Bartłomiej Starosta[1]
ORCID:0000-0002-5554-4596

Mieczysław A. Kłopotek[1]
ORCID:0000-0003-4685-7045

Sławomir T. Wierzchoń[1]
ORCID:0000-0001-8860-392X

Institute of Computer Science of Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland. barstar,klopotek,stw@ipipan.waw.pl

Towards Explainability of Hashtags in the Light of Graph Spectral Clustering Methods

DOI: 10.XXXXXX

Abstract. Hashtags constitute an indispensable part of modern social media world. As more and more hashtags are invented, it becomes a necessity to create clusters of these hashtags. Nowadays, however, the clustering alone does not help the users. They are asking for justification or expressed in the modern AI language, the clustering has to be explainable. We discuss a novel approach to hashtag explanation via a measure of similarity between hashtags based on the Graph Spectral Analysis. The application of this similarity measure may go far beyond the classical clustering task. It can be used to provide with explanations for the hashtags. In this paper we propose such a novel view of the proposed hashtag similarity measure.

Keywords: Graph Spectral Analysis, hashtag similarity, eigenvalue spectrograms, Explainable Artificial Intelligence

1 Introduction

In recent years, several word embedding methods like Doc2Vec [10] or BERT [6] have been proposed. The idea behind those methods is to represent words in a high dimensional space (embedding space)) taking into account the context in which the words occur. Small distance between two words in the embedding space is usually related to their semantic similarity. This property can be exploited under various settings. One example is an explanation of one word in terms of the word or words lying closely to it in the embedding space. If a classification algorithm is based on the representation in the embedding space, then the decision boundaries can be explained in terms of words close to these boundaries. By analogy, the same can be said about the results of clustering performed in the embedding space.

Over the last decades a kind of a new "human" language, the language of hashtags is under development, along with various processing techniques. One could be tempted to try to develop a similar way of understanding hashtags in terms of other hashtags. For example, one could take a collection of texts with hashtags, remove normal words, and then train aforementioned models on "text" consisting of hashtags alone. However, this is not that simple because generally, the hashtags occur very sparsely in e.g. tweets so that mentioned models like Doc2Vec or BERT are barely applicable.

Therefore, we look at another way of hashtag embedding. This embedding should allow to explain one hashtag in terms of other hashtags in a manner similar to word explanation based on e.g. Doc2Vec. By analogy, such an embedding could also be used to explain decision boundaries for classification and clustering methods in terms of hashtags. We demonstrate that Graph Spectral Analysis of documents labeled with hashtags can be a foundation for such an embedding.

Cluster Analysis, particularly Graph Spectral Clustering (GSC), like the entire domain of Artificial Intelligence, has experienced rapid development in recent years. Algorithms of growing complexity and efficiency were provided, but regrettably they are characterized by "black-box nature". Hence their results are hard to understand by human users. Therefore there exists a growing resistance to their application in practical settings. Business and industry requested justifications of decisions suggested by AI systems. This expectation of client of AI systems led to the development of a branch of AI called "Explainable Artificial Intelligence" (XAI) [2], with subbranches including Explainable Clustering [4].

The "black box" problem is more grievant in cluster analysis, compared e.g. to the classification tasks, because the very essence of the concept of "cluster" is not well defined. This is even though the scientific research area of cluster analysis, has nearly a century-long history. Hundreds of clustering algorithms have been developed and countless applications are reported.

Though there exist some approaches for cluster explanations like [9] that are applied to the outcome of spectral clustering, they are not based on the actual principles of spectral clustering, but rather on cluster approximation with some other algorithms. There exist explanation methods to components of spectral clustering, that is to k-means [12] but they are insufficient to explain the outcome of spectral clustering.

The so-called Graph Spectral Analysis (GSA)[1] stands for a novel way of looking into relationships between data objects that are characterized by mutual similarity measures, and hence can be best described by a graph with weights equal to those similarities. The GSA procedure consists of the following steps: First, the similarity matrix is transformed to a combinatorial or normalized Laplacian, which in turn is subject to eigen-decomposition. Eigenvectors constitute a new coordinate system into which the data objects are embedded and thus may be subject of distance-based *data clustering* [11, 18] or *data classification* methods [17, 14]. In particular, this approach is used with hashtags [13].

