6322	94.034	30.1596	40.0134	29.9649
4Keyword	0.0994	1.0878	0.0232	**0.0334**	0.5471	0.1518	1.3669	4.6510	481.1688	33.1056	155.6778	210.9156
P.C.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**0.9899**	0.9697	0.9894	0.9608	**0.9989**	0.9291	0.9982	0.9387	0.9221	0.9470	0.9449	0.9272
3Keyword	0.9452	0.8571	0.9383	0.9082	**0.9781**	0.8620	0.9689	0.9683	0.8514	0.8304	0.8476	0.9589
4Keyword	0.8843	0.6080	0.8709	0.8585	**0.9534**	0.7126	0.9340	0.9059	0.7382	0.6849	0.7703	0.8727
P.E.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.0314	0.0683	0.0326	0.0763	**0.0036**	0.1304	0.0062	0.0998	0.1401	0.1015	0.1068	0.1169
3Keyword	0.1385	0.2964	0.1532	0.1985	0.0488	0.2774	0.0731	**0.0647**	0.2869	0.3375	0.2949	0.0824
4Keyword	0.2808	0.7751	0.3087	0.3249	**0.1070**	0.5851	0.1546	0.1995	0.5342	0.6416	0.4751	0.2563
S.M.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	0.0378	0.0413	0.0393	0.0431	**0.0258**	0.0383	0.0269	0.0326	0.0840	0.0395	0.0899	0.0360
3Keyword	0.0453	0.0993	0.0475	0.0412	**0.0224**	0.0469	0.0259	0.0255	0.0936	0.0569	0.0956	0.0263
4Keyword	0.0527	0.1942	0.0629	0.0524	**0.0227**	0.0541	0.0269	0.0266	0.5563	0.0764	0.1530	0.0289

TABLE. III. EXTERNAL AND INTERNAL INDEXES OF TREC2015 HEALTH KEYWORD PHRASES DATASETS

TREC2015 C.A.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.8750	**1.0000**	**1.0000**	**1.0000**	0.8500	**1.0000**	0.9500	**1.0000**	0.7750
3keyword	**1.0000**	0.7750	**1.0000**	0.7583	**1.0000**	0.7048	**1.0000**	0.4996	**1.0000**	0.8083	**1.0000**	0.5333
4keyword	0.9625	0.6625	0.8313	0.8313	0.9375	0.5563	0.9500	0.4813	**0.9688**	0.8063	**0.9688**	0.4313
5keyword	0.6700	0.5400	0.6800	0.6800	**0.9100**	0.6700	0.8700	0.3750	0.8250	0.6000	0.7650	0.4000
N.M.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.4564	**1.0000**	**1.0000**	**1.0000**	0.3902	**1.0000**	0.7136	**1.0000**	0.2308
3keyword	**1.0000**	0.4479	**1.0000**	0.4530	**1.0000**	**1.0000**	**1.0000**	0.4996	**1.0000**	0.4942	**1.0000**	0.1901
4keyword	0.8865	0.3864	0.6339	0.6339	0.8614	0.3127	0.8828	0.1813	0.9011	0.5739	**0.9011**	0.1552
5keyword	**0.5879**	0.3942	0.5645	0.5645	0.8119	0.4590	0.7591	0.1636	0.6456	0.4027	0.5858	0.2490
Precision (P)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.8974	**1.0000**	**1.0000**	**1.0000**	0.8462	**1.0000**	**1.0000**	**1.0000**	0.7750
3keyword	**1.0000**	**1.0000**	**1.0000**	0.7283	**1.0000**	**1.0000**	**1.0000**	0.7608	**1.0000**	**1.0000**	**1.0000**	0.4943
4keyword	0.9661	0.9661	0.8312	0.4558	0.9637	0.9637	0.9500	0.4913	**0.9939**	**0.9939**	0.9688	0.4620
5keyword	0.6830	0.6830	0.6800	0.5407	0.9321	0.9321	0.8700	0.3949	**0.9366**	**0.9366**	0.7650	0.3993
Recall (R)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.8750	**1.0000**	**1.0000**	**1.0000**	0.8250	**1.0000**	**1.0000**	**1.0000**	0.7750

