Robustness and sensitivity of network-based topic detection

Carla Galluccio[1], Matteo Magnani[2], Davide Vega[2], Giancarlo Ragozini[3], and Alessandra Petrucci[1]

[1] University of Florence, Florence FI 50134, Italy, carla.galluccio@unifi.it
[2] InfoLab, Uppsala University, 752 37 Uppsala, Sweden
[3] University of Naples Federico II, Naples NA 80133, Italy

Abstract. In the context of textual analysis, network-based procedures for topic detection are gaining attention as an alternative to classical topic models. Network-based procedures are based on the idea that documents can be represented as word co-occurrence networks, where topics are defined as groups of strongly connected words. Although many works have used network-based procedures for topic detection, there is a lack of systematic analysis of how different design choices, such as the building of the word co-occurrence matrix and the selection of the community detection algorithm, affect the final results in terms of detected topics. In this work, we present the results obtained by analysing a widely used corpus of news articles, showing how and to what extent the choices made during the design phase affect the results.

Keywords: Text network analysis, community detection, topic detection

1 Introduction

The need to gather information from large textual datasets has led to the development of automated information extraction methods [12,18]. Among these methods, those aimed at identifying topics have become very popular in machine learning and natural language processing [1].

Recently, network-based procedures have gained attention in the context of textual analysis as an alternative to classical topic models for detecting topics in large collections of documents [9]. These methods are based on the idea that any text can be represented as a word co-occurrence network, where topics emerge as groups of strongly connected words. In addition, the network can be used to explore and present the relations between the topics. Although many works have used network-based procedures for detecting topics in textual data, there is a lack of systematic analysis of how different design choices affect the final results in terms of detected topics.

Essentially, a network-based topic discovery process takes the following form:

- pre-processing the text, a step-by-step procedure during which the researcher selects which methods to apply to clean the text and make it ready for the

analysis (e.g. removal of non-alphanumeric characters, removal of stopwords, reduction of terms to a common root);
- forming of the word co-occurrence matrix by defining the context in which two words will be considered semantically related. This is usually done by defining what is meant by "co-occurrence" between words;
- building of the network and selection of the community detection algorithm.

This procedure requires the researcher to make decisions in each of these steps.

In this work, we focus on the two defining steps of this process, as they are unique to network-based approaches: building the word co-occurrence matrix and selecting the community detection algorithm. From our point of view, the definition of the word co-occurrence matrix, which determines the shape of the network, and the community detection algorithm employed are strongly related to the characteristics of the discovered topics. Moreover, the impact of other design choices on text classification has already been studied in a non-network context. For instance, Uysal and Gunal have investigated the impact of text pre-preprocessing on text classification, revealing that choosing an appropriate combination of pre-processing steps may improve the classification accuracy [17].

As an example, Figure 1 shows four different networks built using the same documents. They represent the word co-occurrence matrices of 9 news extracted from the BBC news articles collection [8] concerning business, sport, and tech. More specifically, in the first (Figure 1a) and the third (Figure 1c) networks two words belonging to the same document are adjacent, or co-occur, if they are at most 2 words apart (that is, if between the two words there is at most one word in between). On the other hand, the second (Figure 1b) and the fourth (Figure 1d) networks have been built considering that two words in the same document co-occur if they are at most 10 words apart. Furthermore, in order to identify the topics, we applied the Louvain community detection algorithm [4] on the first and the second networks (Figure 1a and Figure 1b), while on the other two networks we applied Newman's leading eigenvector method for detecting communities [13]. It is possible to observe how the shape of the networks and the detected communities change. For example, we can observe more defined communities in the networks with a window size equal to 10, some communities recognised by one method are split into two by the other, and some nodes are assigned to a different community.

Analysing the effect of the relevant design choices on the final results allows us to identify the fundamental aspects that should be taken into account when using network-based procedures to analyse textual data and discover topics, and those which may require further research. Therefore, the main contribution of this work is to evaluate the relationship between the shape of the network, which changes depending on the word co-occurrence matrix, the community detection algorithm employed, and the features of the discovered topics.

Another unexplored question about network-based topic detection is about its relationship with probabilistic topic models, such as Latent Dirichlet Allocation (LDA) [3]. While this question is also important, before addressing it we need to develop a deeper understanding of optimal design choices for network-based

methods. Therefore, this paper is a first step towards enabling a comparison between these different approaches.

