A Novel Cosine-based Internal and External Validation metrics to assess twitter Data Clustering using Hybrid Topic Models

Noorullah R M and Moulana Mohammed

1Research Scholar, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

1Assistant Professor, Dept. of CSE, Institute of Aeronautical Engineering, Hyderabad, Telangana, India

2Associate Professor, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

E-mail: moulana@kluniversity.in, noorullah.rm@iare.ac.in

Abstract. In document clustering labeled and unlabeled documents are organized into a desired number of coherent and meaningful sub-clusters. Topic models are useful in extracting cluster tendency from Twitter-based data document clusters. Evaluating cluster tendency and performance with a reliable metric is one of the unsolved problems in topic document clustering. In the previous study cluster validity metrics have been proposed under Euclidean distance measure, these metrics underperform in topic models when dealing with numerical databases and for large corpus datasets. In this paper, to assess Twitter data clustering a novel cosine based internal and external validity indices used by considering closeness between documents, the lexical similarity and cluster classification metrics for each topic separately using Hybrid topic models. Experimentally proved the effectiveness of cosine based internal and external validity metrics and results compared with Euclidean metrics using benchmark and Twitter-based datasets.

1. Introduction

In both textual and numerical data clustering, one challenging issue is to organize massive data into sub-clusters without or with prior knowledge. Topics of documents in Topic modeling found by using unsupervised classification [1] or by post clustering without prior information and supervised classification process or preclustering [2] in the presence of prior information. The final result consists of assigning a previously unknown or known category to each document relevant to the main topic. There sufficient literature in the field of topic modeling techniques [3-6], clustering methods and algorithms using hybrid topic models [7] in Twitter data document clustering. Evaluating cluster tendency and performance with a reliable metric is one of the unsolved problems in topic document clustering. Cluster validity measured with internal and external validity indices. The external validity indices [8-11] measure the correspondence between identified clusters and externally provided labels in social voting [28]. The Internal validity indices [12-17] evaluate the goodness of cluster structure with partitioned data by considering compactness and separation of obtained partitioned structure. Internal validity indices are preferred in performance measure because in most of the cases prior information on the number of clusters will not be available. In previous literature, a wide variety of internal and external validity indices have been provided which will be useful in finding the number of topics but not in choosing an appropriate measure and metric to validate cluster and also not by considering the elements in the cluster are well classified are not. Most commonly used measure is especially using Euclidean distance, which shows poor results in high dimensionality document clustering. In this paper, a novel cosine based internal and external validity metrics proposed for internally evaluating the results of a document clustering by considering into account the peculiarity of textual data [18], closeness between documents [19], considering the lexical similarity [20] and also considered cluster classification metrics in classification of elements in the cluster are well classified.
or not. Experimentally effectiveness of proposed clusters validity metrics is evaluated with benchmark and Twitter-based datasets. Rest of the paper is organized as follows: Section 2 presents the theoretical background of clustering in Topic Modelling; Section 3 describes the process description and cluster validation; in Section 4 preliminary experimental evaluation and performance of validation measures; Section 5 a novel cosine based validity metrics for validating document clustering, comparison with validity metrics under Euclidean and cluster classification metrics in cluster validation. Finally, in Section 6 the conclusion and future scope of the work is presented.

2. Theoretical Background Of Clustering In Topic Modeling
For the same dataset, using different algorithms gives different solutions by generating sub-clusters, different choice of input parameters produce different results for the same algorithm which affects the final result in finding the optimal number of topics or clusters in the given topic document. To assess cluster obtained by used algorithm, to decide which algorithm is most suitable for the specific application, and to provide reliability to results suitable evaluation criteria under suitable measure is still needed. In most of algorithms proximities, pairwise distances measured by using Euclidean distance metrics are considered which is suitable for the lower number of dimensionality, it loses its reliability and interpretability at an increase of dimensionality. Clustering algorithm deal with distance, and distance relates to similarity/dissimilarity. The complement to Euclidean metric is cosine based similarity metric in text classification problems which uses both magnitude and direction of vectors, which is non-negative, independent of document length and bounded between [0, 1]. One of the most interesting variations in K-means family is spherical k-means [21], [31] which is based on cosine based similarity used in information retrieval, in which the effect of different length of documents is reduced by normalization. Given two tweet documents $d_i$ and $d_j$ a corpus, then cosine based distance similarity is given as

$$\cos(d_i, d_j) = \frac{d_i^T d_j}{\|d_i\| \|d_j\|}$$

(1)

The cosine is 1 if the documents use the same words and 0 if they have no two terms in common. The effort of different length of documents is lessened by normalization.

3. Process Description And Cluster Validation

3.1 Datasets Description
For the experiment, the datasets collected from Twitter on 20 topics of health-related [24], [27], [32] documents, TREC2014, TREC2015 Keyword Phrases Tweets collected from Twitter are used as described in [7] and Tweets extracted from Twitter related to 25 keyword phrases of TREC2018 [22] as described in Table 1 are also used. Experiments are implemented with Intel core i7 processor @3.4 GHz, 8MB cache, 16GB RAM, 1TB HDD in IDLE (Python 3.8 64bit) environment on these four different datasets and results discussed in ensuing sections.

| S.No. | Datasets | Description of Keyword Phrases |
|-------|----------|-------------------------------|
| 1     | 2Keyword Phrases | Women in Parliaments, Black Bear Attacks |
| 2     | 3Keyword Phrases | Women in Parliaments, Black Bear Attacks, Airport Security |
| 3     | 4Keyword Phrases | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction |
| 4     | 5Keyword Phrases | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals |
| 5     | 6Keyword Phrases | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling |
| 6     | 7Keyword Phrases | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters |
| 7     | 8Keyword Phrases | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy |
| Page | Keyword Phrases                                                                                       |
|------|------------------------------------------------------------------------------------------------------|
| 8    | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy |
| 9    | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition |
| 10   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing |
| 11   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms |
| 12   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy |
| 13   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage |
| 14   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage, women driving in Saudi Arabia |
| 15   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage, women driving in Saudi Arabia, eating invasive species |
| 16   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage, women driving in Saudi Arabia, eating invasive species, protect Earth from asteroids |
| 17   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage, women driving in Saudi Arabia, eating invasive species, protect Earth from asteroids, diabetes and toxic chemicals |
| 18   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage, women driving in Saudi Arabia, eating invasive species, protect Earth from asteroids, diabetes and toxic chemicals, car hacking |
| 19   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage, women driving in Saudi Arabia, eating invasive species, protect Earth from asteroids, diabetes and toxic chemicals, car hacking, social media and teen suicide |
| 20   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage, women driving in Saudi Arabia, eating invasive species, protect Earth from asteroids, diabetes and toxic chemicals, car hacking, social media and teen suicide, federal minimum wage increase |
| 21   | Women in Parliaments, Black Bear Attacks, Airport Security, Wildlife Extinction, Health and Computer Terminals, human smuggling, transportation tunnel disasters, piracy, hydrogen energy, euro opposition, mercy killing, tropical storms, women clergy, college education advantage, women driving in Saudi Arabia, eating invasive species, protect Earth from asteroids, diabetes and toxic chemicals, car hacking, social media and teen suicide, federal minimum wage increase, protect Earth from asteroids, diabetes and toxic chemicals, car hacking, social media and teen suicide, federal minimum wage increase |
3.2 Process Description
On each collected corpus as mentioned above the following steps are implemented:
Step 1: For each Twitter-based datasets collected preprocessing is performed by using Python Gensim library to prepare text documents for Document Clustering and classification.
Step 2: Programs implemented in Python to apply hybrid Topic models [7] under Cosine based and Euclidean distance-based measures.
Step 3: Document clustering and classification performed.
Step 4: Assessment of document cluster with confusion matrices [23] and classification metrics by using novel cosine-based internal and external validity metrics.
Step 5: Results compared with Euclidean metrics with confusion matrices and classification metrics are done.