While the main stream of research concentrates on the usage of a carefully selected subset of eigenvalues and corresponding eigenvectors, the research in this paper took a different path. We investigate properties of the entire set of eigenvalues, as initiated by [1, 5].

This paper aims at an easy introduction to that approach, and based on it, we suggest a way of explaining hashtags by other hashtags which sheds a completely new light on the research reported in [15].

2 Brief introduction to Graph Spectral Clustering

Let us briefly recall that GSC methods are deemed to be a relaxation of cut-based graph clustering methods. A matrix S, called a similarity matrix, represents similarities between pairs of items (e.g. tweets), with entries ranging from 0 to 1. It induces a graph whose nodes correspond to the items. This graph is assumed to be without self-loops (even though an item is most similar to itself). Hence the diagonal of S is assumed to be filled

[1] GSA encompasses Graph Spectral Clustering (GSC) and Graph Spectral Classification (GSL).

with zeros (see e.g. [16] as one example of many). Let n denote the number of items for which S has been computed.

A combinatorial Laplacian L corresponding to the matrix S is defined as

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

where D is the diagonal matrix with $d_{jj} = \sum_{k=1}^{n} s_{jk}$ for each $j = 1, \ldots n$. A normalized Laplacian $\mathcal{L}$ of the graph represented by S is defined as

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

A random walk Laplacian $\mathbb{L}$ of a graph is defined as

$$\mathbb{L} = L D^{-1} = I - S D^{-1} \tag{3}$$

Other Laplacians were also studied [18].

Their relationship to cut based clustering is as follows: The RCut criterion corresponds to finding the partition matrix $P_{RCut} \in \mathbb{R}^{n \times k}$ that minimizes the formula $H'LH$ over the set of all partition matrices $H \in \mathbb{R}^{n \times k}$. Such formulated problem is NP-hard. That is why we relax it by assuming that H is a column orthogonal matrix. In this case the solution is obvious: the columns of P_{RCut} are eigenvectors of L corresponding to k smallest eigenvalues of L. Similarly, the columns of matrix P_{NCut}, representing NCut criterion, are eigenvectors of $\mathcal{L}$ corresponding to k smallest eigenvalues of $\mathcal{L}$. For an explanation and further details see e.g. [11] or [18].

3 Method

As reported in [5], spectra of combinatorial Laplacian of random samples of the same hashtag can be down-scaled to overlap, while those from different classes do not. The same applies also to so-called normalized Laplacians (see Fig. 1).

Consider a similarity matrix S between pairs of items (e.g. tweets). A combinatorial Laplacian L of S was given by the equation (1). Let its eigenvalues be sorted in non-decreasing order $0 = \lambda_1 \leq \cdots \leq \lambda_n$.

Let us recall the function $\lambda_{ssa} : [0, 1] \to \mathbb{R}$ such that [15]

$$\lambda_{ssa}\Big(\frac{n-i}{n-1}\Big) = \frac{\lambda_i}{n} \, . \tag{4}$$

The linear interpolation is applied in-between the points $\frac{n-i}{n-1}$ and $\frac{n-(i+1)}{n-1}$. λ_{ssa} is stable for the group of tweets corresponding to the same hashtag, while it differs for groups of tweets labeled with different hashtags. Fig. 1 shows λ_{ssa}'s for 10 samples of tweets with the same hashtag. Fig. 2 shows λ_{ssa}'s of 34 samples of distinct ones.

Hence, a "distance" between a given new sample and the elements of a class is defined as the area between the λ_{ssa} curves. So if the first subgraph $G1$ is characterized by $\lambda_{ssa,G1}$ curve, and the second subgraph $G2$ is characterized by $\lambda_{ssa,G2}$ curve, then the dissimilarity is computed as

$$dissim(\lambda_{ssa,G2}, \lambda_{ssa,G1}) = \int_0^1 |\lambda_{ssa,G2}(x) - \lambda_{ssa,G1}(x)| dx \, . \tag{5}$$

dissim may be successfully applied to classify a sample into one of the hashtags, as table 1 shows. For each hashtag 100 samples from 30% of its tweets were drawn and classification via the smallest dissimilarity to the hashtag spectra was performed. Reasonable classification was possible for up to 10 hashtags.