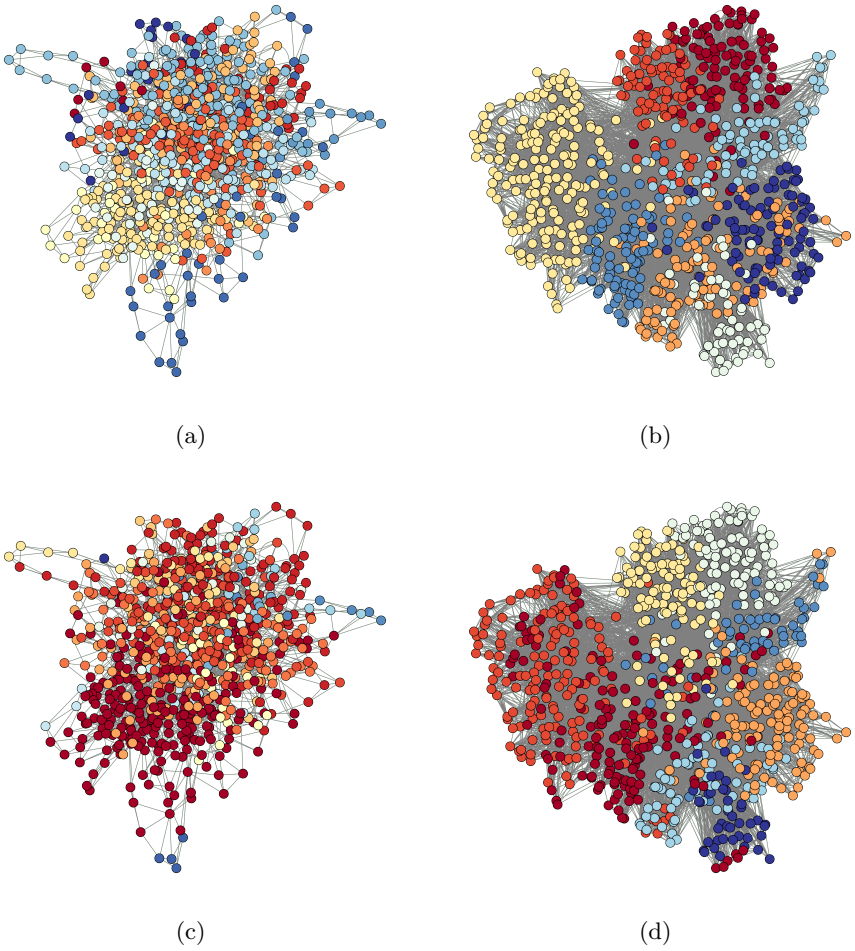

Fig. 1. Example of networks obtained from 9 news of the BBC news article collection. In networks (a) and (c) the window size is equal to 2, while it is equal to 10 in networks (b) and (d). The colours represent the community to which each node belongs according to community detection algorithms: the Louvain algorithm in (a) and (b) and Newman's leading eigenvector method in (c) and (d). Note that the organization of nodes in communities varies between networks. Indeed, while in (b) and (d) the organization in communities is clear, in (a) and (c) the partition is much less defined.

2 State of the art

In recent years, many works have been written about applying community detection methods for topic discovery.

For example, Sayyadi and Raschid find topics as communities in a keyword co-occurrence matrix using the Girvan-Newman community detection algorithm based on the betweenness centrality measure [16]. They build the keyword co-occurrence matrix considering that two keywords are connected if they co-occur in at least one document, and the weight of that link is given by the number of documents in which both keywords co-occur. Then, they compute each word's document frequency and remove the links with a value below a specific threshold.

Another example is given by Salerno *et al.*, who apply the Louvain community detection algorithm for discovering topics on a weighted network in which nodes represent individual words in the vocabulary and links indicate the co-occurrence of a pair of words within a document [15]. The weight of the links between words is determined by the context in which two words co-occur: for example, a co-occurrence within the same sentence carries more weight than a co-occurrence within the same paragraph. Then, they evaluate their results using modularity and comparing the error rate to the results achieved by two baselines: one that classifies documents randomly and another one that classifies documents based on the most common label in the training set. Similar approaches can be found in Dang and Nguyen [6].

Instead, de Arruda *et al.* investigate how specific definitions of the co-occurrence between words favour the emergence of communities of semantically related words, allowing for the identification of relevant topics [7]. In particular, they consider three different ways to define the co-occurrence between two words in the pre-processed text: two words are connected if they are separated by at most a given number of other words; words belonging to the same paragraph are linked together in a clique, disregarding links between words further from each other than the given maximum distance; finally, the statistical significance of co-occurrences with regard to random, shuffled texts is tested. The fast-greedy method is used to find communities of high modularity.