3keyword	**1.0000**	**1.0000**	**1.0000**	0.7333	**1.0000**	**1.0000**	**1.0000**	0.7583	**1.0000**	**1.0000**	**1.0000**	0.5250
4keyword	0.9625	0.9625	0.8312	0.4625	0.8625	0.9625	0.9500	0.4937	**0.9938**	**0.9938**	0.9688	0.4750
5keyword	0.7700	0.7700	0.6800	0.5500	0.9200	0.9200	0.8700	0.3950	**0.9250**	**0.9250**	0.7650	0.4150
F-Score (F)	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**1.0000**	**1.0000**	**1.0000**	0.8861	**1.0000**	**1.0000**	**1.0000**	0.8354	**1.0000**	**1.0000**	**1.0000**	0.7750
3keyword	**1.0000**	**1.0000**	**1.0000**	0.7303	**1.0000**	**1.0000**	**1.0000**	0.7562	**1.0000**	**1.0000**	**1.0000**	0.4931
4keyword	0.9619	0.9619	0.8313	0.4562	0.9621	0.9621	0.9500	0.4883	**0.9937**	**0.9937**	0.9688	0.4649
5keyword	**0.9562**	0.9356	0.6800	0.5347	0.9201	0.9201	0.8700	0.3942	0.9258	0.9258	0.7650	0.4006
D.B.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.8268	0.8784	0.8784	1.0548	**0.6930**	0.7328	0.6939	0.9486	0.8741	0.8871	0.8791	1.5546
3keyword	1.7844	5.7403	1.8031	3.1419	**1.3141**	1.6395	1.3205	1.9608	1.9231	2.7790	1.9400	8.7497
4keyword	2.5904	8.0642	4.5376	4.5376	**2.2253**	6.1540	2.3094	10.2185	2.7870	3.9111	2.8485	7.6919
5keyword	12.2397	16.9242	14.3625	14.3625	**3.2036**	7.6354	3.6686	7.6100	4.7661	12.7339	6.1253	8.8636
C.H.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	3.1217	0.9943	0.8843	0.1132	8.2324	0.6959	0.4773	0.0604	**609.161**	197.676	598.464	20.0786
3keyword	1.1547	0.0125	1.1464	0.0354	3.1818	0.0884	2.1479	0.0329	**183.429**	35.5853	181.120	4.3831
4keyword	215.57	10.9328	44.850	44.850	166.944	7.7988	**220.252**	6.3097	107.242	30.7358	104.351	9.0711
5keyword	24.652	16.2431	18.384	18.384	**122.026**	30.8903	74.6247	5.7980	31.9652	8.3416	25.0822	8.3665
S.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.9868	0.9358	0.6325	0.7642	**0.9944**	0.9228	0.5117	0.4562	0.9357	0.7539	0.2693	0.4328
3keyword	0.9589	0.0844	0.3051	0.8642	**0.9841**	0.5253	0.2400	0.2326	0.7898	0.2884	-0.0279	0.3452
4keyword	**0.7600**	-0.004	0.3390	0.4562	0.6952	-0.0412	-0.0729	-0.0652	0.6528	0.2322	-0.0395	0.2346
5keyword	0.0526	0.0432	-0.0467	0.0456	**0.5608**	0.1192	0.3533	0.2542	0.1978	-0.0553	-0.0642	0.0568
X.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.3658	1.0215	1.0215	1.3430	**0.1731**	1.3944	0.2150	2.1933	1.4691	1.3686	1.4751	2.5907
3keyword	**1.8426**	30.1387	1.9038	8.7804	29.7477	102.25	46.6774	301.5592	61.8303	125.707	117.7518	29.9772
4keyword	13.8912	121.3485	**13.1335**	**13.1335**	354.4803	77.7253	168.7514	102.0204	147.0701	77.3134	155.6775	110.7712
5keyword	1.4976	0.0041	**0.0007**	**0.0007**	0.3305	0.8615	0.0404	2.0835	0.4388	0.0599	0.1613	5.8658
P.C.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.9871	0.9504	0.9504	0.9311	**0.9944**	0.9465	0.9930	0.9150	0.9493	0.9400	0.9486	0.8841
3keyword	0.9374	0.8412	0.9367	0.8887	**0.9770**	0.8611	0.9667	0.8774	0.8092	0.7871	0.8081	0.8872
4keyword	0.8821	0.7395	0.8588	0.8588	**0.9417**	0.6818	0.9255	0.8929	0.7761	0.6586	0.7703	0.8848
5keyword	0.7606	0.6287	0.7778	0.7778	**0.9105**	0.6208	0.8683	0.8800	0.6637	0.4187	0.6734	0.8253
P.E.I.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.0373	0.0967	0.0967	0.1245	**0.0136**	0.0943	0.0173	0.1402	0.1018	0.1086	0.1032	0.1942
3keyword	0.1531	0.3230	0.1559	0.2328	**0.0530**	0.2812	0.0752	0.2283	0.3645	0.4129	0.3663	0.2132
4keyword	0.2854	0.5458	0.3283	0.3283	**0.1276**	0.6388	0.1687	0.2215	0.4628	0.6871	0.4751	0.2316
5keyword	0.4794	0.7980	0.5098	0.5098	**0.2102**	0.8050	0.2982	0.2558	0.7189	1.1633	0.6974	0.3793
S.M.	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	0.0360	0.0429	0.0429	0.0444	**0.0267**	0.0332	0.0281	0.0352	0.0723	0.0391	0.0737	0.0451
3keyword	0.0522	0.0906	0.0514	0.0606	**0.0212**	0.0379	0.0231	0.0287	0.1296	0.0575	0.1316	0.0495
4keyword	0.0609	0.1215	0.0737	0.0737	**0.0236**	0.1017	0.0490	0.0303	0.1479	0.0749	0.1530	0.0271
5keyword	5.2845	0.8762	0.2876	0,0458	0.0281	0.0855	0.1334	**0.0214**	0.2848	0.9381	0.3779	0.0311



