Mieczysław A. Kłopotek
ORCID: 0000-0003-4685-7045

Sławomir T. Wierzchoń
ORCID: 0000-0001-8860-392X

Bartłomiej Starosta
ORCID: 0000-0002-5554-4596

Dariusz Czerski
ORCID: 0000-0002-3013-3483

Piotr Borkowski
ORCID: 0000-0001-9188-5147

Institute of Computer Science of Polish Academy of Sciences
ul. Jana Kazimierza 5,
01-248 Warszawa, Poland

{klopotek,stw,barstar,dcz,p.borkowski}@ipipan.waw.pl

An Evaluation Methodology for Explanations of Clustering Results in Textual Domain

DOI: 10.34739/si.2025.33.04

Abstract. Explainability has become a must for any AI algorithm acting as a black-box, like Graph Spectral Clustering (GSC). Though a success was achieved in developing respective explanation methodologies, a new challenge has to be faced: the evaluation of the (quality of) explanations.

Several evaluation methods for explanations in AI have been developed, but they turn out to have some shortcomings, making them unsuitable for GSC explanation evaluation. In this paper, we recall some of these methods and point out their respective shortcoming. Based on this investigation, we suggest a new explanation methodology oriented towards GSC and similar methods and present a small experiment on the usage of this methodology.

Keywords: Explainable AI, Graph Spectral Clustering, Evaluation of Explanations

1 Introduction

Artificial Intelligence has experienced significant development over recent years, such that even the general public is enthusiastic about its new, astonishing capabilities and application possibilities. This development is spurred and driven by large progress in the domain of Large Language Models (LLM). These emerging capabilities did not occur from nothing, as AI methods were steadily improved over the past decades, providing more and more accurate answers. However, the improved accuracy is achieved at the expense of model complexity, which becomes non-comprehensible, effectively turning the model into a black box.

As vital decisions are going to be based on AI results, like those in business [41], administration [12], court of law [32], military [31] or health care [40], there exists an urgent need to find ways to convince the end users that AI suggestions are rational. This is quite urgent as the missing explanations became a background for fearmongering, especially in cases of failed decision proposals [39].

This problem led to the development of so-called Explainable AI (XAI) that would present reasons for the decision to the end user, see, e.g., [5, 6, 13, 18, 20, 27, 35, 38]. In particular, methods for explaining clusterings [2, 9, 10], and particularly clusterings of textual documents were developed [14, 21, 33, 53].[1] In this way, trust in the AI proposals will be enhanced, usability of AI will be improved, and innovations will be fostered. Finally, accountability can be achieved because AI system decisions can be subject to a detailed analysis of how they were made.

Development of XAI, however, gave rise to the next issue: how to convince that the explanations are valid. Therefore, there is a surge in efforts to create methods of Quality of Evaluation of ML Explanation (QEMLE), see, e.g., [4, 16, 17, 26, 28, 29, 34, 42, 48, 54–58].

In this paper, we will review major trends in QEMLE in Section 2. Then, in Section 3, we will point out problems related to the application of general QEMLE principles to clustering of textual documents using the Graph Spectral Methods. Out of the insights of that section, we propose in Section 4 QEMLE addressing particularly the issues related to QEMLE for Graph Spectral Clustering. We present a small evaluation of this proposal in Section 5. The paper ends with some conclusions presented in Section 6.

[1] An easy-to-read introduction to this subject can be found in the blog "Explainability in Clustering Algorithms: A Survey Paper with Experimental Results" available from `https://medium.com/@srujanaharshinicitd/explainability-in-clustering-algorithms-a-survey-paper-with-experimental-results-65fca1cb9bb5`.

2 Previous work

There exist multiple works ([4, 16, 17, 37, 42]) and surveys on assessing the quality of explanations ([23, 26, 48, 57]). Most quality assessment methods require either direct interaction with the user or the use of predefined datasets containing explanations based on "real-world data."

In this article, we argue that, for document clustering methods, a different approach is needed: an analytical justification of why the explanations are reliable. Let us first outline how the issues have been approached so far.

As summarized by [57], the goal of QEMLE is to assess either interpretability or fidelity of the explanation method, or both. Hereby, these terms are to be understood as follows [28]:

- interpretability: clarity, parsimony, and broadness,
- fidelity: completeness and soundness.

According to [29], the QEMLE can be divided into three categories:

1. application-grounded evaluation,
2. human-grounded evaluation,
3. functionality-grounded evaluation.

Application-grounded evaluation requires experiments with end-users. For predefined complex tasks that reflect real-world applications, we seek to determine to what extent these explanations help industry experts in carrying out these tasks.

Human-grounded evaluation requires experiments with lay humans. Simpler human–subject experiments are conducted that reflect the essence of the target application, but are not that complex. Involving non-experts helps reduce costs and allows for experiments involving a larger group of participants. It is argued that this approach to evaluation depends solely on the quality of the explanation, regardless of the type of explanation or the accuracy of the associated predictions. Functionality-grounded evaluation does not require human experiments. Instead, a comparison is made with some approximation of the formal definition of interpretability, for example, the depth of the decision tree being output as an explanation.

There exist both qualitative and quantitative metrics to evaluate explanations for human experiments. Qualitative metrics include asking about the usefulness of, satisfaction with, confidence in, and trust in provided explanations by means of interviews or questionnaires [19, 54, 56, 58]. Quantitative metrics include measuring human-machine task performance in terms of accuracy, response time needed, likelihood to deviate, or ability to detect errors [34, 54] and physiological responses from humans during experimental tasks [55].