Lancichinetti *et al.* discover topics using the Infomap algorithm on networks built considering that two words are connected if they co-occur in the same document [12]. More specifically, they compute the dot product similarity of each pair of words that co-occur in at least one document in order to compare it against the expectation for a null model where words are randomly shuffled across documents. Then, a threshold is defined for retaining words for which the co-occurrence between them cannot be explained by the null model. However, because Informap is run as a non-overlapping community detection algorithm, to cope with generic words used in multiple topics, they refine the results obtained from applying the community detection algorithm using a latent topic model that allows for non exclusivity.

Some of the most recent contributions in this area are given by Kim and Sayama [11] and Hamm and Odrowski [9]. The former transform the textual data into a vector form by computing the *tf-idf* (term frequency inverse document

frequency) score considering each sentence as a document. Afterwards, they compute the pair-wise cosine similarity of the *tf-idf* vectors to build adjacency matrices of the sentences, and then they use the Louvain community detection algorithm on the sentence networks, where the nodes are the sentences, and the cosine similarity of *tf-idf* representations between every node pair represents the link weight. Hamm and Odrowski apply the Leiden community detection algorithm on undirected weighted networks investigating the effects of the resolution parameter on modularity maximisation [9]. Moreover, they define a measure to identify the most significant words within a topic.

This work contributes to this research line by considering the relationship between the definition of the word co-occurrence matrix, the selection of the community detection algorithms, and the final results.

3 Method and material

In this section, we describe the data and the tested design choices.

Data. For the analysis, we used the corpus of BBC news articles, a collection of documents widely used as a benchmark for machine learning research [8]. The collection is composed of 2,225 complete news articles collected from 2004 to 2005 and divided into five topics: business, entertainment, politics, sport, and tech. The total number of articles and unique words per topic is reported in Table 1. We considered both the headline and the body of each news in the analysis.

Table 1. Number of documents and unique words for each topic of the BBC collection.

Topics	Documents	Unique words
Business	510	10,790
Entertainment	386	11,040
Politics	417	10,636
Sport	511	9,997
Tech	401	11,444

Data pre-processing. We removed non-alphanumeric characters, numbers, and words composed of 1 or 2 characters. Afterwards, we divided the text into tokens, choosing single words (*uni-grams*) as unit of analysis. Then, we removed the stopwords using a list provided with the dataset, and stemmed the text in order to reduce the size of the vocabulary, that is the set of unique words used in the text corpus. Finally, to remove very common words not included in the stopword list, we filtered out words with a value of *tf-idf* less than 0.01 [2]. After the pre-processing stage, the number of unique word tokens was equal to 18,422.

Word co-occurrence matrix. Once we pre-processed the corpus and obtained the vocabulary, we built the word co-occurrence matrices. To generate the word co-occurrence matrices we counted the number of times two words co-occur in the same document within a specific window size.

There are three ways of positioning the window: to the left of the word, to the right, or on either side [5]. Herein, we considered windows of different sizes placed to the right of the words, as usually done in the literature. More specifically, in this work we have considered window sizes equal to 2, 5, 10, 15 and 20.

Furthermore, in the literature different authors apply different filters to the word co-occurrence matrix based on the distribution of the words or their frequency in order to reduce the size of the matrix. For this reason, we decided to test this aspect by using different filters for the word co-occurrence matrices. More specifically, we removed the 100, 500, and 1000 words with the lowest co-occurrence values and the 50, 100, and 500 words with the highest co-occurrence values. We also filtered words with the highest or lowest co-occurrence values considering specific percentages of the total, but the results were similar to those obtained in the first two cases, so we do not report them here.

Afterwards, inspired by Salerno *et al.*, who applied different weights based on the context in which two words co-occur [15], we defined an experimental condition by modifying the co-occurrence values assigned to words within the window size. More specifically, we assigned weights proportional to the words' proximity. For example, for a window size equal to 3, the word adjacent to the target word gets a value equal to 1; the next word takes a value equal to 2/3; then, we assign a value equal to 1/3 to the last word.

Network and community detection algorithm. Starting from the word co-occurrence matrices, interpreted as weighted adjacency matrices, we built the undirected weighted networks on which we applied three different community detection algorithms.