3.3 Performance of Cluster Validation
In Topic modelling selection of appropriate method for implementation and assessment of clustering quality in information retrieval [25], image processing applications [26], [33] is still open challenges. Since the number of topics or clusters is not known ahead, irrespective of the clustering model, the final results needed to be evaluated for cluster validation. To validate cluster, external validation indices and internal validation indices are used. For choosing an optimal clustering algorithm and to assess identified clusters corresponding to externally provided labels external validity indices are used. In addition to these internal validation indices evaluate cluster structure with partitioned data by considering compactness and separation of obtained partitioned structure. It measures intra-cluster homogeneity, inter-cluster separability or both. In the majority of the application, prior information of the number of clusters is not available in such scenarios internal validation indices are best suited for cluster validation. In this paper, both internal validity indices (C.A., NMI, Precision, Recall and F-Score) and internal validity indices (DB, SI, XI, PCI, PEI and SM) used for performance evaluation. In addition to these validity indices classification metrics are also used to check the elements in the cluster are well classified or not topic wise.

4. Preliminary Experimental Evaluation and Performance of Validation Measures
The experiment aims to compare the behaviour of cosine based internal and external validity indices with Euclidean based indices. To perform comparative study different benchmark and real-time
twitter-based datasets are collected. On each dataset, four hybrid topic models [7] algorithms are implemented under Euclidean and Cosine based measures. Results of the five external validity indices and six internal validity indices on every dataset have been calculated and tabulated and sample of compared results are shown in form of tables and graphs.

| Tweets Dataset | CLUSTERING ACCURACY (CA) |
|----------------|----------------------------|
|                | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases      | 1.000 | 0.850 | 0.575 | 0.500 | 1.000 | 0.800 | 0.675 | 0.500 |
| 3KPhrases      | 1.000 | 0.500 | 0.467 | 0.375 | 1.000 | 0.625 | 0.542 | 0.442 |
| 4KPhrases      | 0.888 | 0.644 | 0.494 | 0.356 | 0.931 | 0.625 | 0.481 | 0.394 |
| 5KPhrases      | 0.615 | 0.495 | 0.360 | 0.310 | 1.000 | 0.465 | 0.620 | 0.335 |
| 6KPhrases      | 0.521 | 0.408 | 0.329 | 0.338 | 0.767 | 0.454 | 0.383 | 0.342 |
| 7KPhrases      | 0.445 | 0.407 | 0.332 | 0.300 | 0.861 | 0.321 | 0.407 | 0.268 |
| 8KPhrases      | 0.644 | 0.644 | 0.644 | 0.644 | 0.813 | 0.316 | 0.397 | 0.288 |
| 9KPhrases      | 0.497 | 0.406 | 0.286 | 0.275 | 0.767 | 0.317 | 0.369 | 0.289 |
| 10KPhrases     | 0.538 | 0.353 | 0.273 | 0.223 | 0.593 | 0.280 | 0.383 | 0.288 |
| 11KPhrases     | 0.450 | 0.266 | 0.309 | 0.198 | 0.714 | 0.268 | 0.323 | 0.239 |
| 12KPhrases     | 0.456 | 0.350 | 0.319 | 0.210 | 0.679 | 0.329 | 0.425 | 0.231 |
| 13KPhrases     | 0.423 | 0.221 | 0.252 | 0.250 | 0.508 | 0.288 | 0.346 | 0.202 |
| 14KPhrases     | 0.373 | 0.261 | 0.239 | 0.220 | 0.645 | 0.252 | 0.377 | 0.213 |
| 15KPhrases     | 0.293 | 0.207 | 0.263 | 0.200 | 0.331 | 0.175 | 0.226 | 0.148 |
| 16KPhrases     | 0.411 | 0.253 | 0.263 | 0.223 | 0.570 | 0.295 | 0.377 | 0.220 |
| 17KPhrases     | 0.378 | 0.210 | 0.222 | 0.213 | 0.550 | 0.288 | 0.301 | 0.244 |
| 18KPhrases     | 0.310 | 0.265 | 0.275 | 0.193 | 0.515 | 0.258 | 0.403 | 0.206 |
| 19KPhrases     | 0.359 | 0.222 | 0.322 | 0.197 | 0.570 | 0.299 | 0.382 | 0.245 |
| 20KPhrases     | 0.343 | 0.235 | 0.213 | 0.210 | 0.524 | 0.275 | 0.421 | 0.205 |
| 21KPhrases     | 0.540 | 0.150 | 0.610 | 0.145 | 0.542 | 0.139 | 0.298 | 0.137 |
| 22KPhrases     | 0.472 | 0.148 | 0.501 | 0.150 | 0.482 | 0.135 | 0.310 | 0.147 |
| 23KPhrases     | 0.477 | 0.160 | 0.503 | 0.141 | 0.480 | 0.145 | 0.283 | 0.145 |
| 24KPhrases     | 0.477 | 0.148 | 0.485 | 0.142 | 0.478 | 0.143 | 0.316 | 0.142 |
| 25KPhrases     | 0.573 | 0.153 | 0.468 | 0.143 | 0.574 | 0.134 | 0.302 | 0.146 |

VN: Visual NMF VL: Visual LDA VLS: Visual LSI VPL: Visual PLSA

Results of all datasets of external and internal validity indices under cosine and Euclidean are tabulated for four hybrid topic models. Some sample results are presented in tabular and graphical forms. In Table 2 External validity index (Cluster Accuracy) of 2 keyword phrases to 25 keyword phrases of TREC2018 datasets and in Table 3, all external and internal validity indices of TREC2014 dataset are shown. From these results interpreted that cosine based external and internal validity indices perform better than that of Euclidean in majority of keyword phrases. Particularly performs well when smaller keyword phrases, as keyword phrases size increases result values are decreasing under both metrics, but still consistency is maintained in case of cosine based metrics. Higher values of results are represented in bold format.

4.1 Document clusters validation by using Cosine based Measures

For evaluating compactness and separation of formed clusters usually Euclidean measure deployed in previous studies and for external validity indices in most of the cases. Using this measure may be inconsistent with the criterion for getting partition for a specific algorithm. With this motivation, in this paper novel cosine based derived metrics are used in document clustering algorithms using hybrid topic models, hybrid framework [29], [30] and also in validating formed clusters by using these metrics. In addition to that clusters to have high cohesion and well distinguished, both compactness and separation are considered.