In the last category, i.e., functionality-grounded evaluation, the evaluations of explanations depend on the types of explanation, which themselves can be divided into:

- model-based explanations (e.g., decision trees),
- attribution-based explanations (e.g., the ranked lists of attributes contributing to the result), and
- example-based explanations (e.g., selecting instances that are well predicted or not well predicted by the model as explanations).

Respective evaluation metrics are summarized in the paper [57].

Model-based explanation quality is assessed by measuring the complexity of the model used for explanation. Measures encompass

- model size (e.g., number of nodes or tree depth for a decision tree, number of rules and rule length in rule-based models),
- model run-time (count of Boolean or arithmetic operations executed by the explanation model for an input),
- output susceptibility (to changes of values of input variables),
- complexity of linear approximation (number of parameters to approximate the output locally by a piece-wise linear model),
- degree of interaction between input features,
- model-to-explanation agreement (the percentage of agreements between the original model and the explanation model – not applicable if the explanation model is the same as explanation model).

Methods of evaluation of attribution-based explanations have the following brands:

- correlation between original input data and input data to the explanation model, same as output,
- sensitivity (at business-meaningful level) of output to the value range of input features,
- n features removal sensitivity (variance after removal of n features equals to the contributions of the n removed variables).
- one variable removal sensitivity.

Example-based explanations summarize a model by a set of representative examples. Under these circumstances, the quality of the explanation is measured via:

- the size of the set of examples,
- the diversity of examples.

There also exists a tendency to develop explanation evaluation methods targeting particular needs while stress is made on the elimination of human engagement. LLM technology was recently applied for the evaluation of causal and counterfactual explanations, [7], evaluation of convincingness, clarity, and accuracy of explanations taking into account aspects like privacy-preservation [43], explanations of dashboard contents [11], etc. [49] proposes sensitivity analysis of explanations in the computer vision domain. [1] proposes a framework for evaluating explanation methods in healthcare. [15] proposes to evaluate explanations based on synthetic ground truth explanations to avoid the need for human interaction. Another method of evaluation without real ground truth was proposed in [36]. Interestingly, [10] proposes to assign a small set of exemplars that nicely characterize each cluster.

A historical overview of evaluation methods for explanations may be found in [30].

3 Issues in Evaluating Explanations for Clustering of Textual Documents

The evaluation methodologies mentioned in the literature are not well suited for evaluating explanations for Graph Spectral Clustering (GSC) methods [24, 51] applied to textual documents. The basic advantage of using GSC for textual documents is the reduction of the

clustering embedding space dimension even by a factor of 1,000. Graph Spectral Clustering methods can be viewed as a relaxation of cut-based graph clustering methods.

Let S be a (symmetric) similarity matrix between pairs of items (e.g., documents), representing a weighted graph whose nodes correspond to the items (documents), while the weights are the similarities between items. It is generally assumed that this graph is deemed as one without self-loops (in spite of the fact that an item is most similar to itself). Hence, the diagonal of S is assumed to be filled with zeros (see e.g., [50] as one example of many). In the domain of text mining, the similarity matrix is usually based on either a graph representation of relationships (links) between items (text documents) or such a graph is induced by (cosine) similarity measures between items. However, mixed object representations (text and links) have also been studied [52]. By convention, all diagonal elements of the matrix S are equal to zero. We concentrate here on the document texts.

A combinatorial, or unnormalised, Laplacian L corresponding to this matrix is defined as

$$L = D - S, \tag{1}$$

where D is the diagonal matrix with $d_{ii} = \sum_{\ell=1}^{n} s_{i\ell}$ for each $i \in [n]$.

A normalized Laplacian $\mathfrak{L}$ of the graph represented by S is defined as

$$\mathfrak{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} S D^{-1/2}. \tag{2}$$

D^{-1} is a pseudo-inverse of D. Though other Laplacians are also discussed in the literature, we focus on the above two here.

Whichever Laplacian is used, the partition of a data-set into k clusters is performed as follows. One computes the eigen-decomposition of the respective Laplacian, getting n eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$ (always $\lambda_1 = 0$) and corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Then one embeds the documents in the k-dimensional space spanned by the k eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ corresponding to k lowest eigenvalues. That is, one assigns each document i the coordinates $[v_{i,1}, \ldots, v_{i,k}]$. This shall be called L-embedding if the combinatorial Laplacian L is used, and N-embedding if the normalized Laplacian $\mathfrak{L}$ is used.[2] Then one clusters the documents in this embedding using, e.g., k-means algorithm. Detailed descriptions can be found, e.g., in [24, 51]. When the clustering is finalized, then for each item i we possess its coordinates $[v_{i,1}, \ldots, v_{i,k}]$ and its membership in some cluster C_j, but we cannot tell why i belongs to C_j because none of the coordinates $[v_{i,1}, \ldots, v_{i,k}]$ has anything to do with the contents of the document i, in particular with its term frequency (tf, tfidf) or any other content representation. Therefore an explanation methodology designed for GSC had to be developed. Several methodologies emerged, including [3, 22, 44–46].

The majority of the methods follows the scheme:

1. The original documents are placed in some original embedding space, like Term Vector Space (TVS), GloVe, BERT.
2. Cosine similarity is computed yielding the matrix S.
3. Laplacian of S is computed and clustering is performed.

[2] Please note that the L-embedding approximates the so-called Ratio Cut (or RCut) and the N-embedding approximates the Normalized Cut (or NCut), [25].

4. The theoretical foundations are provided yielding equivalence between clustering in the original space (which is ineffective), clustering in an intermediate (usually a kernel) space, (which is for theoretical purposes only, not used in practical computation), and a clustering in GSC space (in which the actual clustering is performed).
5. Based on this equivalence, the explanation of clusters from GSC is computed as if the clustering was performed in the original space (yielding a vector of words best characterizing each cluster, or document membership in a cluster, etc.)