Since almost all the works reported in Section 2 applied modularity optimisation algorithms, we decided to use the Louvain community detection algorithm as one of the most popular among them. Then, to investigate the performance of a different kind of approach we employed a spectral algorithm, namely Newman's leading eigenvector method. The rationale behind this choice is that if the network obtained after the pre-processing phase presents clearly separated topics, different algorithms should find similar results, while for networks with a less clear community structure the specific types of community that each different method is designed to identify would potentially lead to significantly different results. Finally, we argue that despite the absence of methods finding overlapping communities in the literature on network-based topic detection, in theory these methods are the most appropriate. In general, we cannot exclude that a word belongs to multiple topics at the same time, but using a partitioning method prevents the identification of such cases. As a consequence, we also tested the SLPA algorithm as a method designed to discover overlapping community [19].

4 Results and discussion

In this section, we present the results of our experiments, focusing on how the different choices we made in the definition of the word co-occurrence matrix and the selection of the community detection algorithm affect the features of the detected topics.

4.1 The effect of the window size

The main result we observe is that the number of communities obtained by the three algorithms is generally higher for smaller window sizes. Indeed, as the window size increases, the number of communities the algorithms find decreases, remaining constant for a window size greater than 5.

Figure 2 shows the number of communities found applying the three algorithms on the word co-occurrence matrices without filters: here, the number of communities identified by the non-overlapping community detection algorithms, that is, the Louvain and Newman's leading eigenvector methods, is always greater than the number of communities identified by SLPA for window sizes greater than 2. In particular, SLPA finds only one community with these settings.

4.2 Filters on the word co-occurrence matrix

The results remain stable when we remove the words with the lowest co-occurrence values from the word co-occurrence matrix. Instead, removing the words with the highest co-occurrence values changes the number of detected communities only for a window size equal to 2: the Louvain community detection algorithm found 47 communities, Newman's algorithm found 27 communities, while the SLPA found 112 communities. The results for window sizes greater than 2 remain stable.

4.3 Weighting scheme

Finally, we assessed the effect of using a different weighting scheme within the window sizes. We evaluated this aspect in the condition without any filters on the word co-occurrence matrix. In this case, results were significantly different from those obtained in the other experimental conditions for the Louvain and the SLPA community detection algorithms, with a number of communities ranging from 10 to 51 for the former and from 30 to 179 for the latter. However, also in this case the number of communities decreases when we increase the window size.

4.4 Selection of the community detection algorithm

Regarding the community detection algorithm, the Louvain algorithm showed the most interesting results. In almost all the experimental conditions, this algorithm found a number of communities equal to the number of the actual topics in

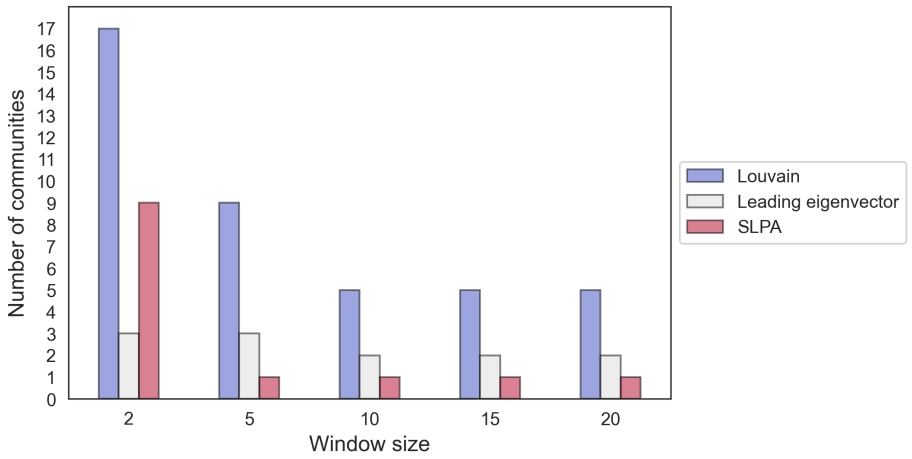

Fig. 2. Number of communities per window size and community detection algorithm. Observe that the number of communities decreases as the window size increases.

the document collection for window sizes greater than 5. Moreover, as shown in Figures 3(a) – 3(c), the communities are coherent with the content of the actual topics in the BBC document collection, with each community representing mainly one topic.

Note that Figure 3 was built by matching the communities' words with the actual topics' words, enabling possible overlapping. Therefore, in the representation of the correspondence between communities' words and topics' words, generic words such as "month" or "show" could be included in more than one topic.