Consider corpus $X = \{d_1, d_2, \ldots, d_n\} \subseteq \mathbb{K}^p$ consists of n document vectors in 'p' terms space of dimension. With the help of hybrid clustering algorithm k number of clusters Cq (where q=1, 2, ...,k) have been identified, such that each document has one of the labels identifying the k different
clusters. These clustering algorithms aim to maximize intra-cluster proximities and minimize inter-cluster proximities. Let $d_i$, $d_i'$ and $d_j$ be three documents in a corpus $X$, with $d_i$ and $d_i'$ belongs to the same cluster and $d_j$ belongs to other clusters. Compactness and separation can be calculated as follows:

Compactness
\[
(C_y) = \sum_{d_i,d_j \in C_y} \text{proximities}(d_i,d_j)
\]

Separation
\[
(C_y,C_q) = \sum_{d_i \in C_y, d_j \in C_q} \text{proximities}(d_i,d_j)
\]

(2)

(3)

Where proximities ($\cdot$) usually denote the Euclidean distance.

In this paper, external validity indices include clustering accuracy (CA), Normalized Mutual Information (NMI), Precision (P), Recall (R), and F-Score (F) under cosine-based metrics and derived internal validity indices with cosine similarity as mentioned below. Davis-Bouldin Index (DB), Silhouette Index (XI), Partition Entropy Index (PEI) and Separation Measure (SM) are considered for evaluating. In internal validity indices, Davis-Bouldin index (DB) depends on both data and algorithm given as

\[
DB = \frac{1}{N} \sum_{i=1}^{k} D_i
\]

\[
D_i = \max_{a \neq b} R_{ab} - \frac{S_a + S_b}{M_{ab}}
\]

(4)

Eq. (4) can be rewritten with cosine dissimilarity as

\[
DB_{\text{cosine}} = \frac{1}{N} \sum_{i=1}^{k} (1 - \cos(D_i))
\]

(5)

Silhouette index (S.I) is given as

\[
S(i) = \begin{cases} 
1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\
0 & \text{if } a(i) = b(i) \\
\frac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i)
\end{cases}
\]

(6)

By considering cosine similarity, Eq. (6) can be written as

\[
S(i)_{\text{cosine}} = \begin{cases} 
1 - \cos\left(\frac{a(i)}{b(i)}\right) & \text{if } a(i) < b(i) \\
0 & \text{if } a(i) = b(i) \\
\cos\left(\frac{b(i)}{a(i)} - 1\right) & \text{if } a(i) > b(i)
\end{cases}
\]

(7)

By using these equations, calculated values of validity indices are tabulated. Higher values are represented in bold form in the tables mentioned below. From these tabulated values, comparison between cosine and Euclidean representations is given in graphical form in following sections.

| TREC2014 | Cosine Based | Euclidean Based |
|----------|--------------|-----------------|
| C.A.     | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases| 1.000 | 0.975 | 1.000 | 0.750 | 1.000 | 0.975 | 0.975 | 0.700 |
| 3KPhrases| 1.000 | 0.908 | 1.000 | 0.483 | 0.983 | 0.891 | 0.983 | 0.483 |
| 4KPhrases| 1.000 | 0.725 | 1.000 | 0.450 | 0.850 | 0.825 | 0.968 | 0.443 |

| N.M.I.   | Cosine Based | Euclidean Based |
|----------|--------------|-----------------|
| VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases| 1.000 | 0.831 | 1.000 | 0.188 | 1.000 | 0.831 | 0.831 | 0.118 |
| 3KPhrases| 1.000 | 0.716 | 1.000 | 0.090 | 0.929 | 0.687 | 0.929 | 0.076 |
| 4KPhrases| 1.000 | 0.439 | 1.000 | 0.153 | 0.636 | 0.583 | 0.901 | 0.161 |

| Precision (P) | Cosine Based | Euclidean Based |
|---------------|--------------|-----------------|
| VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases| 1.000 | 1.000 | 1.000 | 0.794 | 1.000 | 1.000 | 1.000 | 0.814 |
| 3KPhrases| 1.000 | 1.000 | 1.000 | 0.460 | 1.000 | 1.000 | 1.000 | 0.460 |

Table 3 TREC2014 Dataset External and Internal Validity Indices
| 4KPhrases | 0.993 | 0.993 | **1.000** | 0.441 | 0.670 | 0.670 | 0.968 | 0.486 |
| Recall(R) | Cosine Based | Euclidean Based | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases | **1.000** | 1.000 | 0.975 | 0.675 | **1.000** | 1.000 | 0.875 | 0.550 |
| 3KPhrases | **1.000** | 1.000 | **1.000** | 0.458 | **1.000** | 1.000 | 0.983 | 0.458 |
| 4KPhrases | 0.993 | 0.993 | **1.000** | 0.443 | 0.706 | 0.706 | 0.968 | 0.500 |
| F-Score(F) | Cosine Based | Euclidean Based | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases | **1.000** | 1.000 | 0.987 | 0.729 | **1.000** | 1.000 | 0.933 | 0.656 |
| 3KPhrases | **1.000** | 1.000 | **1.000** | 0.458 | **1.000** | 1.000 | 0.983 | 0.458 |
| 4KPhrases | 0.993 | 0.993 | **1.000** | 0.440 | 0.656 | 0.656 | 0.968 | 0.489 |
| D.B. | Cosine Based | Euclidean Based | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases | **0.690** | 0.765 | 0.690 | 1.229 | 0.929 | 2.874 | 0.932 | 2.025 |
| 3KPhrases | 1.306 | 1.567 | **1.316** | 4.108 | 1.845 | 2.110 | 1.878 | 5.976 |
| 4KPhrases | **1.855** | 3.876 | 1.875 | 6.184 | 3.570 | 3.903 | 2.848 | 5.465 |
| S.I. | Cosine Based | Euclidean Based | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases | **0.998** | 0.800 | 0.869 | 0.654 | 0.894 | 0.859 | 0.090 | 0.432 |
| 3KPhrases | **0.983** | 0.557 | 0.165 | 0.145 | 0.764 | 0.470 | 0.153 | 0.524 |
| 4KPhrases | **0.962** | 0.103 | 0.065 | 0.042 | 0.163 | 0.252 | -0.04 | 0.243 |
| X.I. | Cosine Based | Euclidean Based | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases | **0.038** | 1.807 | 0.065 | 1.616 | 1.970 | 1.235 | 3.60 | 1.638 |
| 3KPhrases | 14.38 | 20.47 | 25.94 | 17.63 | 94.03 | 30.15 | 40.01 | 29.96 |
| 4KPhrases | 0.547 | 0.151 | 1.366 | 4.651 | 481.16 | 33.10 | 155.6 | 210.9 |
| P.C.I. | Cosine Based | Euclidean Based | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases | **0.998** | 0.929 | 0.998 | 0.938 | 0.922 | 0.947 | 0.944 | 0.927 |
| 3KPhrases | **0.978** | 0.862 | 0.968 | 0.968 | 0.851 | 0.830 | 0.847 | 0.958 |
| 4KPhrases | **0.953** | 0.712 | 0.934 | 0.905 | 0.738 | 0.684 | 0.770 | 0.872 |
| P.E.I. | Cosine Based | Euclidean Based | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases | **0.003** | 0.130 | 0.006 | 0.099 | 0.140 | 0.101 | 0.106 | 0.116 |
| 3KPhrases | 0.048 | 0.277 | 0.073 | **0.064** | 0.286 | 0.337 | 0.294 | 0.082 |
| 4KPhrases | **0.107** | 0.585 | 0.154 | 0.199 | 0.534 | 0.641 | 0.475 | 0.256 |
| S.M. | Cosine Based | Euclidean Based | VN | VL | VLS | VPL | VN | VL | VLS | VPL |
| 2KPhrases | **0.025** | 0.038 | 0.0269 | 0.032 | 0.084 | 0.039 | 0.089 | 0.036 |
| 3KPhrases | **0.022** | 0.046 | 0.0259 | 0.025 | 0.093 | 0.056 | 0.095 | 0.026 |
| 4KPhrases | **0.022** | 0.054 | 0.0269 | 0.026 | 0.556 | 0.076 | 0.153 | 0.028 |