It should be stressed that GSC clustering methods, as described in [24, 51], until recently were considered black-box methods. Only recently, explanation methods were elaborated in the mentioned references. Hence, evaluation of such explanations remains a hot topic. Neither Human-grounded evaluation nor Application-grounded evaluation are feasible as the explanation methodology still waits for being incorporated in widely used (clustering) systems. Hence, no human subject experiments are feasible.

Methods of Functionality-grounded evaluation are not applicable because:

- Investigations of explanation complexity are pointless because the explanations are simplistic in nature: a simple list of ranked top words, characterizing a cluster.
- Tests of leaving out some words from the documents are pointless due to the nature of natural language, as here the same concept can be expressed by different words.
- Stability tests are also not very helpful because of the nature of the underlying classical clustering algorithm (k-means). k-means seeks in a stochastic way the local minimum of the loss function, and therefore clustering results may (slightly) differ from run to run. Also, the GSC methods are in fact approximations of graph cut methods, constituting another source of instability between runs.
- Other methods from this area are not applicable either.

4 A Proposal of Evaluation Methods for Explanations for Clustering of Textual Documents

The problems mentioned above urged us to develop a new approach to the evaluation of explainability, which may be called assumptions-based. The validity of explanations relies on a chain of nearly equivalence relations (see [45]). So the quality of explanations relies in fact on the degree of fitness to the various assumptions (e.g., the cluster balance, impacting both the explanation process and the clustering process). The new methodology means then defining the deviation degrees of individual steps and their combination.

Hence we can distinguish here the following types of quality measures:

- evaluation of the final result of explanation process (Sec 4.1),
- evaluation of fitness of the assumptions of individual explanation steps to the actual data (Sec. 4.3),
- evaluation of the final result of explanation process versus the "ideal" process (Sec. 4.2),
- evaluation of sensitivity of the explanation to the leave-one-out testing (Sec. 4.4).

4.1 Evaluation of the final result of explanation process

After the explanation process each cluster C_i is assigned an ordered list of words L_i, where the ordering depends on the importance of a given word for the cluster center. This importance may take the form of, e.g.,

- the weight assigned to the word when forming the cluster center out of the vectors of words, or, more simplistically
- the count of occurrences of the word within documents assigned to the cluster.

The simplest statistics is the presence of explanation words in the cluster documents which can be summarized in different ways, like

- the positions of the words from the documents from cluster C_i on the explanation list L_i of a given cluster; in particular one seeks the word from a document d_j that has the highest rank $r_i(d_j)$ on L_i; one then computes either the average of $r_i(d_j)$ over all documents $d_j \in C_i$, or computes the histogram, or some other statistics;
- the positions of the words from the documents on the explanation list of a given cluster $r_i(d_j)$ versus presence in the other clusters $r_{i'}(d_j)$, $i' \neq i$; one would expect that $r_i(d_j)$ is numerically lower than $r_{i'}(d_j)$; one would then compute averages, histograms etc. of the difference $r_{i'}(d_j) - r_i(d_j)$ or of $sgn(r_{i'}(d_j) - r_i(d_j))$;
- the positions of the words from the explanation list of a given cluster L_i versus presence on the explanation lists of other clusters $r_{i'}(L_{ij})$, $i' \neq i$; one would expect that $r_i(L_{ij})$ is numerically lower than $r_{i'}(L_{ij})$; one would then compute averages, histograms etc. of the difference $r_{i'}(L_{ij}) - r_i(L_{ij})$ or of $sgn(r_{i'}(L_{ij}) - r_i(L_{ij}))$.

In the above, usually one will not consider all the words, but rather restrict oneself to n top words for each cluster. This razor will generally reduce the potential noise from rare words. Furthermore, one may remove from considerations general stop words or stop words specific to application area.

4.2 Evaluation of the final result of explanation process versus the "ideal" process

One takes two explanations, one based on the result of the clustering and another based on intrinsic clustering and computes measures mentioned above (Sec. 4.1) for each of them. Then one compares how much do these measures disagree between the two clusterings.

The disagreement may be presented in form of histograms or some aggregated statistics.

Note that primarily to such a comparison, matching of actual clusters and intrinsic clusters has to be performed. We apply the Hungarian algorithm [8].

While the results of the method mentioned in Sec. 4.1 depend solely on the inner consistency of the clustering and on natural properties of the data, the results obtained in this section are affected by the ability of the algorithm to discover the clusters "from human perspective" and the method requires external information via human labeling.

4.3 Evaluation of fitness of the assumptions of individual explanation steps to the actual data

This approach relies on details of the clustering process and include (in case of the clustering methods described in [45]):

- comparison of the actual k-means fitness measure against the ideal one (best result over a long series of clusterings),
- verification of the assumption of how much the cluster sizes differ from uniform clusters,
- histograms of similarities to the one's own cluster center versus other (closest) cluster centers.

The first point is related to the known property of k-means of sticking in local optimization minima. As the global minimum of k-means is hardly ever known, it is advantageous to use for quality evaluations results of long-run k-means as the "ideal" reference.

The second point is closely related to the effect described in [47]. It turns out that, in case of some GSC methods, k-means tends to detect small groups of documents at large distance from the remaining documents which is disadvantageous for understanding the clustering. Upon occurrence of this effect, methods proposed in [47] may help to eliminate this effect.

As the last point is concerned, it is well-known that one can get clusters very similar to one another (especially in the explanation aspect) in case that the chosen number of clusters is too high compared to true clusters. In this case the k-means algorithm may try to split intrinsic clusters in smaller ones that will then be quite similar, both in terms of contents and in terms of their explanations.