To better understand these results, take as an example the communities found by the Louvain community detection algorithm for a window size equal to 10 (Figure 3a). First, the size of communities is quite balanced, with a number of words ranging from 3162 to 4283. Then, from an inspection of the words with the highest node degree within each community, we observed that they are coherent with the topic they represent. So, for example, among the top 15 words with the highest node degree in the first community there are words such as "show", "film", "record", "star", and "music", coherent with the topic "entertainment". We observed the same for window sizes equal to 15 and 20.

Instead, in the cases in which the Louvain algorithm finds more than 5 communities, namely for window sizes equal to 2 and 5, we observed that there are always 5 bigger communities coherent with the original topics and a variable number of smaller communities. Moreover, the largest communities generally include a number of words greater than 2000, whereas the smallest are composed of hundreds, tens, or just a few words.

To provide a more detailed analysis of the communities identified by the Louvain algorithm under different settings, we computed the Adjusted Rand Index (ARI) [10], a metric for comparing disjoint clustering solutions. Table 2

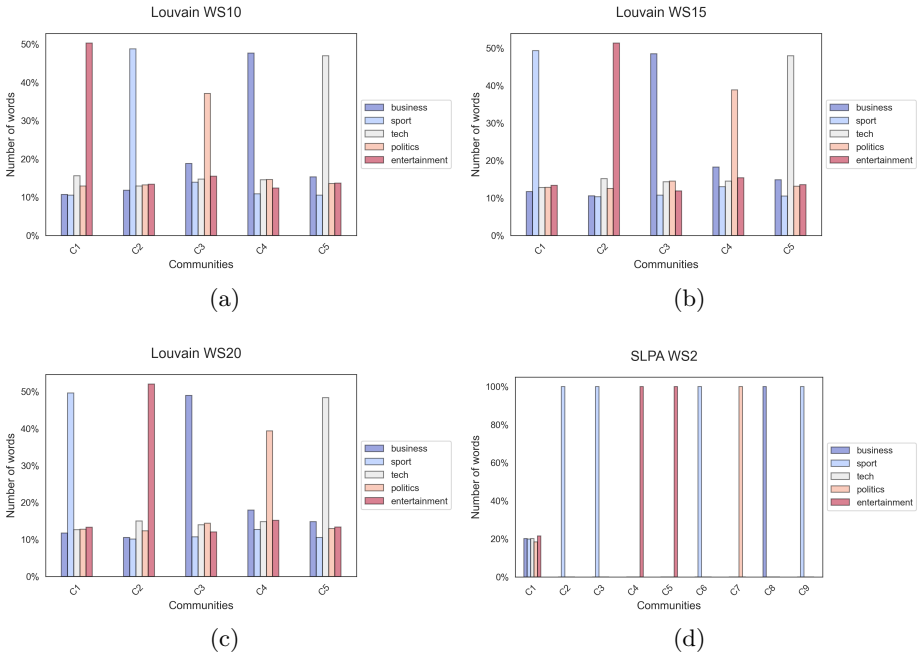

Fig. 3. Matching of the communities' words to actual topics' words for the Louvain community detection algorithm for window sizes equal to 10 (a), 15 (b), and 20 (c), and for the SLPA algorithm for window size equal to 2 (d). In this case, no filter was applied to the word co-occurrence matrix. On the y axis we reported the percentage of words in each community belonging to each actual topic (every group of bars sum up to 100%). "C" means "community", while "WS" means "window size".

shows the ARI for different window sizes. Observe that the ARI is generally high, particularly between the partitions obtained considering window sizes greater than 5. More specifically, for window sizes greater than 5, ARI values range from 0.604 to 0.878, showing high similarities, but also that the algorithm finds the same number of communities but the communities are not identical.

The lowest ARI values are associated to the partitions obtained using smaller window sizes, requiring an additional analysis to show how these communities relate to those found with larger window sizes. Therefore, we computed the contingency table between the partitions obtained with window sizes equal to 5 and 10, respectively, to better understand the tendency of the algorithm to merge communities related to the same topic by increasing the window size. The table is not reported here for space reasons, but it shows that some of (but not all) the clusters obtained using a window size equal to 5 are assimilated into some of the larger clusters found in the partition obtained using a window size equal to 10.

The two other algorithms failed to find a reasonable number of communities, with the SLPA algorithm finding only one community for window sizes greater

Table 2. Value of the ARI computed between all the partitions obtained by the Louvain community detection algorithm applied to networks built from the different word co-occurrence matrices. Here, "WS" means "window size".