4-topics

8-topics

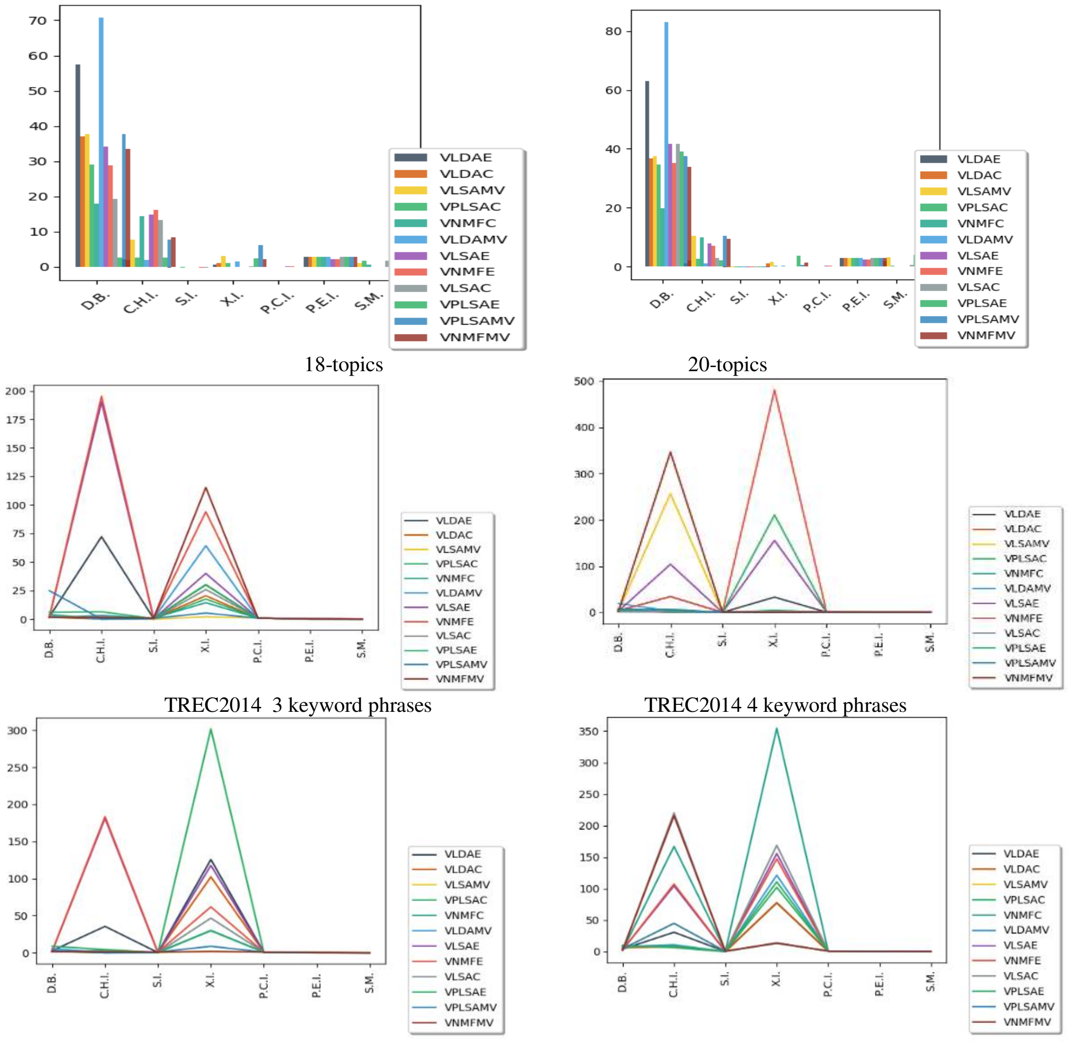
TREC 2014 4 Keyword Phrases

TREC 2014 3 Keyword Phrases

TREC 2015 4 Keyword Phrases

TREC 2015 5 Keyword Phrases

Fig. 4. Bar Graphs of External Validity Indexes of 4-Topics, 8-Topics, 12-Topics, 20-Topics, TREC2014-4, 3Keyword Phrases, and TREC2015-4 and 5 Keyword Phrases Datasets.



18-topics

20-topics

TREC2014 3 keyword phrases

TREC2014 4 keyword phrases

TREC2015 3 keyword phrases

TREC2015 4 keyword phrases

Fig. 5. Line Graphs of Internal Indexes of TREC2014 4, 3 Keyword Phrases and TREC2015 4, 3Keyword Phrases.

D. Convergence Study

Table IV shows the total number of iterations required for the execution of hybrid visual topic models with an error tolerance of 0.000001 for convergence of 2-topics to 20-topics of twitter health datasets and TREC2014 and TREC2015 2, 3, 4 and 5-keyword phrases of data sets under multi-viewpoints, cosine and Euclidean metrics.