4.4 Evaluation of sensitivity of the explanation to the leave-one-out testing

Generally, leave-one-out testing is deemed to run comparisons when dismissing single documents from the data. Our approach is, however, to leave entire cluster out and perform a clustering into a lower number of clusters. The expectation is that ranks of words shall not differ too much between the original clustering and the clustering in reduced document set.

So the procedure would be: (1) Cluster the original data into k clusters, (2) Create an explanation for the clustering. (3) For each cluster (3a) leave this cluster out of the data, (3b) cluster the remaining data into $k-1$ clusters, (3c) Match the obtained clusters and the original ones, e.g. using the Hungarian Algorithm, (3c) compare the rankings of explaining terms in the newly obtained clusters and in the original ones. (4) Compute a summary statistics for the loop over clusters.

Note that, as a side effect, the stability of the clustering can be verified in the process (by checking the degrees of agreement between membership of corresponding clusters).

5 Experimental evaluation

5.1 Experimental setup for testing the evaluation methods

The team has collected over years tweets which will be the basis for the verification process. We define an intrinsic cluster as the set of documents that has one single hashtag assigned to them. Documents with two or more hashtags are dismissed from consideration.

Sets of 10 intrinsic clusters from cluster pool will be randomly selected. Then the clustering and explanation process will be performed. Subsequently, the evaluations, described in this section was performed.

In the experiment, the results of which are shown in Section 5.2 below, a set of tweets associated with the following 10 hashtags was used: #90dayfiance, #tejran, #ukraine, #tejasswiprakash, #nowplaying, #anjisalvacion, #puredoctrinesofchrist, #1, #lolinginlove, #bbnaija.

In the experiment with 4 hashtags, the following ones were used: #nowplaying, #anjisalvacion, #puredoctrinesofchrist, #lolinginlove.

We tested how clearcut are the clusters with respect to words characterizing them. For each cluster, its centroid was computed as the count of word occurrences in tweets belonging to a cluster. We then computed the pseudo-distance between the clusters. It runs as follows:

1. For each cluster, we calculate the centroid (i.e., a sorted list by frequency of all tweets in a given grouping): $c_i = [w1, w2, ...]$
2. We remove all stopwords and words starting with '@' from the tweet text. Thereafter we keep 20 top frequency words for each cluster.
3. Given two lists of words representing the centroids $c_1 = [w1, w2, ..., w20]$ and $c_2 = [ww1, ww2, ..., ww20]$ we calculate the rank distance $dist(c_1, c_2)$ as follows: we check the ranks of the words ww1, ww2, and ww20 in list c_1, and then average these ranks.

Notes:

- Note 1: If a given word does not appear in list c_1, we assign it a rank of N=20.
- Note 2: The above $dist()$ function is not symmetric, so we present the full distance matrix between centroids.
- Note 3: We assumed that $dist(x, x) = 0$.

If the two lists completely agree, then the pseudo-distance amounts to 10.5. If they are completely without common elements, the distance is 20.

We also introduced computation of pseudo-distance between cluster center and a document as follows: A distance between a document and a cluster center is the rank of a term from the document ranking highest on the centroid of the cluster. For all documents from one cluster, the average of the above was taken when comparing with a cluster centroid. One expects that for a good explanation this average should be low for the cluster centroid of the same cluster from which the documents stem, and high when comparing with other cluster centers.

5.2 Some results

We performed a number of clustering, explanation and explanation evaluation tasks. In one of the experiments with 10 clusters we used the GSC clustering method using `CountVectorizer` vectorizer with parameters `+SW` (preserve stopwords) and `+HT` (keep hashtags) and with precomputed `affinity` (cosine similarity matrix as input), and normalized Laplacian. The resulting explanations for the mentioned clustering task result are presented in Table 1.

We performed two evaluations of clustering explanations: (1) The similarity of explanations between different clusters, presented in table 2; (2) The similarity of document contents of a cluster to the cluster explanations for all clusters, presented in Table 3.

Table 1: The explanations for the clustering task

Centroid	Explanation words and their occurrence counts
Centroid nr: 0	[('$goo', 9), ('event', 9)]
Centroid nr: 1	[('singles', 25), ('extended', 22), ('remix', 18), ('baby', 15), ('nikko', 14), ('spears', 14), ('pitt', 14), ('alisha', 14), ('dj', 14), ('production', 14), ('dave', 14), ('bee', 14), ('gees', 14), ('johnny', 14), ('&', 12), ('williams', 11), ('rose', 9), ('culture', 8), ('surender', 8), ('talk', 8)]
Centroid nr: 2	[('plz', 47), ('rescue', 47), ('pet', 47), ('vulnerable', 36), ('starving', 7)]
Centroid nr: 3	[('shoutout', 1), ('shoutouts', 1)]
Centroid nr: 4	[('unknown', 31), ('mp3', 23), ('feat', 17), ('radio', 14), ('deep', 14), ('girl', 14), ('remix', 12), ('meremix', 12), ('&', 11), ('sweet', 9), ('extended', 8), ('mart', 8), ('finally', 8), ('2021', 8), ('boutique', 8), ('vip', 8), ('mix', 8), ('music', 7), ('howard', 7), ('jones', 7)]
Centroid nr: 5	[('anji', 616), ('version', 330), ('people', 292), ('king', 240), ('kjv', 236), ('james', 235), ('like', 219), ('god', 217), ('just', 195), ('dont', 194), ('love', 188), ('life', 185), ('tejasswi', 175), ('app', 169), ('live', 169), ('day', 158), ('1', 156), ('happy', 150), ('–', 146), ('way', 140)]
Centroid nr: 6	[('madonna', 28), ('remix', 20), ('production', 14), ('rudil', 14), ('lacey', 14), ('holidayremix', 12), ('rfb', 9), ('mix', 9), ('extended', 9), ('alisha', 7), ('causing', 7), ('nano', 7), ('shifer', 7), ('nikko', 7), ('sreason', 7), ('rain', 7), ('stargazingrfb', 6), ('commotionrfb', 6), ('fourtunatoextended', 6), ('cultureremix', 6)]
Centroid nr: 7	[('pass', 24), ('sold', 24), ('free', 24), ('mint', 24)]
Centroid nr: 8	[('servilely', 5), ('fearfully', 5), ('evaluation', 5), ('pugnacious', 5), ('clutch', 4), ('virgin', 4), ('degrease', 4), ('congo', 4), ('flatterer', 4), ('isomer', 4), ('middleage', 4), ('principalities', 4), ('hinting', 4), ('publicity', 4), ('consoles', 4), ('outburst', 4), ('matrixes', 4), ('roister', 4), ('statutes', 4), ('climatically', 4)]
Centroid nr: 9	[('just', 281), ('like', 278), ('people', 272), ('dont', 229), ('tejasswi', 182), ('love', 179), ('version', 175), ('king', 145), ('james', 137), ('im', 136), ('kjv', 136), ('god', 129), ('make', 122), ('good', 118), ('said', 108), ('want', 100), ('know', 99), ('1', 97), ('&', 95), ('time', 92)]