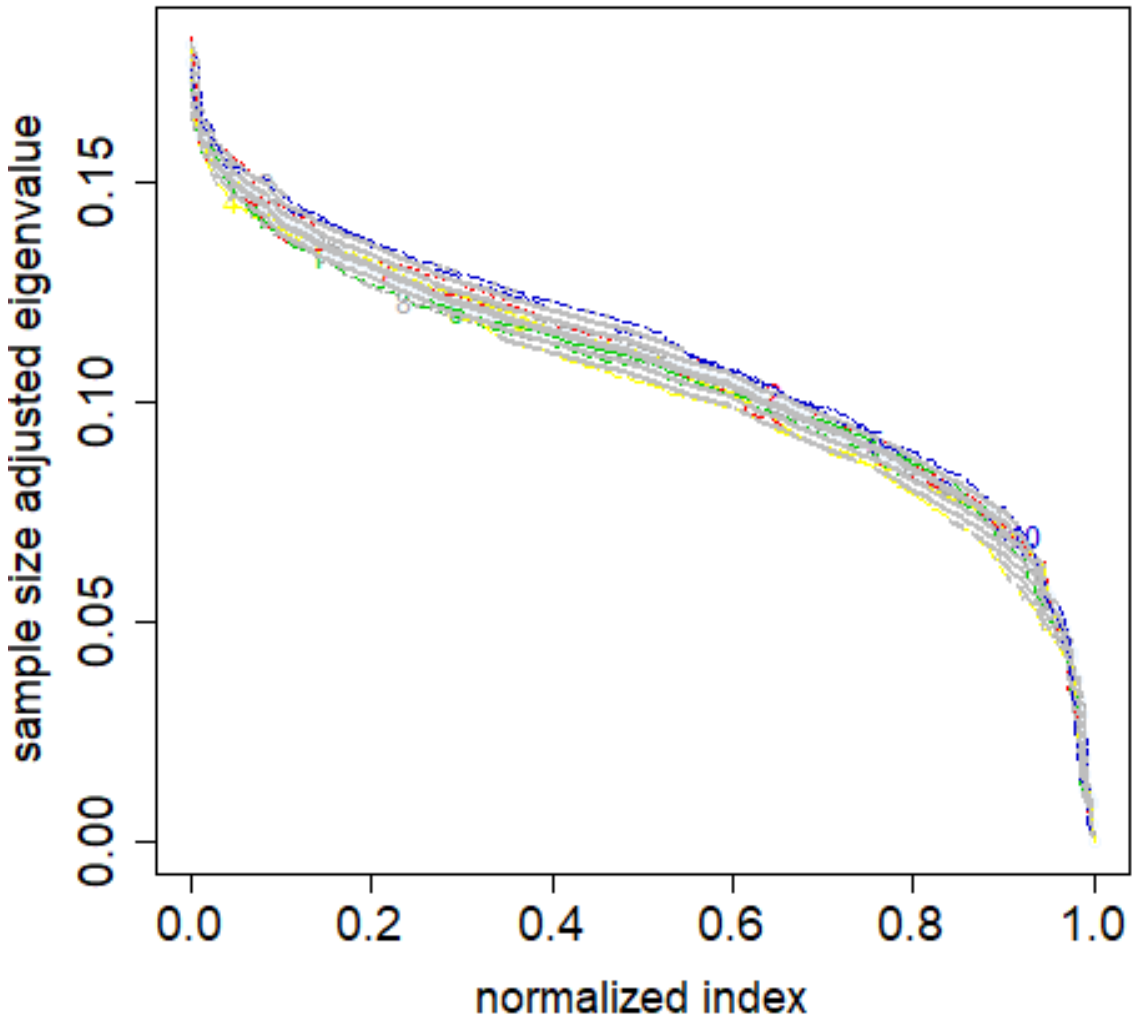

Figure 1. Normalized spectrograms for samples of one single hashtag. Source: research presented in [3]