Bar graphs are represented in Fig. 6 for 2-topics to 20-topics health datasets and TREC2014 and TREC2015 keyword phrases. On overall observations in all types of datasets, the total number of iterations required within error tolerance is less under multi-viewpoints cosine distance metric than Euclidean and cosine based metric. There are only a few cases where less number of iterations required under Euclidean distance such as 13-topics, 15-topics, 16-topics, and 20-topics.

E. Computational Complexity Analysis

Computational time in seconds and allocated memory in Kb are measured for evaluating the computational complexity of visual topic models. Table V presents the total computation time of visual topic models for various subsets, include 5-topics, 10-topics, 15-topics, and 20-topics of datasets, TREC2014, and TRC2015 all keyword phrases datasets. It presents the computation time using

Euclidean, cosine based and multi-viewpoint cosine based distance metrics for comparative analysis purpose. It is noted that VNMF under multi-viewpoint is taking less amount of time when the number of topics is less than five. As the number of topics increasing VLSA under cosine based metric taking less time than other models. Hence, VLSA under cosine based metric is the time-efficient model for topics clustering.

Fig. 7 shows the computational time analysis using bar graphs under Euclidean, cosine based and multi-viewpoints cosine based metrics. In the cosine metric, it is observed that VPLSA has taken less computational time than other models for 5-topics to 20-topics health data and TREC2014 and TREC2015 datasets.

TABLE. IV. TOTAL NUMBER OF ITERATIONS FOR CONVERGENCE

Under Multi-viewpoints

2topics	**3topics**	**4topics**	**5topics**	**6topics**	**7topics**	**8topics**	**9topics**	**10topics**	**11topics**	**12topics**	**13topics**	**14topics**	**15topics**	**16topics**	**17topics**	**18topics**	**19topics**	**20topics**
30	39	**37**	**59**	57	**63**	83	72	**71**	81	**90**	75	**86**	100	123	125	**110**	**111**	134

Under Cosine

2topics	**3topics**	**4topics**	**5topics**	**6topics**	**7topics**	**8topics**	**9topics**	**10topics**	**11topics**	**12topics**	**13topics**	**14topics**	**15topics**	**16topics**	**17topics**	**18topics**	**19topics**	**20topics**
29	**30**	41	68	**50**	69	**75**	**63**	92	**73**	106	80	111	90	107	113	118	112	116

Under Euclidean

2topics	**3topics**	**4topics**	**5topics**	**6topics**	**7topics**	**8topics**	**9topics**	**10topics**	**11topics**	**12topics**	**13topics**	**14topics**	**15topics**	**16topics**	**17topics**	**18topics**	**19topics**	**20topics**
32	44	46	68	63	69	86	75	82	89	92	**74**	94	**90**	**101**	**96**	154	116	**114**

TREC2014 Keyword Phrases									TREC2015 Keyword Phrases											
Multi-viewpoint			Cosine			Euclidean			Multi-viewpoint				Cosine				Euclidean			
2key	**3key**	**4key**	**2key**	**3key**	**4key**	**2key**	**3key**	**4key**	**2key**	**3key**	**4key**	**5key**	**2key**	**3key**	**4key**	**5key**	**2key**	**3key**	**4key**	**5key**
26	**50**	**56**	59	56	80	76	57	70	**30**	**48**	**50**	**56**	35	75	54	70	81	62	80	69

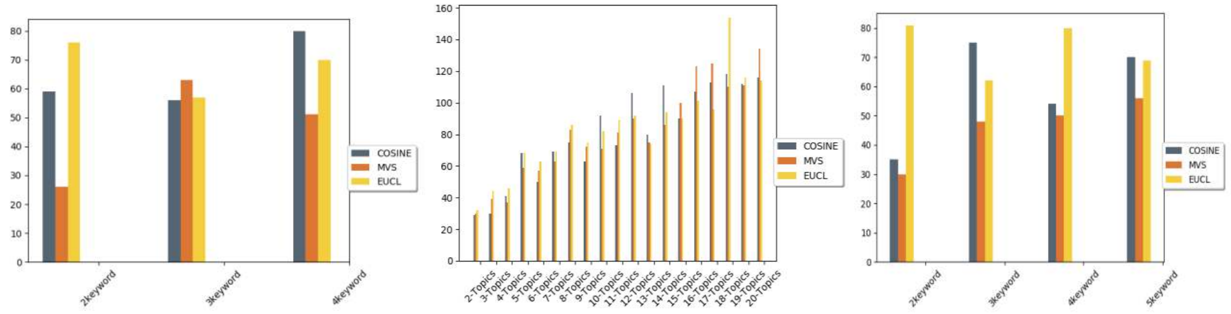


Fig. 6. Total Number of Iterations for Convergence of 2-Topics to 20-Topics, TREC2014, and TREC2015 Keyword Phrases.