Table 2: Pseudo-distances between clusters

	$cl0$	$cl1$	$cl2$	$cl3$	$cl4$	$cl5$	$cl6$	$cl7$	cl	$cl9$
$cl0$	0.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00
$cl1$	20.00	0.00	20.00	20.00	18.00	20.00	16.40	20.00	20.00	19.75
$cl2$	20.00	20.00	0.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00
$cl3$	20.00	20.00	20.00	0.00	20.00	20.00	20.00	20.00	20.00	20.00
$cl4$	20.00	18.35	20.00	20.00	0.00	20.00	18.75	20.00	20.00	19.45
$cl5$	20.00	20.00	20.00	20.00	20.00	0.00	20.00	20.00	20.00	12.75
$cl6$	20.00	16.90	20.00	20.00	17.95	20.00	0.00	20.00	20.00	20.00
$cl7$	20.00	20.00	20.00	20.00	20.00	20.00	20.00	0.00	20.00	20.00
$cl8$	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	0.00	20.00
$cl9$	20.00	19.95	20.00	20.00	19.95	12.30	20.00	20.00	20.00	0.00

Table 3: Average pseudo-distances between documents and clusters

	$cl0$	$cl1$	$cl2$	$cl3$	$cl4$	$cl5$	$cl6$	$cl7$	cl	$cl9$
$dc0$	1.00	20.00	20.00	20.00	20.00	19.95	20.00	20.00	20.00	19.99
$dc1$	20.00	4.06	20.00	20.00	13.02	19.56	4.91	20.00	20.00	19.57
$dc2$	20.00	20.00	1.00	20.00	20.00	19.98	20.00	20.00	20.00	19.95
$dc3$	20.00	20.00	20.00	1.50	20.00	20.00	20.00	20.00	20.00	20.00
$dc4$	20.00	13.48	20.00	20.00	9.96	18.82	9.73	20.00	19.98	18.65
$dc5$	20.00	20.00	20.00	20.00	19.04	11.35	19.73	20.00	19.94	13.55
$dc6$	20.00	10.50	20.00	20.00	15.00	19.83	1.91	20.00	19.98	19.84
$dc7$	20.00	20.00	20.00	20.00	19.70	19.81	20.00	1.00	19.99	19.83
$dc8$	20.00	20.00	20.00	20.00	20.00	19.99	20.00	20.00	19.37	19.99
$dc9$	20.00	19.92	20.00	20.00	18.79	12.73	19.82	20.00	19.93	11.76

As shown in Table 2, the explanations of distinct clusters differ strongly from one another (nearly all pseudo-distances of about 20). Only clusters 9 and 5 seem to have close explanations.

As shown in Table 3, cluster explanations fit very well the actual cluster document contents while clearly separating from other clusters.

In another experiment, performed with 4 clusters, we used the GSC clustering method using `CountVectorizer` vectorizer with parameters `+SW` (preserve stopwords), `+HT` (keep hashtags), with precomputed `affinity` (as cosine similarity matrix), and with normalized Laplacian. The resulting explanations for the mentioned clustering task result are presented in Table 4.

We performed two evaluations of clustering explanations: (1) The similarity of explanations between different clusters, presented in Table 5; (2) The similarity of document contents of a cluster to the cluster explanations for all clusters, presented in Table 6.

Table 4: The explanations for the clustering task with 4 clusters

Centroid	Explanation words and their occurrence counts
Centroid nr: 0	[('radio', 37), ('info', 24), ('global', 18), ('com', 18), ('singles', 18), ('airplay', 17), ('email', 17), ('goglobalradio@gmail', 17), ('remix', 16), ('extended', 15), ('unknown', 14), ('girl', 14), ('deep', 14), ('rudil', 14), ('feat', 13), ('meremix', 12), ('yourfb', 12), ('youextended', 12), ('peter', 10), ('la', 9)]
Centroid nr: 1	[('anji', 166), ('dalampasigan', 44), ('–', 37), ('people', 33), ('feelstheconcert', 30), ('life', 29), ('salvacion', 28), ('happy', 25), ('things', 22), ('mv', 21), ('love', 20), ('birthday', 19), ('success', 16), ('day', 16), ('need', 16), ('talk', 15), ('dont', 14), ('withanji', 14), ('cheers', 13), ('20th', 13)]
Centroid nr: 2	[('version', 161), ('james', 118), ('king', 116), ('kjv', 115), ('god', 77), ('app', 62), ('1', 54), ('standard', 48), ('esv', 48), ('english', 47), ('live', 44), ('proverbs', 42), ('hand', 40), ('christ', 34), ('shall', 34), ('radio', 34), ('listen', 33), ('2', 32), ('lord', 32), ('3', 32)]
Centroid nr: 3	[('redrafting', 4), ('sabotages', 3), ('recurred', 3), ('unclimbable', 3), ('teachings', 3), ('orate', 3), ('entombed', 3), ('demonstratively', 3), ('fax', 3), ('servilely', 3), ('bathing', 3), ('attenuator', 3), ('doubled', 3), ('gullies', 3), ('mention', 3), ('lathe', 3), ('exmember', 3), ('temples', 3), ('chaplain', 3), ('accumulating', 3)]