Table 1. Classification errors and F1 measure for most distant hashtags. Source: research presented in [3]

no. of hashtags	2	3	4	5	6	7	8	9	10	11
error %	0.00	1.33	1.75	0.60	4.83	7.00	9.75	9.22	8.20	9.82
F1*100	100.00	98.67	98.25	99.40	95.18	93.03	89.99	90.53	91.71	90.03
no. of hashtags	12	13	14	15	16	17	18	19	20	21
error %	14.17	16.77	19.07	18.47	20.75	23.53	26.22	30.47	26.90	29.38
F1*100	85.74	83.18	80.86	81.00	78.90	76.04	73.61	69.37	72.89	70.60

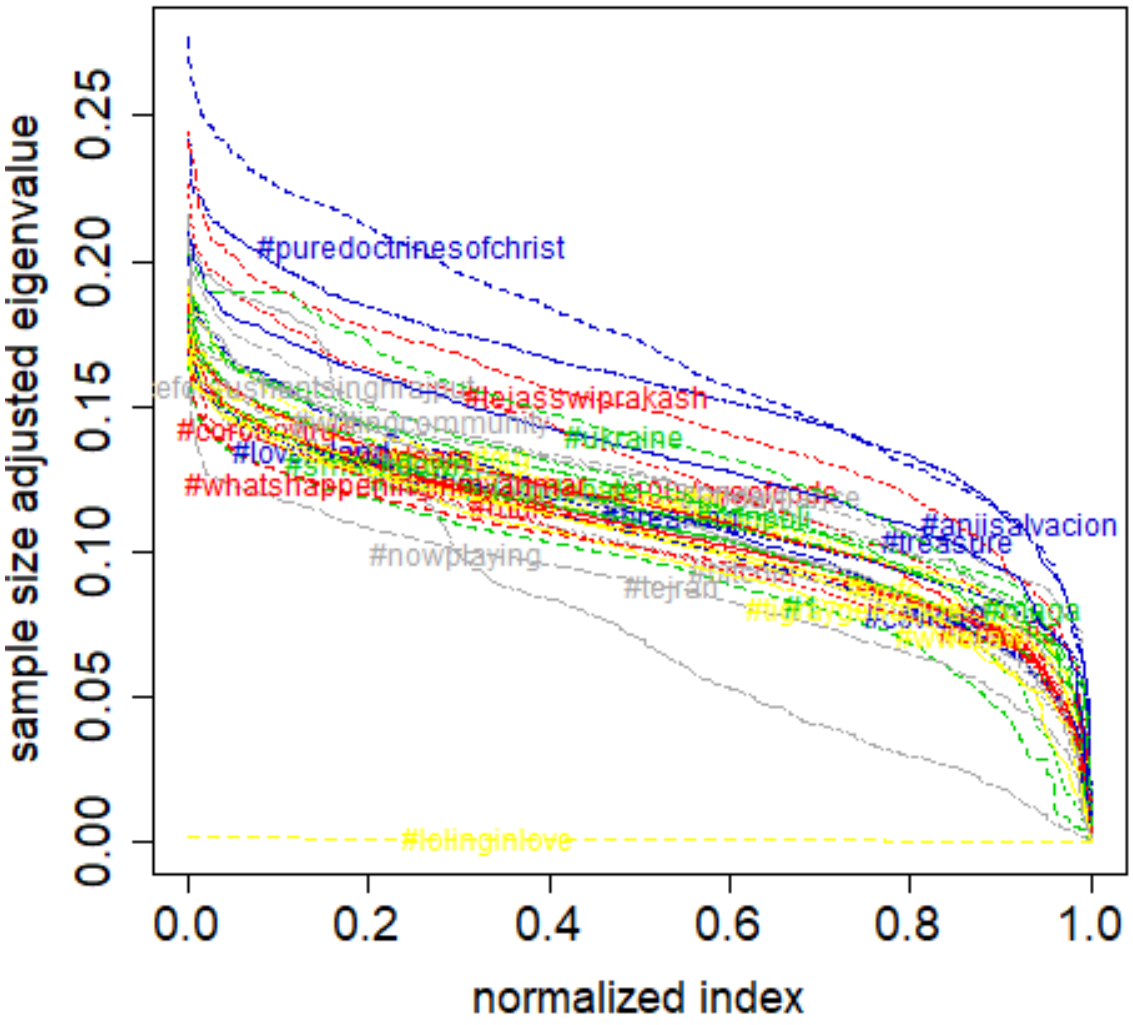

Figure 2. Normalized spectrograms for samples of various hashtags. Source: research presented in [3]