TABLE. V. TOTAL TIME (SEC) TAKEN OF HEALTH TWEETS DATASETS FROM TWITTER

Tweets Dataset	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
5-Topics	0.080	0.118	0.054	0.860	0.336	0.313	**0.045**	1.042	0.394	0.289	0.153	1.094
10-Topics	0.331	0.767	0.147	5.860	0.703	0.475	**0.061**	7.722	0.716	0.589	0.137	6.786
15-Topics	1.716	1.621	0.265	24.031	2.689	1.056	**0.206**	22.158	2.683	0.986	**0.177**	22.291
20-Topics	1.704	0.856	0.130	37.928	5.492	0.845	**0.125**	39.860	4.358	1.154	0.124	38.964
TREC2014	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	**0.032**	0.087	0.048	0.041	0.052	0.084	0.051	0.076	0.148	0.261	0.103	0.102
3Keyword	**0.117**	0.304	0.122	0.198	0.156	0.135	0.137	0.192	0.159	0.261	0.136	0.146
4Keyword	**0.129**	0.252	0.130	0.453	0.214	0.257	0.194	0.842	0.215	0.159	0.161	0.726
Trec2015	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	**0.305**	0.365	0.357	0.356	0.376	0.498	0.388	0.367	0.326	0.312	0.355	0.321
3keyword	**0.052**	0.174	0.060	0.271	0.134	0.286	0.084	0.279	0.115	0.160	0.094	0.277
4keyword	**0.071**	0.445	0.059	0.503	0.193	0.132	0.073	0.430	0.144	0.184	0.061	0.648
5keyword	0.267	0.796	0.177	0.932	0.269	0.296	**0.110**	1.204	0.423	0.564	0.140	0.453

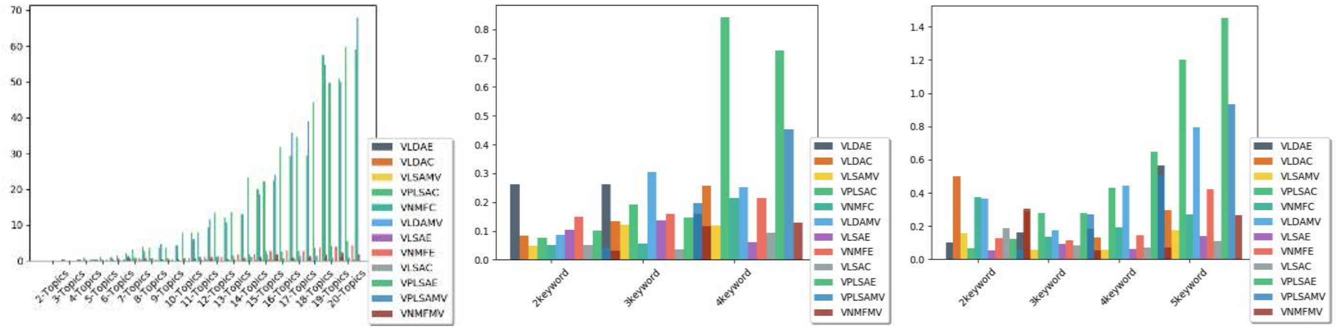


Fig. 7. Total Time is Taken (sec) for the Execution of 2-Topics to 20-Topics of Twitter Datasets and TREC2014 Keyword Phrases under Multi-Viewpoint Cosine based, Cosine based and Euclidean Distances.

A memory requirement is another important criterion for calculating computational complexity. In the experimental analysis, the memory allocated for successful running visual topic models are measured in terms of Kb for 2-topics to 20-topics, TREC2014 and TREC2015 datasets under three different similarity measures, i.e., Euclidean, cosine based, and multi-viewpoint cosine distance values are tabulated in Table VI. It is observed that VLSA under cosine based is more memory efficient than other visual hybrid models.

Fig. 8 shows the memory requirement comparison for visual topic models, and it is evaluated for 2-topics to 20-topics, TREC2014, and TREC2015 using Euclidean, cosine based, and multi-viewpoints cosine based distance that of other visual topic models, though VLSA and VNMF visual model performs well in other aspects i.e., results visually, which are more suitable for our further work in developing smart healthcare applications. Our proposed visual topic models outperform with other traditional topic models in two aspects; VLSI under cosine has taken less amount of space than other visual topic models in most of the health topics. VLSI performs well concerning space complexity than other models.