VN: Visual NMF  VL: Visual LDA  VLS: Visual LSI  VPL: Visual PLSA

4.1.1 Graphical representation of experimental results of External and Internal Validity indices under cosine
Figure 1. External validity indices of 2Topics to 20Topics Twitter Dataset under Cosine Metric

External validity indices (CA, NMI, Precision, Recall and F-Score) under cosine of 2topics to 20 topics health datasets are represented as spiral graphs shown in Figure 1(a) to 1(e). All external validity indices values lies [0, 1]. Any external validity index value near to value 1 performs better clustering. From Figure 1(a) Accuracy index results for 2topics to 20 topics are shown, from this spiral graph interpreted that Visual NMF and Visual LSI algorithm perform well. At 7T, 8T, 11T and 12T visual NMF perform better than the other three methods. By observing NMI external index results shown in Figure 1(b) for most of the topics Visual LSI method perform well, whereas, for 7T, 8T, 11T and 13T Visual NMF performance is better than other methods. In Figure 1(c) precision values are shown, from this inferred Visual NMF performs well in most of the topics except 3T to 6T, and 10T Visual LSI performs well. Recall values are shown in Figure 1(d) from which conclusion drawn except for 3T to 6T, for rest of topics Visual NMF performance is good. In those topics, Visual LSI performs well. In Figure 1(e) F-Score values are represented from these results inferred that both Visual NMF and Visual LSI perform well. On overall performance, both Visual NMF and Visual LSI perform well when compared to the other two methods for all five external indices values under cosine based metric.
Figure 2 TREC2018 Dataset Internal Validity Indices under Cosine Metric

Figure 2(a) shows the performance of Davis-Bouldin (DB) internal index values under the cosine metric of TREC2018 keyword phrases. Its values range from 0 to 40 shown on the Y-axis. In the case of this index, the minimum value will perform better clustering results. Form this graph on observation, for most of the keyword phrases visual LSI performs better than other methods. In case of 7keywords, 8keywords, 13keywords, 16keywords, 19keywords and 20keywords Visual NMF performed better than other methods.

Silhouette index (SI) values range from \(-1\) to \(+1\). If this index value is nearer to \(+1\) then cluster performance will be best. If values decrease from \(+1\) to \(-1\) its performance also decreases. From the bar graph shown in Figure 2(b) results can interpret Visual NMF under cosine performs well in all TREC2018 keyword phrases. Partition coefficient index (PCI) values lie between 0 and 1. Values nearer to 1 will be treated as best. From Figure 2(d), based on PCI values under cosine metric, in case of 7keywords, and 10 keywords Visual LSI perform better, for 11keyword phrases Visual NMF perform well and in rest of all keyword phrases Visual PLSI methods perform well. Figure 2(e) shows performance values of partition coefficient internal index values which range from 0 to \(\log(c)\). In this case its value range from 0 to 3 as indicated on the Y-axis line graph. The minimum value will be considered for higher performance in clustering. From this graph, Visual LSI method performs well for 7 to 10 keyword phrases, for 11 and 12 keyword phrases Visual NMF and for rest of keyword phrases Visual PLSI perform better than other methods. Separation Measure internal index value is smaller then it will have greater performance. In this case, its value ranges from 0 to 10 as represented on the Y-axis. From this line graph as shown in Fig. 2(f), 7keyword phrases, 8keywords, 11keywords and 13keywords Visual LSI perform well and in rest of keyword phrases, Visual PLSI under cosine metric performs better than other methods.

4.2 Comparative study of cosine based validation with Euclidean distance-based cluster validation

Figure 3 External validity indices Comparative results of 2Topics to 20Topics Twitter Dataset
External validity indices (CA, NMI, Precision, Recall and F-Score) comparative results of 2 topics to 20 topics health datasets are represented in the form of spiral graphs as shown in Figure 3(a) to 3(e). All external validity indices values lies [0, 1]. Any external validity index value near to value 1 performs better clustering. From Figure 3(a) Accuracy index results for 2 topics to 20 topics are shown, from this spiral graph Visual LSI algorithm under cosine based metric perform well. At 7T, 8T, 11T and 12T visual NMF under cosine performs better than the other three methods. By observing NMI external index results in Figure 3(b) for most of the topics Visual LSI under cosine metric performs well, whereas, for 7T, 8T, 11T and 12T Visual NMF under cosine performance is better than other methods. In Figure 3(c) precision values are shown, from this results inferred Visual LSI under cosine metric performs well in most of the topics except 7T, 8T, 12T, and 13T Visual NMF under cosine perform, whereas at 14 Visual NMF under Euclidean perform well when compared to all other methods. Recall values are shown in Figure 3(d) from which conclusion drawn that both Visual NMF and Visual LSI under cosine metric perform equally. In Figure 3(e) F-Score values are represented from this can inferred that both Visual NMF and Visual LSI under cosine metric perform well. On overall performance, both Visual NMF and Visual LSI under cosine metric perform well when compared Euclidean for all five external indices value.

Figure 4(a) shows performance comparative results of Davis-Bouldin (DB) internal index values under Cosine and Euclidean metrics of TREC2018 keyword phrases. Its values range from 0 to 70 shown on the Y-axis. In the case of this index, the minimum value will be considered for better clustering results. Form this graph on observation, 2keywords to 6keywords, 9keywords to 12keywords, 14keywords and 17keywords visual LSI under cosine perform well, for 18keyword phrases visual NMF perform best and rest of keyword phrases Visual NMF under Cosine performs better than other models. Silhouette index (SI) values range from -1 to +1. If this index value is nearer to +1 then cluster performance will be best. If values decrease from +1 to -1 its performance also decreases. From the line graph as shown in Figure 4(b) interpreted that Visual NMF under cosine performs well in all TREC2018 keyword phrases, except 5keyword phrases where Visual NMF under Euclidean perform well.
In Figure 4(c) Xie-Beni index (XI) internal validity index value under cosine and Euclidean metric are represented. Its values range from 0 to 110 as represented on the Y-axis. The minimum value of this index will be considered as the best performance. From this line graph, in case of 2 keyword phrases to 5 keyword phrases visual LDA under Euclidean performs well, and rest of keyword phrases of TREC2018 datasets visual PLSI under cosine metric performs better than other methods and also better than Euclidean distance metric.