4 Results

Table 2. Closeness of hashtags based on eigenvalue spectrum. Source: research presented in [3]

hashtag	s.hashtag	min.dist	avg.dist	std.dist	subs.dist	subs.err	rel.subs.dist
#1	#tejran	0.0080	0.0275	0.0201	0.0037	0.0022	0.1342
#100days ofcode	#treasure	0.0069	0.0258	0.0210	0.0054	0.0028	0.2128
#90dayfiance	#maga	0.0051	0.0230	0.0211	0.0036	0.0014	0.1599
#aewdynamite	#demdebate	0.0015	0.0177	0.0213	0.0032	0.0013	0.1838
#anjisalvacion	#tejass wiprakash	0.0108	0.0481	0.0244	0.0034	0.0018	0.0706
#auspol	#coronavirus	0.0011	0.0175	0.0216	0.0029	0.0021	0.1655
#bbnaija	#whatshap peningin myanmar	0.0044	0.0217	0.0208	0.0015	0.0008	0.0716
#bitcoin	#whatshap peningin myanmar	0.0019	0.0204	0.0213	0.0023	0.0013	0.1162
#blacklives matter	#demdebate	0.0018	0.0179	0.0213	0.0020	0.0012	0.1171
#breaking	#justicefor sushantsinghrajput	0.0016	0.0180	0.0209	0.0051	0.0039	0.2827
#cdnpoli	#covid_19	0.0024	0.0190	0.0212	0.0039	0.0025	0.2052
#coronavirus	#auspol	0.0011	0.0176	0.0216	0.0015	0.0005	0.0876
#covid	#coronavirus	0.0018	0.0175	0.0215	0.0051	0.0022	0.2950
#covid_19	#cdnpoli	0.0024	0.0182	0.0212	0.0058	0.0040	0.3183
#covid19	#wweraw	0.0016	0.0186	0.0216	0.0020	0.0009	0.1092
#demdebate	#aewdynamite	0.0015	0.0180	0.0213	0.0026	0.0012	0.1463
#endsars	#blacklives matter	0.0026	0.0175	0.0213	0.0031	0.0015	0.1770
#justicefor sushantsinghrajput	#breaking	0.0016	0.0180	0.0209	0.0043	0.0021	0.2396
#lolinginlove	#nowplaying	0.0809	0.1135	0.0186	0.0003	1.1618e-05	0.0033
#loveisland	#demdebate	0.0038	0.0184	0.0211	0.0019	0.0009	0.1032
#maga	#cdnpoli	0.0040	0.0207	0.0212	0.0065	0.0038	0.3135
#mufc	#coronavirus	0.0015	0.0180	0.0216	0.0023	0.0006	0.1284
#nowplaying	#1	0.0301	0.0453	0.0137	0.0104	0.0042	0.2294
#nufc	#mufc	0.0016	0.0183	0.0215	0.0039	0.0008	0.2148
#puredoctrines ofchrist	#anjisalvacion	0.0133	0.0602	0.0250	0.0040	0.0016	0.0679
#smackdown	#auspol	0.0026	0.0186	0.0213	0.0040	0.0014	0.2149
#tejass wiprakash	#ukraine	0.0093	0.0385	0.0232	0.0048	0.0028	0.1268
#tejran	#1	0.0080	0.0320	0.0193	0.0039	0.0015	0.1226
#tigraygenocide	#bitcoin	0.0035	0.0214	0.0209	0.0043	0.0023	0.2039
#treasure	#ukraine	0.0057	0.0278	0.0217	0.0030	0.0009	0.1103
#ukraine	#treasure	0.0057	0.0309	0.0221	0.0054	0.0019	0.1760
#whatshap peningin myanmar	#bitcoin	0.0019	0.0205	0.0213	0.0031	0.0010	0.1530
#writing community	#maga	0.0042	0.0222	0.0209	0.0056	0.0032	0.2558
#wweraw	#covid19	0.0016	0.0194	0.0215	0.0057	0.0035	0.2939

We performed our research for a small collection of hashtags extracted from Twitter tweets. Their names are listed in the first column of Table 2. It is a random sample of about 1% of English tweets retrieved from the stream endpoint of Twitter API.