TABLE. VI. ALLOCATED MEMORY (KB) OF HEALTH TWEETS DATASETS FROM TWITTER

Tweets Dataset	Multi-viewpoint Cosine				Cosine				Euclidean			
	VNMF	VLDA	VLSA	VPLSA	VNMF	VLDA	VLSA	VPLSA	VNMF	VLDA	VLSA	VPLSA
5-Topics	14988	72	**0**	648948	37776	36736	6988	750960	29440	14312	1024	752208
10-Topics	43976	8084	2460	5268368	24944	52	**0**	6808840	62504	34640	2040	6090784
15-Topics	52248	17980	**4420**	23677472	88388	22912	4424	21312380	87696	22244	5284	21319836
20-Topics	26376	**156**	3812	68096596	80436	284	3816	58952964	79044	164	3812	57937436
TREC2014	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2Keyword	692	**0**	**0**	**0**	**0**	**0**	68	**0**	13312	5656	1024	3336
3Keyword	14468	43880	6892	146116	**0**	**0**	**0**	123588	14708	26916	11516	85612
4Keyword	24104	24544	1024	266248	6036	12	1140	397804	8276	684	**0**	348036
TREC2015	**Multi-viewpoint Cosine**				**Cosine**				**Euclidean**			
	VNMF	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**	**VNMF**	**VLDA**	**VLSA**	**VPLSA**
2keyword	13196	4388	2052	2528	12652	17308	1024	2132	**0**	**0**	68	64
3keyword	2180	68	**0**	133840	100	1524	1056	143400	2920	300	4	102472
4keyword	1140	1504	**0**	227408	3392	**0**	**0**	208868	3720	**0**	**0**	397396
5keyword	12788	64648	2488	501600	5304	**0**	**0**	620012	5340	**0**	**0**	611308

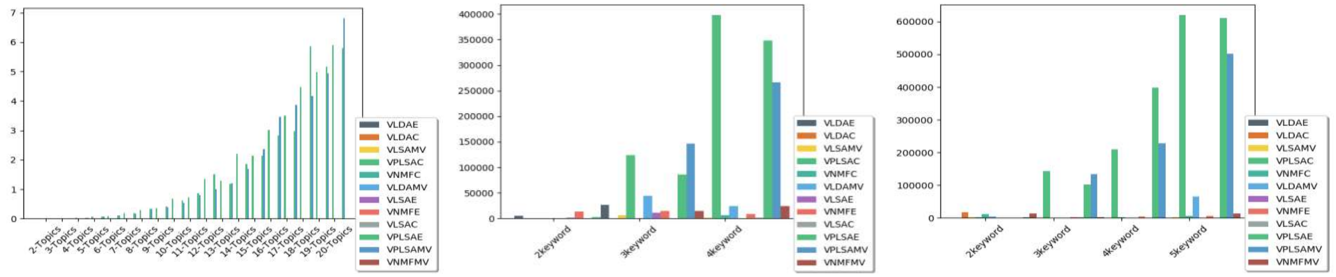


Fig. 8. Total Time is Taken (sec) for the Execution of 2-Topics to 20-Topics of Twitter Datasets and TREC2014 Keyword Phrases under Multi-Viewpoint Cosine based, Cosine based and Euclidean Distances.

First, cluster tendency or prior knowledge about social data is unknown in existing topic models, whereas our models assess the prior information of social data clusters visually without external interference. Secondly, our visual topic models effectively deal with a large amount of unlabeled social data in determining the number of clusters (or topics) visually. In our work, cluster validity is assessed by both internal and external cluster index measures with Euclidean, cosine based, and multi-viewpoints cosine based similarity measures.

VI. Conclusion and Future Work

Topic models are capable of finding hidden conceptual topics from such a vast number of terms of tweet documents. The topic models, LSI, PLSI, NMF, LDA, and intJNMF determines the topic clusters without knowledge of cluster tendency. Hybrid topic models overcome the problem of health cluster tendency and improve the performance of topic clustering. The empirical analysis of proposed hybrid topic model techniques is performed based on the parameters, such as convergence speed, time, and space computational complexities. In our experimental study, cosine based hybrid topics models much succeeded for the detection of hidden concepts or topics from tweets documents and observed numerical improvement with an increased rate of 40% to 45% over 2-topics to 10-topics and 35% to 30% over 11-topics to 20-topics. Among proposed visual topic models, VLSA and VNMF under multi-viewpoints cosine show the best performance in finding both numbers of topics for unlabeled twitter data and tweets clustering internal and external validity indexes, total numbers of iterations for convergence and less time is taken for execution. Hence, VNMF and VLSA show better performance in finding both the number of topics as well as clustering results for unlabeled twitter data and tweets. Based on the observation of space complexity in experimental, it needs to be improved as scalable visual topic models in future work for performing effective big data health clustering.

Acknowledgment

This work is supported by the Science & Engineering Research Board (SERB), Department of Science and Technology, Government of India for the Research Grant of DST project number ECR/2016/001556.

References

[1] J. Vosecky, D. Jiang, K. W.-T. Leung, K. Xing, and W. Ng, "Integrating social and auxiliary semantics for multifaceted topic modelling in twitter, " ACM Transactions on Internet Technology (TOIT), vol. 14, no. 4, pp. 27, 2014.

[2] Chen Z, and Liu B, "Mining topics in documents: standing on the shoulders of big data, " in Proc. 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1116-1125, Aug. 2014.

[3] Amrutha Benny and Mintu Phili, " Keyword Based Tweet Extraction and Detection of Related Topics, " ICICT2014, vol. 46, pp. 364-371, 2015.

[4] McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," Computer Science Department Faculty Publication Series, pp. 3, 2005.

[5] Vandana Singh, and Sanjay Kumar Dubey, " Opinion mining and analysis: A literature review, " 2014 IEEE conference, pp. 25-26, Sep. 2014.