Partition coefficient index (PCI) values lie between 0 and 1. Maximum values will be considered as better performance, values nearer to 1 will be treated as best. From Figure 4(d), based on PCI comparative result values under cosine and Euclidean metrics, interpret that 2 keyword phrases to 6 keyword phrases visual PLSI under Cosine metric performance are good, for 8 keyword phrases, 10 keyword phrases, 14 keyword phrases, and 17 keyword phrases visual NMF under Euclidean metric and for rest of keyword phrases visual LSI under cosine metric perform well.

Figure 4(e) shows performance comparative values of partition coefficient internal index values which range from 0 to \( \log c \). In this case, its value range from 0 to 3 as indicated on the Y-axis line graph. The minimum value will be considered for higher performance in clustering. From this graph, infer that for 2 to 4 keyword phrases, 6 keyword phrases visual PLSI under Cosine, for 5 keyword phrases, and 7 keyword phrases visual LSI under cosine, and for 8 keyword phrases, 10 keyword phrases, 14 keyword phrases, and 17 keyword phrases visual NMF under Euclidean and rest of keyword phrases visual LSI under Euclidean perform better.
Separation Measure internal index value is smaller then it will have greater performance. In this case, its value ranges from 0 to 10 as represented on the Y-axis. From this line graph as shown in Figure 4(f), 2 keyword phrases to 5 keyword phrases visual LDA under Euclidean perform better and for rest of keyword phrases, visual PLSI under cosine metric perform better than other models and also compared to Euclidean distance.

In Figure 5(a) to 5(d) external validity indices (Accuracy, NMI, Recall and F-Score) comparative results are shown. All external validity index value lies between 0 and 1. If values are nearer to 1, it indicates good clustering done, and appropriated keywords are placed in the appropriate cluster. From these bar graphs, interpret in all external validity indices visual NMF, visual LSI and visual LDA under cosine metrics perform well and their values are near to 1.
Figure 6(a) shows comparative performance values of Davis-Bouldin (DB) internal index values under cosine and Euclidean metrics of TREC2015 keyword phrases. Its values range from 0 to 15 shown on the Y-axis. In the case of this index, the minimum value will perform better clustering results. Form this graph on observation, inferred that visual NMF under cosine metric performs well when compared to the Euclidean metric for all models. Silhouette internal index (SI) values range from -1 to +1. If this index value is nearer to +1 then cluster performance will be best. If values decrease from +1 to -1 its performance also decreases. From the line graph as shown in Figure 6(b) Visual NMF under cosine metric performs well in all TREC2015 keyword phrases than that of Euclidean distance metric. In Figure 6(c) Xie-Beni index (XI) internal validity index value under cosine and Euclidean metric are represented. Its values range from 0 to 300 as represented on the Y-axis. The minimum value of this index will be considered as the best performance. From this line graph results interpreted for 2keyword phrases and 3keyword phrases visual NMF performs well, whereas for 4keyword phrases and 5keyword phrases of TREC2015 visual LDA performs well. In all cases performs is better under cosine based validity index than Euclidean metric based. Partition coefficient index (PCI) values lie between 0 and 1. Maximum values will be considered as better performance, values nearer to 1 will be treated as best. From Figure 6(d), based on PCI comparative result values under Cosine and Euclidean metrics, Visual NMF and Visual LSI both methods values are greater than that of other values under Cosine metric based validity indices. Figure 6(e) shows performance comparative values of partition coefficient internal index values which range from 0 to log\(_c\). In this case, its value range from 0 to 1.2 as the number of keywords considered are only four which is indicated on the Y-axis line graph. The minimum value will be considered for higher performance in clustering. From this graph, visual NMF values under cosine metric are greater than that of Euclidean distance-based metrics. Separation Measure internal index value is smaller then it will have greater performance. In this case, its value ranges from 0 to 1 as represented on the Y-axis. From this line graph as shown in Figure 6(f), for 2keyword phrases and 3keyword phrases visual NMF, in case of 4keyword phrases, and 5 keyword phrases visual LDA. Both methods have better values under Cosine based metric validity index values than that of Euclidean based metric validity index values.

Figure 6 TREC2015 Internal Validity Indices comparative Results
4.3 Cluster Classification Metrics to check elements in cluster

4.3.1 External validity indices under Cosine metric based on Cluster Classification Metrics

| Cl # | P      | R      | F      | SU | Cl # | P      | R      | F      | SU  |
|------|--------|--------|--------|-----|------|--------|--------|--------|-----|
| 1    | 0.550  | 0.550  | 0.550  | 40  | 1    | 0.250  | 0.250  | 0.250  | 40  |
| 2    | 0.350  | 0.350  | 0.350  | 40  | 2    | 0.300  | 0.300  | 0.300  | 40  |
| 3    | 0.625  | 0.625  | 0.625  | 40  | 3    | 0.325  | 0.325  | 0.325  | 40  |
| 4    | 0.525  | 0.525  | 0.525  | 40  | 4    | 0.275  | 0.275  | 0.275  | 40  |
| 5    | 0.675  | 0.675  | 0.675  | 40  | 5    | 0.225  | 0.225  | 0.225  | 40  |
| 6    | 0.700  | 0.700  | 0.700  | 40  | 6    | 0.200  | 0.200  | 0.200  | 40  |
| 7    | 0.200  | 0.200  | 0.200  | 40  | 7    | 0.225  | 0.225  | 0.225  | 40  |

Accuracy: Weighted Average (W.A.), Macro Average (M.A.), and Weighted Average

In section 4.2 and 4.3 cluster validity indices are calculated based on confusion matrix and the number of clusters. In previous studies also cluster validation done using confusion matrices but not considered the elements in the cluster are well classified or not. In this paper, cluster validation done by considering both confusion matrices and classification metrics to see that elements in cluster are well classified or not. Cluster classification metrics are tabulated for all datasets for all four models under Cosine based and Euclidean metrics. Some sample results are represented from table 4 to table 9 for different datasets. In table 4, external validity indices Precision (P), Recall (R), F-Score (F), Accuracy, Macro Average (M.A.) and Weighted Average (W.A.) of 7 Topics Twitter datasets based on Cluster Classification under Cosine metric are presented. Here, seven topics treated as seven clusters and external validity indices results for each cluster are represented by considering every document in that particular cluster where Support (SU) represent number of documents present in that cluster.