Table 2 gives an overview of stability of hashtag dissimilarities. For each hashtag from the column `hashtag`, the column `s.hashtag` represents the closest hashtag. Their closeness is characterized by the subsequent columns:

1. `min.dist` being the dissimilarity to it.

2. `avg.dist` presents the average dissimilarity of the given hashtag from the remaining ones,
3. `std.dist` shows the standard deviation of dissimilarity.
4. `subsamp.dist` represents the average dissimilarity to 100 samples from the same hashtag
5. `subsamp.err` being the standard deviation of this measure
6. `rel.subsamp.dist` is the quotient of `subsamp.dist` / `avg.dist`.

`rel.subsamp.dist` demonstrates that in fact the samples from the same hashtag are closer to one another than to other hashtags.

The hashtag *#lolinginlove* seems to be most distant from all the other hashtags on average, while *#blacklivesmatter* seems to be close to many other hashtags from the list, in particular demonstration related hashtags. The hashtag *#puredoctrinesofchrist* seems also to be distant from the other, though it is quite near to *#anjisalvacion*. *#covid* has a characteristic quite similar to *#coronavirus* which should not be surprising.

The average dissimilarity to other hashtags is 5 to 10 time bigger than the sampling variation within a single hashtag.

A careful look at the hashtag similarities reveals a potential behind it that has not been explored before. Note that from a human point of view, the examples demonstrate that one hashtag explains the meaning of the other. Therefore, in general, hashtag similarity can be used to explain a hashtag in terms of other hashtags. For example, *#blacklivesmatter* may be explained as a term with commonalities with *#demdebate* and *#endsars*. *#whatshappeninginmyanmar* on the other hand seems to be related to *#bitcoin*, *#bbnaija*.

This way of looking at explaining hashtags has a striking new feature: As in this study we used tweets with only single hashtags, the obtained similarity measure is not based on the co-occurrence of hashtags, but rather on their contextual occurrence. So a new meaning to the explanation of one hashtag by other ones is invented. They are on a higher conceptual level. This is in contrast to the traditional way of explaining the contents of a cluster by the most frequent words or phrases in the textual content. Here, to explain one hashtag, we use another hashtag that was not present in the texts attached to the previous one. It differs also from the word embedding approaches like Doc2Vec or BERT because the mentioned embeddings assign similar vectors to words occurring close to one another in text, while the hashtags in our embedding do not need to occur and usually do not occur in the same text.

Still, another usage of the explanatory power of this new similarity measure can be an explanation of the entire set of documents. One can for example seek the most distinct hashtags and present them as a brief explanation of the entire document set. One can proceed as follows: The first explaining hashtag would be one with the highest sum of dissimilarities to other hashtags. The other can be selected based on the highest sum of dissimilarities to those already chosen. The following list of 5 best explaining hashtags was obtained in this way for this collection: *#lolinginlove*, *#puredoctrinesofchrist*, *#anjisalvacion*, *#nowplaying*, *#tejran*.

5 Conclusions

We have proposed to provide explanations for topical groups of objects, like tweets, via a characteristic spectrum of combinatorial Laplacian. It appears to be quite a stable descriptor of samples from the same population, while discriminating different populations. With this proposal, we go beyond the similarity measurement of hashtags, suggested in [15].

In particular, a group of documents sharing the same hashtag can be characterized by a combinatorial Laplacian spectrogram. In this paper, we have demonstrated that such spectrograms can be used to look for similar or dissimilar hashtags. In this sense such a spectrogram can be seen as a special vehicle for hashtag embeddings in hashtag processing, playing similar role as word embeddings related to Doc2Vec or BERT like methods in natural language processing. It has to be stressed that this method of hashtag embedding constitutes a novelty compared to the mentioned and other conventional methods of word embedding. While the conventional methods are focusing on word co-occurrence, our method does not require hashtag co-occurrence for embedding computation.