[6] Menna Allah Hassan, " A comparative study of classification algorithm in e-health environment, " 2016 IEEE conference (ICDIPC), April 2016.

[7] Z. He, S. Xie, R. Zdunek, G. Zhou, and A. Cichocki, "Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering," IEEE Transactions on Neural Networks, vol. 22, no. 12, pp. 2117–2131, 2011.

[8] Kolini, Farzan, Janczewski, and Lech, "Clustering and Topic Modelling: A New Approach for Analysis of National Cybersecurity Strategies," PACIS 2017 Proceedings no. 126, pp.1-12, 2017.

[9] James C Bezdek " VAT: A tool for visual assessment of cluster tendency, " IJCNN2002, Feb 2002, DOI: 10.1109/IJCNN.2002 .1007487.

[10] Wuhan, "TF-IDF based Feature Words Extraction and Topic Modeling for Short Text," ICMSS2018, Jan. 2108, DOI: 10.1145/3180374.3181354.

[11] Alexandra Schofield " Pulling out the stop: Rethinking stop words removal for topic models, " 15th conference of ACM, pp:432-36, April 2107.

[12] Amritanshu Agrawal " What is Wrong with Topic Modeling? (and How to Fix it Using Search-based SE), " IEEE transaction on software engineering, Aug 2016, DOI: 10.1016/j.infsof.2018.02.005.

[13] Dredze M "How Social Media Will Change Public Health," IEEE Intelligent Systems, Vol.27, no.4, pp.81–84, Aug. 2012.

[14] Dr Choudhury M, Gamon M, Counts S, and Horvitz E, "Predicting Depression via Social Media," in Proc. 7th International Conference on Weblogs and Social Media, 2013.

[15] Kumar, D., Bezdek, J. C., Palaniswami, M., Rajasegarar, S., Leckie, C., & Havens, T. C. (2016). A Hybrid Approach to Clustering in Big Data. IEEE Transactions on Cybernetics, 46(10), 2372–2385. DOI:10.1109/tcyb.2015.2477416.

[16] K. Rajendra Prasad and M. Suleman Basha, "Improving the performance of speech clustering method," 10th International Conference on Intelligent Systems and Control (ISCO), 2016, DOI: 10.1109/ISCO.2016.7726878.

[17] Eswara Reddy B, and K. Rajendra Prasad, "Improving the performance of visualized clustering method," International Journal of Systems Assurance Engineering and Management, Vol. 7, pp.102-111, Dec.2016.

[18] Y. Hu, A. John, F. Wang, and S. Kambhampati, "Et-LDA: Joint topic modelling for aligning events and their twitter feedback, " in AAAI Conference on Artificial Intelligence (AAAI 2012), vol. 12, Toronto, Ontario, Canada, pp. 59–65, July 2012.

[19] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modelling based on interactive nonnegative matrix factorization," IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pp. 1992–2001, 2013.

[20] X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang, "Learning topics in short texts by non-negative matrix factorization on term correlation matrix," in Proceedings of the SIAM International Conference on Data Mining (SIAM 2013, San Diego, California, USA: SDM, July 2013.

[21] M. Albakour, C. Macdonald, and I. Ounis, "On scarcity and drift for effective real-time filtering in microblogs," in Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM2013), pp.419–428, October 2013.

[22] J. Li, Z. Tai, R. Zhang, W. Yu, and L. Liu, "Online busty event detection from microblog," in Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference, pp. 865–870, Dec. 2014.

[23] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models, " The International AAAI Conference on Web and Social Media (ICWSM), vol. 10, pp. 130–137, May 2010.

[24] Yan, X, and Guo, J, " Clustering Short Text Using Ncut-weighted Non-negative Matrix Factorization, " in Proc. CIKM 2012, Miami, HI, USA, pp. 2259–2262, 2012.

[25] Yan, X, and Guo, J, "Learning Topics in short text Using Ncut-weighted non-negative matrix Factorization on term correlation matrix," http://xiaohuiyan.com/papers/TNMF-SDM-13.pdf

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

[27] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, " Indexing by latent semantic analysis, " Journal of American Society for Information Sciences, Vol.41, no.6, pp:391-407, 1990.

[28] Singular Value Decomposition [Online] Available: web.mit .edu/be.400/www/SVD/Singular_Value_Decomposition.htm.

[29] Thomas Hofmann, "Probabilistic latent semantic indexing," SIGIR, ACM, pp:50-57, 1999.

[30] Dempster A, Laird N, and Rubin D, "Maximum likelihood from incomplete data via the EM algorithm," J. royal statistics society B 39, pp: 1-38, 1977.

[31] Madhuri Avula, Narasimha Prasad Lakkakula, and Murali Prasad Raja" Bone cancer detection from mri scan imagery using mean pixel intensity" IEEE 8th Asia Modelling Symposium, pp:141-146, 2014.

[32] PR Anisha, C Kishor Kumar Reddy, and LV Narasimha Prasad" A pragmatic approach for detecting liver cancer using image processing and data mining techniques" IEEE International Conference on Signal Processing and Communication Engineering Systems, pp: 352-357, 2015.