Table 5: Pseudo-distances between 4 clusters

	$cl0$	$cl1$	$cl2$	$cl3$
$cl0$	0.00	20.00	19.05	20.00
$cl1$	20.00	0.00	20.00	20.00
$cl2$	19.80	20.00	0.00	20.00
$cl3$	20.00	20.00	20.00	0.00

Table 6: Average pseudo-distances between documents and 4 clusters

	$cl0$	$cl1$	$cl2$	$cl3$
$dc0$	8.94	19.77	17.95	19.98
$dc1$	18.95	3.10	16.96	19.90
$dc2$	18.29	19.21	5.65	19.93
$dc3$	20.00	20.00	19.81	18.54

We see from Table 5, the explanations of distinct clusters differ strongly from one another (near all pseudo-distances of about 20). And again, in Table 6, cluster explanations fit very well the actual cluster document contents while clearly separating from other clusters.

Results from other experiments are available from the authors. We performed the experiments with normalized Laplacian, using various methods related to handling of negative similarities like adding a number to the primary similarity, adding a number to the primary similarity and normalizing $((s+n)/(1+n))$ exponential transformations $(exp(-(1-(s+n))/2))$, cosine transformation $(cos(arccos(s)/(1+n)$, normalizing by maximum cosine $(cos((\pi/2) * arccos(s)/maxarccos(S)))$. Generally, the results were similar in nature though in some cases, when the clustering performance was poor, then also the explanations were not convincing.

6 Conclusions

Explainability has become a must for any AI algorithm acting as a black-box, like Graph Spectral Clustering (GSC). Though a success was achieved in developing respective explanation methodologies, a new challenge has to be coped with: the evaluation of the (quality of) explanations.

A number of evaluation methods for explanations in AI have been developed. We reviewed some of them in this paper. Regrettably, these methods do not fit the evaluation needs of GSC explanation evaluation. Both Human-grounded evaluation and Application-grounded evaluation would require as a pre-requisite implementations of the discussed GSC clustering explanation methods in some popular production data mining systems so that evaluators could be engaged without in-depth knowledge. But such systems are currently not available to our knowledge, as the explanation methods themselves are quite recent. On the other hand, existent methods of Functionality-grounded evaluation are not usable either, as they concentrate on aspects like explanation complexity, word-level leave-out-out or stability tests that are in no sense informative for GSC XAI.

Therefore we presented proposals of new evaluation methods for this class of clustering algorithm explanations that avoid human interaction and at the same time cover important aspects of GSC, including the final result of explanation process, both for its inner consistency and idealized outcome, its intermediate steps, sensitivity to disturbances in data at cluster level.

We presented a small evaluation of this proposal. The idea behind these explanation evaluations seems to be appealing.

References

1. Agrawal, K., El Shawi, R., Ahmed, N.: eXplainable artificial intelligence-Eval: A framework for comparative evaluation of explanation methods in healthcare. Digit Health (Sep 2025). https://doi.org/10.1177/20552076251368045
2. Alvarez-Garcia, M., Ibar-Alonso, R., Arenas-Parra, M.: A comprehensive framework for explainable cluster analysis. Information Sciences **663**, 120282 (2024). https://doi.org/10.1016/j.ins.2024.120282
3. Argov, T., Wagner, T.: Spex: A spectral approach to explainable clustering (2025), https://arxiv.org/abs/2511.00885
4. Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., Ghassemi, M.: The road to explainability is paved with bias: Measuring the fairness of explanations. In: 2022 ACM Conference on Fairness Accountability and Transparency. p. 1194–1206. FAccT '22, ACM (Jun 2022), http://dx.doi.org/10.1145/3531146.3533179
5. Barredo Arrieta, et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82 – 115 (2020), https://doi.org/10.1016/j.inffus.2019.12.012
6. Bobek, S., Kuk, M., Szeląż, M., Nalepa, G.: Enhancing cluster analysis with explainable AI and multidimensional cluster prototypes. IEEE Access **10**, 101556–101574 (2022)
7. Bona, F.B.D., Dominici, G., Miller, T., Langheinrich, M., Gjoreski, M.: Evaluating explanations through llms: Beyond traditional user studies (2024), https://arxiv.org/abs/2410.17781
8. Cormen, T.H., Leiserson, C., Rivest, R., Stein, C.: Introduction to algorithms, (4th ed.), chap. 25.3: The Hungarian algorithm for the assignment problem, p. 723–739. The MIT Press (2022)
9. Dasgupta, S., Frost, N., Moshkovitz, M., Rashtchian, C.: Explainable k-means and k-medians clustering. arXiv preprint arXiv:2002.12538 (2020), https://arxiv.org/abs/2002.12538
10. Davidson, I., Livanos, M., Gourru, A., Walker, P., Velcin, J., Ravi, S.S.: Explainable clustering via exemplars: Complexity and efficient approximation algorithms. arXiv preprint arXiv:2209.09670 (2022), https://arxiv.org/abs/2209.09670
11. Deriyeva, A., Paassen, B.: Evaluation of llm-based explanations for a learning analytics dashboard (2025), https://arxiv.org/abs/2511.11671
12. Fonseca, R.G.: Optimizing organizational efficiency through AI-driven administrative technology. World Journal of Advanced Research and Reviews **26**(01), 2156–2158 (2025). https://doi.org/https://doi.org/10.30574/wjarr.2025.26.1.1212
13. Gamlath, B., Jia, X., Polak, A., Svensson, O.: Nearly-tight and oblivious algorithms for explainable clustering. CoRR **abs/2106.16147** (2021), https://arxiv.org/abs/2106.16147
14. Guan, R., Zhang, H., Liang, Y., Giunchiglia, F., Huang, L., Feng, X.: Deep feature-based text clustering and its explanation. IEEE Transactions on Knowledge and Data Engineering **34**(8), 3669–3680 (2022). https://doi.org/10.1109/TKDE.2020.3028943
15. Guidotti, R.: Evaluating local explanation methods on ground truth. Artificial Intelligence **291**, 103428 (2021). https://doi.org/https://doi.org/10.1016/j.artint.2020.103428, https://www.sciencedirect.com/science/article/pii/S0004370220301776

16. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Front. Comput. Sci., Sec. Theoretical Computer Science, 06 February 2023 **5** (2023). https://doi.org/https://doi.org/10.3389/fcomp.2023.1096257
17. Holzinger, A., Carrington, A.M., Müller, H.: Measuring the quality of explanations: The system causability scale (SCS). comparing human and machine explanations. CoRR **abs/1912.09024** (2019), http://arxiv.org/abs/1912.09024
18. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W.: Explainable AI Methods - A Brief Overview, chap. 1, pp. 13–38. Springer International Publishing, Cham (2022). https://doi.org/url{10.1007/978-3-031-04083-2_2}
19. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B.: An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support Syst. **51**(1), 141–154 (Apr 2011), https://doi.org/10.1016/j.dss.2010.12.003
20. Khan, R., Dofadar, D., Alam, M.G.R.: Explainable customer segmentation using k-means clustering. IEEE Explore pp. 0639–0643 (12 2021). https://doi.org/10.1109/UEMCON53757.2021.9666609
21. Kim, W., Nam, K., Son, Y.: Categorizing affective response of customer with novel explainable clustering algorithm: The case study of amazon reviews. Electronic Commerce Research and Applications **58**, 101250 (2023). https://doi.org/https://doi.org/10.1016/j.elerap.2023.101250, https://www.sciencedirect.com/science/article/pii/S1567422323000157
22. Kłopotek, M., Wierzchoń, S.T., Starosta, B., Borkowski, P., Czerski, D.: Towards explainable graph spectral clustering for BERT embeddings. Journal of Automation, Mobile Robotics, and Intelligent Systems **20**(1), 53–65 (2026). https://doi.org/10.14313/jamris-2026-005
23. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1) (2021). https://doi.org/10.3390/e23010018
24. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007). https://doi.org/10.1007/s11222-007-9033-z
25. Luxburg, U.v.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007), http://dx.doi.org/10.1007/s11222-007-9033-z
26. Löfström, H., Hammar, K., Johansson, U.: A Meta Survey of Quality Evaluation Criteria in Explanation Methods, p. 55–63. Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-07481-3_7
27. Makarychev, K., Shan, L.: Explainable k-means. don't be greedy, plant bigger trees! In: Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, June 2022. pp. 1629–1642 (2022). https://doi.org/10.1145/3519935.3520056
28. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. CoRR **abs/2007.15911** (2020), https://arxiv.org/abs/2007.15911
29. Michel, A.H.: The black box, unlocked: Predictability and understand-ability in military AI; (2020), https://unidir.org/files/2020-09/BlackBoxUnlocked.pdf
30. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys **55**(13s), 1–42 (Jul 2023). https://doi.org/10.1145/3583558, http://dx.doi.org/10.1145/3583558
31. Nitzl, C., Cyran, A., Krstanovic, S., Borghoff, U.M.: The use of artificial intelligence in military intelligence: an experimental investigation of added value in the analysis process. Frontiers in Human Dynamics **7** (May 2025). https://doi.org/10.3389/fhumd.2025.1540450, http://dx.doi.org/10.3389/fhumd.2025.1540450
32. Nowotko, P.M.: AI in judicial application of law and the right to a court. Procedia Computer Science **192**, 2220–2228 (2021). https://doi.org/https://doi.org/10.1016/j.procs.2021.