**Table 5 External validity indices based on Cluster Classification Metrics of 10 Topics Twitter Datasets**

| Cl # | P      | R      | F      | SU | Cl # | P      | R      | F      | SU  |
|------|--------|--------|--------|-----|------|--------|--------|--------|-----|
| 1    | 0.800  | 0.800  | 0.800  | 50  | 1    | 0.240  | 0.240  | 0.240  | 50  |
| 2    | 0.400  | 0.600  | 0.480  | 20  | 2    | 0.200  | 0.150  | 0.171  | 20  |
| 3    | 0.714  | 0.385  | 0.500  | 65  | 3    | 0.246  | 0.246  | 0.246  | 65  |
| 4    | 0.457  | 0.457  | 0.457  | 35  | 4    | 0.157  | 0.314  | 0.210  | 35  |
| 5    | 0.822  | 0.529  | 0.643  | 70  | 5    | 0.257  | 0.129  | 0.171  | 70  |
| 6    | 0.550  | 0.244  | 0.338  | 45  | 6    | 0.229  | 0.178  | 0.200  | 45  |
| 7    | 0.323  | 0.700  | 0.442  | 30  | 7    | 0.133  | 0.133  | 0.133  | 30  |
| 8    | 0.543  | 0.543  | 0.543  | 35  | 8    | 0.111  | 0.143  | 0.125  | 35  |
| 9    | 0.247  | 0.514  | 0.343  | 35  | 9    | 0.229  | 0.229  | 0.229  | 35  |
| 10   | 0.000  | 0.000  | 0.000  | 15  | 10   | 0.100  | 0.133  | 0.114  | 15  |

Accuracy: Weighted Average (W.A.), Macro Average (M.A.), and Weighted Average

In table 5, external validity indices of 10 Topics Twitter datasets for Visual NMF and Visual LDA hybrid topic models based on Cluster Classification metrics under Cosine metric are tabulated. Here, ten topics treated as ten clusters and external validity indices results for each cluster are represented by considering every document in that particular cluster where Support (SU) represent number of documents present in that cluster.
4.3.2 Comparative Results of External validity indices based on Cluster Classification

In this paper, comparative study of external validity indices based on cluster classifications metrics also performed for different hybrid topic models under Cosine based and Euclidean based metrics. Experimental results are tabulated for all types of datasets mentioned in the datasets description section. Sample of comparative results of external validity indices based on cluster classification for 20 keyword phrases of TREC2018 datasets are mentioned in table 6 to table 9.

Table 6 Comparative results of validity indices (Visual NMF) based on Cluster Classification for 20 Keyword Phrases of TREC2018 Datasets

| Visual NMF under Cosine metric | Cl # | P   | R   | F   | SU | Visual NMF under Eucl metric | Cl # | P   | R   | F   | SU |
|-------------------------------|------|-----|-----|-----|----|-------------------------------|------|-----|-----|-----|----|
| 1                             | 0.740| 0.740| 0.740| 50  | 1  | 0.740                        | 0.740| 0.740| 50 |
| 2                             | 0.300| 0.375| 0.333| 40  | 2  | 0.300                        | 0.375| 0.333| 40 |
| 3                             | 0.306| 0.578| 0.400| 45  | 3  | 0.306                        | 0.578| 0.400| 45 |
| 4                             | 0.200| 0.200| 0.200| 35  | 4  | 0.200                        | 0.200| 0.200| 35 |
| 5                             | 0.440| 0.367| 0.400| 30  | 5  | 0.440                        | 0.367| 0.400| 30 |
| 6                             | 0.543| 0.224| 0.317| 85  | 6  | 0.543                        | 0.224| 0.317| 85 |
| 7                             | 0.733| 0.440| 0.550| 50  | 7  | 0.733                        | 0.440| 0.550| 50 |
| 8                             | 0.880| 0.400| 0.550| 55  | 8  | 0.880                        | 0.400| 0.550| 55 |
| 9                             | 0.289| 0.371| 0.325| 35  | 9  | 0.289                        | 0.371| 0.325| 35 |
| 10                            | 0.000| 0.000| 0.000| 15  | 10 | 0.000                        | 0.000| 0.000| 15 |
| 11                            | 0.000| 0.000| 0.000| 30  | 11 | 0.000                        | 0.000| 0.000| 30 |
| 12                            | 0.325| 0.650| 0.433| 20  | 12 | 0.325                        | 0.650| 0.433| 20 |
| 13                            | 0.129| 0.360| 0.189| 25  | 13 | 0.129                        | 0.360| 0.189| 25 |
| 14                            | 0.600| 0.514| 0.554| 35  | 14 | 0.600                        | 0.514| 0.554| 35 |
| 15                            | 0.400| 0.286| 0.333| 70  | 15 | 0.400                        | 0.286| 0.333| 70 |
| 16                            | 0.771| 0.600| 0.675| 45  | 16 | 0.771                        | 0.600| 0.675| 45 |
| 17                            | 0.436| 0.480| 0.453| 50  | 17 | 0.436                        | 0.480| 0.453| 50 |
| 18                            | 0.086| 0.086| 0.086| 35  | 18 | 0.086                        | 0.086| 0.086| 35 |
| 19                            | 0.511| 0.920| 0.657| 25  | 19 | 0.511                        | 0.920| 0.657| 25 |
| 20                            | 0.067| 0.040| 0.050| 20  | 20 | 0.067                        | 0.040| 0.050| 25 |

Accuracy: 0.388 800
M.A: 0.388 382 800
W.A: 0.446 388 392 800

CL#: Cluster Number; P: Precision; R: Recall; F: F-Score; SU: Support; M.A.: Macro Average; W.A.: Weighted Average

Comparative results of validity indices of 20 keyword phrases of TREC2018 for Visual NMF hybrid topic model are mentioned in Table 6. From these results inferred that both results are the same under two distance metrics for all twenty clusters.

Table 7 Comparative results of validity indices (Visual LDA) based on Cluster Classification for 20 Keyword Phrases of TREC2018 Datasets

| Visual LDA under Cosine metric | Cl # | P   | R   | F   | SU | Visual LDA under Eucl metric | Cl # | P   | R   | F   | SU |
|-------------------------------|------|-----|-----|-----|----|-------------------------------|------|-----|-----|-----|----|
| 1                             | 0.171| 0.120| 0.141| 50  | 1  | 0.160                        | 0.160| 0.160| 50 |
| 2                             | 0.109| 0.150| 0.126| 40  | 2  | 0.167                        | 0.125| 0.143| 40 |
| 3                             | 0.100| 0.111| 0.105| 45  | 3  | 0.171                        | 0.133| 0.150| 45 |
| 4                             | 0.160| 0.144| 0.133| 35  | 4  | 0.100                        | 0.200| 0.133| 35 |
| 5                             | 0.086| 0.100| 0.092| 30  | 5  | 0.150                        | 0.100| 0.120| 30 |
| 6                             | 0.167| 0.100| 0.125| 85  | 6  | 0.212                        | 0.212| 0.212| 85 |
| 7                             | 0.167| 0.100| 0.125| 50  | 7  | 0.160                        | 0.080| 0.107| 50 |
| 8                             | 0.171| 0.218| 0.192| 55  | 8  | 0.171                        | 0.109| 0.133| 55 |
| 9                             | 0.100| 0.086| 0.092| 35  | 9  | 0.156                        | 0.200| 0.175| 35 |
| 10                            | 0.120| 0.200| 0.150| 15  | 10 | 0.080                        | 0.133| 0.100| 15 |
| 11                            | 0.100| 0.067| 0.080| 30  | 11 | 0.133                        | 0.067| 0.089| 30 |
| 12                            | 0.067| 0.050| 0.057| 20  | 12 | 0.080                        | 0.200| 0.114| 20 |
| 13                            | 0.114| 0.160| 0.133| 25  | 13 | 0.109                        | 0.240| 0.150| 25 |
In Table 7, comparative results of validity indices are shown under cosine and Euclidean based metrics for Visual LDA model. From these results interpreted Euclidean based metric on average for all clusters perform better than cosine based metric.