Potential applications seem to be as ingredients in explainable classification and clustering tasks as well as explainable data visualization and hashtag recommendation, that was investigated in [8, 7].

We investigated the possibility of a novel way of hashtag embedding so that explainability of hashtags, of classification and clustering results in terms of other hashtags can be achieved. We demonstrated that Graph Spectral Analysis of documents labelled with hashtags can be a foundation for such an embedding. In particular, in the context of hashtag recommendation, the explainability of hashtags in terms of other hashtags is of interest. Note that in this study we used tweets with only single hashtags so that the similarity measure is not based on the co-occurrence of hashtags, but rather on their contextual occurrence. This attaches a new meaning to the explanation of one hashtag by other ones. They are on a higher conceptual level. One can think that if in a period of time a group of people, inducing a group style of writing, is interested in several hashtags, then it is to be assumed that these hashtags have something in common. This constitutes the foundation for the reasonability of explanation by our method.

The explainable characterization of hashtags suggested here differs from the main stream Graph Spectral Analysis performed so far as standard GSA concentrated on selected eigenvalues and eigenvectors.

Further research is necessary as to what causes this spectral behavior for similar and different hashtags.

References

1. Alzate, C., Suykens, J.A.: Multiway spectral clustering with out-of-sample extensions through weighted kernel pca. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(2), 335 – 347 (February 2010)
2. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82 – 115 (2020)
3. B. Starosta, M.A. Kłopotek, S.W.: Hashtag similarity based on laplacian eigenvalue spectrum. to appear in proc. pp-rai'2023 - 4th polish conference on artificial intelligence , Łódź, poland 2023.
4. Bandyapadhyay, S., Fomin, F.V., Golovach, P.A., Lochet, W., Purohit, N., Simonov, K.: How to find a good explanation for clustering? (2021). https://doi.org/10.48550/ARXIV.2112.06580
5. Borkowski, P., Kłopotek, M.A., Starosta, B., Wierzchoń, S.T., Sydow, M.: Eigenvalue based spectral classification. PLoS ONE **18**(4), e0283413 (2023). https://doi.org/https://doi.org/10.1371/journal.pone.0283413
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019). https://doi.org/10.48550/ARXIV.1810.04805
7. Gupta, V., Hewett, R.: Unleashing the power of hashtags in tweet analytics with distributed framework on Apache Storm (2018). https://doi.org/10.48550/ARXIV.1812.01141
8. Jeon, M., Jun, S., Hwang, E.: Hashtag recommendation based on user tweet and hashtag classification on twitter. In: Web-Age Information Management. pp. 325–336. Springer International Publishing, Cham (2014)
9. Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W., Muller, K.R.: From clustering to cluster explanations via neural networks. IEEE Transactions on Neural Networks and Learning Systems pp. 1–15 (2022). https://doi.org/10.1109/tnnls.2022.3185901, `https://doi.org/10.1109%2Ftnnls.2022.3185901`
10. Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation (2016). https://doi.org/10.48550/ARXIV.1607.05368

11. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007)
12. 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
13. Schmidt, A., et al.: Using spectral clustering of hashtag adoptions to find interest-based communities. In: 2018 IEEE International Conference on Communications (ICC). pp. 1–7 (2018). https://doi.org/10.1109/ICC.2018.8422244
14. Sevi, H., Jonckheere, M., Kalogeratos, A.: Generalized spectral clustering for directed and undirected graphs (2022). https://doi.org/10.48550/ARXIV.2203.03221
15. Starosta, B., Kłopotek, M., Wierzchoń, S., Czerski, D.: Hashtag discernability – competitiveness study of graph spectral and other clustering methods. In: accepted for the 18th Conference on Computer Science and Intelligence Systems FedCSIS 2023 (IEEE #57573) Warsaw, Poland, 17–20 September, 2023 (2023)
16. 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)
17. Tu, J., Mei, G., Picciallib, 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 (2022)
18. Wierzchoń, S., Kłopotek, M.: Modern Clustering Algorithms, Studies in Big Data, vol. 34. Springer Verlag (2018)