08.235, https://www.sciencedirect.com/science/article/pii/S1877050921017324, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021
33. Penta, A., Pal, A.: What is this cluster about? explaining textual clusters by extracting relevant keywords. Knowledge-Based Systems **229**, 107342 (2021). https://doi.org/https://doi.org/10.1016/j.knosys.2021.107342
34. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.M.: Manipulating and measuring model interpretability. CoRR **abs/1802.07810** (2018), http://arxiv.org/abs/1802.07810
35. Prabhakaran, K., Dridi, J., Amayri, M., Bouguila, N.: Explainable k-means clustering for occupancy estimation. Procedia Comput. Sci. **203**(C), 326–333 (jan 2022). https://doi.org/10.1016/j.procs.2022.07.041, https://doi.org/10.1016/j.procs.2022.07.041
36. Rawal, K., Fu, Z., Delaney, E., Russell, C.: Evaluating model explanations without ground truth. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. p. 3400–3411. FAccT '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3715275.3732219, https://doi.org/10.1145/3715275.3732219
37. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead (2019), https://arxiv.org/abs/1811.10154
38. Sabbatini, F., Calegari, R.: Explainable clustering with cream. In: Proc. of the Conference on Principles of Knowledge Representation and Reasoning KR2023. pp. 593–603 (09 2023). https://doi.org/10.24963/kr.2023/58
39. Samuel, J., Khanna, T., Esguerra, J., Sundar, S., Pelaez, A., Bhuyan, S.: The rise of artificial intelligence phobia! unveiling news-driven spread of AI fear sentiment using ml, nlp, and llms. IEEE Access **13**, 125944–125969 (2025). https://doi.org/10.1109/ACCESS.2025.3588179, publisher Copyright: © 2013 IEEE.
40. Shah, R., Chircu, A.M.: IOT and AI in healthcare: A systematic literature review. Issues in Information Systems **19**(3), 33–41 (2018). https://doi.org/https://doi.org/10.48009/3_iis_2018_33-41
41. Soni, N., Sharma, E.K., Singh, N., Kapoor, A.: Artificial intelligence in business: From research and innovation to market deployment. Procedia Computer Science **167**, 2200–2210 (2020). https://doi.org/https://doi.org/10.1016/j.procs.2020.03.272, https://www.sciencedirect.com/science/article/pii/S1877050920307389, international Conference on Computational Intelligence and Data Science
42. Sovrano, F., Vitali, F.: An objective metric for explainable AI: how and why to estimate the degree of explainability. CoRR **abs/2109.05327** (2021), https://arxiv.org/abs/2109.05327
43. Soyarar, E., Aydogan, R., Buzcu, B., Calvaresi, D.: Llm-based evaluation methodology of explanation strategies. In: Calvaresi, D., Najjar, A., Omicini, A., Aydogan, R., Carli, R., Ciatto, G., Tiribelli, S., Främling, K. (eds.) Explainable, Trustworthy, and Responsible AI and Multi-Agent Systems. pp. 85–103. Springer Nature Switzerland, Cham (2026)
44. Starosta, B., Kłopotek, M.A., Wierzchoń, S.T.: Approaches to Explainability of output of graph spectral clustering methods. In: Mikułowski, D., Niewiadomski, A. (eds.) Design and Implementation of Artificial Intelligence Systems, pp. 5–31. Intelligent Systems and Information Technology, University of Siedlce (2025)
45. Starosta, B., Kłopotek, M.A., Wierzchoń, S.T., Czerski, D., Sydow, M., Borkowski, P.: Explainable graph spectral clustering of text documents. PLoS One **20(2):e0313238** (Feb 2025). https://doi.org/10.1371/journal.pone.0313238
46. Starosta, B., Wierzchoń, S.T., Borkowski, P., Czerski, D., Sydow, M., Laskowski, E., Kłopotek, M.A.: Rough sets for explainability of spectral graph clustering (2025), https://arxiv.org/abs/2512.12436

47. Starosta, B., Wierzchoń, S.T., Borkowski, P., Czerski, D., Sydow, M., Laskowski, E., Kłopotek, M.A.: Rough sets for explainability of spectral graph clustering (2026), https://arxiv.org/abs/2512.12436
48. Tamajka, M.: How to measure the quality of explanations of AI predictions (2022), https://kinit.sk/how-to-measure-the-quality-of-explanations-of-ai-predictions/
49. Tan, H.: Evaluating sensitivity consistency of explanations. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 182–191 (2025). https://doi.org/10.1109/WACV61041.2025.00028
50. Tu, J., Mei, G., Piccialli, F.: An improved nyström spectral graph clustering using k-core decomposition as a sampling strategy for large networks. Journal of King Saud University - Computer and Information Sciences **34**(6, Part B), 3673–3684 (2022). https://doi.org/https://doi.org/10.1016/j.jksuci.2022.04.009, https://www.sciencedirect.com/science/article/pii/S1319157822001379
51. Wierzchoń, S., Kłopotek, M.: Modern Clustering Algorithms, Studies in Big Data, vol. 34. Springer Verlag (2018)
52. Xu, Z., Ke, Y.: Effective and efficient spectral clustering on text and link data. In: CIKM '16: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 357—-366 (October 2016). https://doi.org/https://doi.org/10.1145/2983323.2983708
53. Zhao, Y., Liang, S., Ren, Z., Ma, J., Yilmaz, E., de Rijke, M.: Explainable user clustering in short text streams. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 155–164. SIGIR '16, Association for Computing Machinery, New York, NY, USA (2016), https://doi.org/10.1145/2911451.2911522
54. Zhou, J., Chen, F.E.: Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent. ; Human–Computer Interaction, Springer: Berlin/Heidelberg, Germany (2018)
55. Zhou, J., Sun, J., Chen, F., Wang, Y., Taib, R., Khawaji, A., Li, Z.: Measurable decision making with gsr and pupillary analysis for intelligent user interface. ACM Trans. Comput.-Hum. Interact. **21**, 33 (2015). https://doi.org/https://doi.org/10.1145/2687924
56. Zhou, J., Arshad, S.Z., Yu, K., Chen, F.: Correlation for user confidence in predictive decision making. In: Proceedings of the 28th Australian Conference on Computer-Human Interaction. p. 252–256. OzCHI '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/3010915.3011004
57. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5) (2021). https://doi.org/10.3390/electronics10050593
58. Zhou, J., Li, Z., Hu, H., Yu, K., Chen, F., Li, Z., Wang, Y.: Effects of influence on user trust in predictive decision making. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. p. 1–6. CHI EA '19, Association for Computing Machinery, New York, NY, USA (2019), https://doi.org/10.1145/3290607.3312962