Table 8 Comparative results of validity indices (Visual LSI) based on Cluster Classification for 20 Keyword Phrases of TREC2018 Datasets

| Cl # | P   | R   | F   | SU | Cl # | P   | R   | F   | SU |
|------|-----|-----|-----|----|------|-----|-----|-----|----|
| 1    | 0.200 | 0.140 | 0.165 | 50 | 1    | 0.840 | 0.420 | 0.560 | 50 |
| 2    | 0.200 | 0.150 | 0.171 | 40 | 2    | 0.340 | 0.425 | 0.378 | 40 |
| 3    | 0.380 | 0.422 | 0.400 | 45 | 3    | 0.375 | 0.333 | 0.333 | 45 |
| 4    | 0.200 | 0.143 | 0.167 | 35 | 4    | 0.600 | 0.429 | 0.500 | 35 |
| 5    | 0.300 | 0.300 | 0.300 | 30 | 5    | 0.100 | 0.067 | 0.080 | 30 |
| 6    | 0.660 | 0.388 | 0.489 | 85 | 6    | 0.486 | 0.200 | 0.283 | 85 |
| 7    | 0.240 | 0.120 | 0.160 | 50 | 7    | 0.133 | 0.080 | 0.100 | 50 |
| 8    | 0.051 | 0.327 | 0.400 | 55 | 8    | 0.422 | 0.345 | 0.080 | 55 |
| 9    | 0.200 | 0.229 | 0.213 | 35 | 9    | 0.540 | 0.771 | 0.635 | 35 |
| 10   | 0.222 | 0.667 | 0.333 | 15 | 10   | 0.000 | 0.000 | 0.000 | 15 |
| 11   | 0.000 | 0.000 | 0.000 | 30 | 11   | 0.171 | 0.200 | 0.185 | 30 |
| 12   | 0.200 | 0.350 | 0.255 | 20 | 12   | 0.267 | 0.600 | 0.369 | 20 |
| 13   | 0.086 | 0.240 | 0.126 | 25 | 13   | 0.133 | 0.080 | 0.100 | 25 |
| 14   | 0.080 | 0.057 | 0.067 | 35 | 14   | 0.388 | 0.943 | 0.550 | 35 |
| 15   | 0.235 | 0.286 | 0.258 | 70 | 15   | 0.329 | 0.329 | 0.329 | 70 |
| 16   | 0.400 | 0.489 | 0.440 | 45 | 16   | 0.333 | 0.222 | 0.267 | 45 |
| 17   | 0.089 | 0.080 | 0.084 | 50 | 17   | 0.686 | 0.480 | 0.565 | 50 |
| 18   | 0.014 | 0.114 | 0.114 | 35 | 18   | 0.540 | 0.771 | 0.635 | 35 |
| 19   | 0.150 | 0.120 | 0.133 | 25 | 19   | 0.171 | 0.240 | 0.200 | 25 |
| 20   | 0.200 | 0.400 | 0.267 | 25 | 20   | 0.364 | 0.800 | 0.500 | 25 |

Accuracy 0.249 800 Accuracy 0.375 800
M.A. 0.234 0.251 0.227 800 M.A. 0.361 0.387 0.348 800
W.A. 0.273 0.249 0.248 800 W.A. 0.398 0.375 0.361 800

Comparative results of external validity indices of 20 keyword phrases of TREC2018 for Visual LSI hybrid topic model are mentioned in Table 8. From these results, Euclidean results are better than that of cosine based on accuracy, Macro Average and Weighted Average.
Table 9: Comparative results of validity indices (Visual PLSI) based on Cluster Classification for 20 Keyword Phrases of TREC2018 Datasets

| CL # | P  | R  | F  | SU | Visual PLSI under Euclidean metric |
|------|----|----|----|----|---------------------------------|
| 1    | 0.200 | 0.180 | 0.189 | 50 | 1.000 | 0.200 | 0.200 | 0.200 | 50 |
| 2    | 0.140 | 0.175 | 0.156 | 40 | 1.000 | 0.080 | 0.050 | 0.062 | 40 |
| 3    | 0.111 | 0.111 | 0.111 | 45 | 0.267 | 0.267 | 0.267 | 0.267 | 45 |
| 4    | 0.160 | 0.229 | 0.188 | 35 | 0.067 | 0.086 | 0.075 | 0.075 | 35 |
| 5    | 0.100 | 0.233 | 0.140 | 30 | 0.114 | 0.133 | 0.123 | 0.123 | 30 |
| 6    | 0.300 | 0.106 | 0.157 | 85 | 0.176 | 0.176 | 0.176 | 0.176 | 85 |
| 7    | 0.114 | 0.080 | 0.094 | 50 | 0.167 | 0.100 | 0.125 | 0.125 | 50 |
| 8    | 0.164 | 0.164 | 0.164 | 55 | 0.200 | 0.127 | 0.156 | 0.156 | 55 |
| 9    | 0.100 | 0.057 | 0.073 | 35 | 0.100 | 0.057 | 0.073 | 0.073 | 35 |
| 10   | 0.100 | 0.267 | 0.145 | 15 | 0.060 | 0.200 | 0.092 | 0.092 | 15 |
| 11   | 0.114 | 0.133 | 0.123 | 30 | 0.160 | 0.133 | 0.145 | 0.145 | 30 |
| 12   | 0.029 | 0.050 | 0.036 | 20 | 0.133 | 0.100 | 0.144 | 0.144 | 20 |
| 13   | 0.120 | 0.120 | 0.120 | 25 | 0.100 | 0.160 | 0.123 | 0.123 | 25 |
| 14   | 0.160 | 0.114 | 0.133 | 35 | 0.114 | 0.114 | 0.114 | 0.114 | 35 |
| 15   | 0.165 | 0.200 | 0.181 | 70 | 0.200 | 0.200 | 0.200 | 0.200 | 70 |
| 16   | 0.133 | 0.089 | 0.107 | 45 | 0.127 | 0.156 | 0.140 | 0.140 | 45 |
| 17   | 0.171 | 0.120 | 0.141 | 50 | 0.171 | 0.120 | 0.141 | 0.141 | 50 |
| 18   | 0.140 | 0.200 | 0.165 | 35 | 0.133 | 0.114 | 0.123 | 0.123 | 35 |
| 19   | 0.120 | 0.120 | 0.120 | 25 | 0.160 | 0.160 | 0.160 | 0.160 | 25 |
| 20   | 0.200 | 0.120 | 0.150 | 25 | 0.080 | 0.160 | 0.107 | 0.107 | 25 |
| Accuracy | 0.141 | 800 | Accuracy | 0.145 | 800 |
| M.A.  | 0.142 | 0.143 | 0.135 | 800 | M.A. | 0.141 | 0.141 | 0.136 | 800 |
| W.A.  | 0.158 | 0.141 | 0.140 | 800 | W.A. | 0.153 | 0.145 | 0.146 | 800 |

CL#: Cluster Number; P: Precision; R: Recall; F: F-Score; SU: Support; M.A.: Macro Average; W.A.: Weighted Average

In Table 9, comparative results of validity indices based on cluster classification metrics are shown under Cosine and Euclidean based metrics. From these results interpreted that Euclidean based metric on average for all clusters to perform better than cosine based metric.