[33] TREC2015 https://trec.nist.gov/pubs/trec24/trec2015.html

[34] TREC2014 https://trec.nist.gov/pubs/trec23/trec2014.html.

[35] C. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval ", online edition, vol. 1, Cambridge, 2008, https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

[36] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," Advances in Neural Information Processing Systems 13, NIPS 2000, Denver, CO, USA, pp. 556–562, 2000.

[37] R. Nugroho, J. Yang, Y. Zhong, C. Paris, and S. Nepal, "Deriving topics in twitter by exploiting tweet interactions," Proceedings of the 4th IEEE International Congress on Big Data, New York, USA: IEEE Services Comp. Community, July 2015.

[38] R. Pochampally and V. Varma, "User context as a source of topic retrieval in twitter," Workshop on Enriching Information Retrieval (with ACM SIGIR). Beijing, China: ACM, pp. 1–3, July 2011.

[39] Paul M, and Girju R, "A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics," Proc.24th AAAI-10 Conference on Artificial Intelligence, Atlanta, USA, 2010.

[40] Robertus Nugroho, Jian Yang, Weiliang Zhao, Cecile Paris, and Surya Nepal, "What and With Whom? Identifying Topics in Twitter Through Both Interactions and Text ", Journal of Latex Class Files, Vol.14, No.8, August 2015.

[41] Pattanodom, M., I am-On, N., and Boongoen, T. " Clustering data with the presence of missing values by ensemble approach," 2016 Second Asian Conference on Defense Technology (ACDT). doi:10.1109/acdt.2016.7437660.

[42] Alessia Amelio and Clara Pizzuti, " Is Normalized Mutual Information a Fair Measure for Comparing Community Detection Methods? , " IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.

[43] 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, 7, 58407–58418. DOI:10.1109/access.2019.2914097.

[44] Li, Z., Shang, W., & Yan, M. (2016). News text classification model based on-the topic model. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).doi:10.1109/icis.2016.7550929.

[45] I. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data USING clustering," Machine Learning, vol. 42, nos. ½, pp. 143-175, Jan. 2001.

[46] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering," Proc. IEEEInt'lConf.DataMining(ICDM),pp.107-114,2001.

[47] H.Zha, X.He, C.Ding, H.Simon, and.Gu, "SpectralRelaxation for K-Means Clustering," Proc. Neural Info. Processing Systems (NIPS), pp. 1057-1064,2001.

[48] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug.2000.

[49] I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 269-274,2001.

[50] Y. Gong and W. Xu, Machine Learning for Multimedia Content Analysis. Springer-Verlag,2007.

[51] A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-Page Clustering," Proc. 17th Nat'l Conf.Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI), pp. 58-64, July 2000.

[52] D. Ienco, R.G. Pensa, and R. Meo, "Context-Based Distance Learning for Categorical Data Clustering," Proc. Eighth Int'l Symp. IntelligentDataAnalysis(IDA),pp.83-94,2009.

[53] P. Lakkaraju, S. Gauch, and M. Speretta, "Document Similarity Based on Concept Tree Distance," Proc. 19th ACM Conf. Hypertext and Hypermedia, pp. 127-132,2008.

[54] H. Chim and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229, Sept.2008.

[55] Ujjwal Maulik and Sanghamitra Bandyopadhyay, " Performance Evaluation of Some Clustering Algorithms and Validity Indices," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, December 2002.

[56] Tapana Mekaroonkamon and Sarwan Wongsa, "A Comparative Investigation of the Robustness of Unsupervised Clustering Techniques for Rotating Machine Fault Diagnosis with Poorly-Separated Data," 8th International Conference on Advanced Computational Intelligence Chiang Mai, Thailand; February 14-16, 2016.

[57] Dan A. Simovici, Member, IEEE, and Szymon Jaroszewicz, "An Axiomatization of Partition Entropy," IEEE Transactions on Information Theory, vol. 48, no. 7, July 2002.

[58] Kelemen, O., Tezel, O., Ozkul, E., Tiryaki, B. K., and Agayev, E, "A comparison of validity indices on fuzzy C-means clustering algorithm for directional data," 2017 25th Signal Processing and Communications Applications Conference (SIU).doi:10.1109/siu.2017.7960557.

[59] Kuo-Lung Wu, "An analysis of robustness of partition coefficient index. 2008 IEEE International Conference on Fuzzy Systems," IEEE World Congress on Computational Intelligence.doi:10.1109/fuzzy .2008.4630393.

[60] Vergani, A. A., & Binaghi, E, "A Soft Davies-Bouldin Separation Measure," IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). DOI:10.1109/fuzz-IEEE.2018.8491581.

[61] K. Rajendra Prasad, Moulana Mohammed, R M Noorullah,, " Visual Topic Models for Healthcare Data Clustering", Evolutionary Intelligence, Oct, 2019, pp:1-17, https://doi.org/10.1007/s12065-019-00300-y.