5 Conclusion And Future Scope

A cosine based validation metrics proposed in this paper has the advantage of considering both in implementation of hybrid topic models clustering algorithms and in the validation of formed clusters. Nearness among documents in terms of topics also quantified by not only closeness between two different documents but also their lexical similarity. In this point of view proposed cosine based metrics are more desirable than Euclidean metrics where simply the distance between two clusters will be considered in document clustering. In this paper in cluster validation compactness, separation, number of clusters and classification metrics are considered which will evaluate the classification by considering every element in all clusters of a corpus. Experimentally proved proposed novel cosine based internal and external validity indices work well in cluster validation and in improving the effectiveness of cluster than that of Euclidean validity metrics. However, in case of high sparsity other aspects such as density should also be considered in the evaluation. Performance can be optimized by increasing scalability of their execution in semi-distributed environment and dealing with dynamically changing large datasets in text documents clustering applications.

References

[1] Saeedeh Momtazi (2018). Unsupervised Latent Dirichlet Allocation for supervised question classification. 27918R Information Processing and Management, 54(3): 380-393. DOI:10.1016/j.ipm.2018.01.001.
[2] Filipe Rodrigues, Mariana Lourenc, Bernardete Ribeiro, Francisco, Pereira (2018). Learning Supervised Topic Models for Classification and Regression from Crowds IEEE Transactions on Pattern Analysis and Machine Intelligence,1:1-15. DOI: 10.1109/TPAMI.2017.2648786.
[3] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). Indexing by latent semantic analysis. Journal of American Society for Information Sciences, 41(6):391-407.
[4] Thomas Hofmann (1999). Probabilistic latent semantic indexing. SIGIR, ACM, pp:50-57.
[5] Blei D.M, Ng A.Y, Jordan M.I. (2003). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3:993–1022. https://dl.acm.org/doi/10.5555/944919.944937.

[6] Choo J, Lee C, Reddy C.K, Park H (2013). Utopian: User-driven topic modelling based on interactive non-negative matrix factorization. IEEE Transactions on Visualization and Computer Graphics, 19(12):1992–2001. https://doi.org/10.1109/TVCG.2013.212.

[7] Rajendra Prasad K, Moulana Mohammed, Noorullah RM (2019). Visual Topic Models for Healthcare Data Clustering. Evolutionary Intelligence, 1:1-17. https://doi.org/10.1007/s12605-01900300-y.

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

[9] Alessia Amelio, Clara Pizzuti (2015). Is Normalized Mutual Information a Fair Measure for Comparing Community Detection Methods?. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. DOI: 10.1109/ICSN.2016.7449821.

[10] Dan A, SimoviciE, SzmonJaroszewicz (2002). AnK Axiomatization of Partition Entropy. IEEE Transactions on Information Theory, 48(7):2138-2142. DOI:10.1109/TIT.2002.1013159.

[11] Xue 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.

[12] Li Z, Yan M (2016). A cosine based validation measure for document clustering data with the presence of missing values by ensemble approach. Second Asian Conference on Defense Technology. DOI:10.1109/acdt.2016.7437660.

[13] Tapana Mekaroonkamon, Sarwan Wongsa (2016). 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, Thailand. DOI:10.1109/ICACI.2016.7449821.

[14] Dan A, SimoviciE, Szmon Jaroszewicz (2002). AnK Axiomatization of Partition Entropy. IEEE Transactions on Information Theory, 48(7):2138-2142. DOI:10.1109/TIT.2002.1013159.

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

[16] Kuo-Lung Wu (2008). An analysis of the robustness of partition coefficient index. IEEE International Conference on Fuzzy Systems. DOI:10.1109/fuzzy.2008.4630393.

[17] Vergani A, Binaghi E (2018). A Soft Davies-Bouldin Separation Measure. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). DOI:10.1109/fuzz-ieee.2018.8491581.

[18] Simona Babil, Michelangelo, Maria Spano (2016). A cosine based validation measure for document clustering. JADT 2016. Corpus ID: 195322770.

[19] Oghbaie M, Mohammadi Zanjireh M (2018). Pairwise document similarity measure based on present term set.J Big Data, 5(2). https://doi.org/10.1186/s40537-018-0163-2.

[20] Muhammad Rafi et. al. (2016). Exploiting document level semantics in document clustering. International Journal of Advanced Computer and Applications (IJACSA). 7(6): 462-469. DOI:10.14569/IJACSA.2016.070660.

[21] Shi Zhong (2005). Efficient online Spherical K-means Clustering, IEEE International Conference on Neural Network. DOI: 10.1109/JCNN.2005.1556436.

[22] Dataset TREC2018 Keyword phrases https://trec.nist.gov/pubs/trec24/trec2018.html.

[23] WongkotSriurai, PhayungMeasad, ChoochartHaruechaisak (2010). Hierarchical Web Page Classification Based on a Topic Model and NeighboringPages Integration. (IJCSI) International Journal of Computer Science and Information Security, 7(2). Cite as: arXiv:1510[cs.LG].

[24] Nallamala S.H. Mishra P and Konuru SV, Brest cancer detection using machine learning, International journal of Recent Technology and Engineering. 7(5S4), 2019, pp:478-481.

[25] Vamsidhar, E., Saichandana, B., & Harikiran, J. (2018). A novel approach for feature selection and classifier optimization compressed medical retrieval using hybrid cuckoo search. Indian Journal of Electrical Engineering and Informatics, 6(4), 410-417. DOI:10.11591/ijeei.v6i4.584.

[26] P. Jahnvi. VamsidharEniredly, C. Kalthikeyan (2019). Facial expression detection of all emotions and face recognition system, International Journal of Emerging Trends in Engineering Research. 7, pp:778-783. DOI: 10.30534/ijeter/2019/087122019.

[27] Devisetty, Vidyullatha.P,”Sentiment analysis of tweets using rapid miner tool”, International Journal of Innovative Technology and Exploring Engineering. 2019.
[28] Mohammed Ismail B, K. Bhanu Prakash, M. Nagabhushana Rao “Collaborative Filtering-Based Recommendation of Online Social Voting” International journal of Engineering and Technology 7 (3), pp:1504-1507 July 2018.

[29] K.Srinivas, Dr.MohammedIsmail.B “Testcase Prioritization with Special Emphasis On Automation Testing Using Hybrid Framework” Journal of Theoretical and Applied Information Technology 96(13) pp:4180-4190, July 2018.

[30] Dr.K.V.D.Kiran,”Performance analysis of hybrid hierarchical K-means algorithm using correspondence analysis for thyroid drug data”.Journal of Advanced Research in Dynamical and Control Systems,Volume 10, Issue 12 Special Issue, 2018, Pages 698-712. ISSN:1943023X

[31] Dr.K.V.D.Kiran,”Computational approach to overcome over-lapping of clusters by fuzzy k-means”.International Journal of Recent Technology and Engineering,Volume 7, Issue 4, 2018, Pages350-355,ISSN:22773878

[32] Rahul Shahne, Mohammed Ismail, CSR Prabhu “Survey on Deep Learning Techniques for Prognosis and Diagnosis of Cancer from Microarray Gene Expression Data” Journal of computational and theoretical Nanoscience 16 (12), 5078-5088, Dec 2019.

[33] Dr.Amarendra K.et.al,Image steganography using LSB, International Journal of Science Technology and Research, Volume 8, Issue 12, December 2019.