	WS2	WS5	WS10	WS15	WS20
WS2	1				
WS5	0.348	1			
WS10	0.289	0.651	1		
WS15	0.279	0.624	0.828	1	
WS20	0.277	0.604	0.793	0.878	1

than 2 in all the experiments. Even in those cases where SLPA finds more than one community, the communities are not balanced, with almost all the words within one of the detected communities. Figure 3(d) shows the results we obtained applying the SLPA algorithm on the word co-occurrence matrix without filters using a window size equal to 2. Note that in the first community there are 18,402 words, while in the others the number of words ranges from 1 to 5. As an overlapping community detection algorithm, we also tried to use the K-clique algorithm [14] with different values for the k parameter, but we did not manage to obtain results because of the presence of large dense subgraphs, making this approach computationally intractable.

5 Conclusions

In this work we assess the effect of different design choices in network-based procedures for topic detection. In particular, we tested different ways of building the word co-occurrence matrix found in the literature and the selection of different community detection algorithms.

Our findings show that, for all tested algorithms, increasing the window size initially decreases the number of communities, which becomes stable for window sizes equal to or greater than 5 depending on the algorithm. This suggests that some of topics identified in the literature may have been influenced by this design choice, and leads to the consideration that the window size should be regarded as an important hyperparameter in future studies.

In addition, considering the number of detected topics applying different filters on the word co-occurrence matrix, we observe that the Louvain community detection algorithm generally performs better than the other tested algorithms. Indeed, considering the information available on the actual number of topics in the BBC document collection, the Louvain algorithm always detects the correct number of topics for a window size greater than 5, whereas the other two algorithms fail. This does not lead to a rejection of our hypothesis that overlapping community detection methods are more appropriate to find topics in word co-occurrence networks: it is still possible that the Louvain algorithm could

correctly cluster together words belonging to a single topic, while arbitrarily including multi-topic words in only one of the communities where they should have been included. However, we can conclude that some of the typical overlapping community detection methods are not able to identify significant topics under the experimental settings tested in this paper. The fact that these settings are taken from the literature suggests that more research should be done to identify pre-processing schemes leading to networks better suited to the application of these methods. One feature of the networks obtained in our experiments that may have determined the poor results of the tested methods is their high density, suggesting that stronger filtering schemes should be considered.

Finally, regarding the weighting scheme, our results show that while weighting the links can significantly affect the results, finding a good setting is not straightforward, with the number of communities suddenly becoming very high after imposing the basic scheme considered in this paper. This shows that this aspect should be analysed in more depth, also testing different combinations of pre-processing steps to select the words and to define co-occurrence weights and values.

In summary, on the one hand our preliminary results confirm what is stated in the literature, where network-based procedures for topic discovery show promising results; on the other hand, they highlight how different design choices, such as choosing specific algorithms or window sizes, applying filters on the word co-occurrence matrix, or defining different weighting schemes, may significantly affect the results in terms of detected topics.

Most importantly, this study highlights a number of aspects deserving additional attention. First, as further developments, we plan to extend our study considering additional community detection algorithms, to evaluate which methods are appropriate depending on the applied pre-processing steps. Second, additional ways to define the word co-occurrence matrix should also be studied, to enable the application of a broader range of algorithms and consequently the discovery of different types of communities. Third, we plan to define additional measures aimed at evaluating the quality of the detected topics, going beyond the basic measure of word overlapping used in this paper. Finally, we aim to assess the effects of these design choices on different kinds of texts. For example, we can expect different window sizes to be relevant for shorter documents, such as social media posts, and different vocabulary sizes to lead to networks with different sizes and densities.

Acknowledgements

The authors acknowledge the financial support provided by the “Dipartimenti Eccellenti 2018-2022” ministerial funds. This work has also been partly funded by eSSENCE, an e-Science collaboration funded as a strategic research area of Sweden, and by EU CEF grant number 2394203 (NORDIS - NORdic observatory for digital media and information DISorder).

References

1. Alghamdi, R., Alfalqi, K.: A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 6, 147–153 (2015).
2. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: A brief survey of text mining: Classification, clustering and extraction techniques. arXiv:1707.02919, 1–13 (2017).
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res., 3, 993–1022 (2003).
4. Blondel, V.D., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech., 1–12 (2008).
5. Bullinaria, J.A., Levy, J.P.: Extracting semantic representations from word co-occurrence statistics: A computational study. Behav. Res. Methods, 39, 510–526 (2007).
6. Dang, T., Nguyen, V.T.: ComModeler: Topic Modeling Using Community Detection. In: Tominski, C., von Landesberger, T. (eds.) EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association, pp. 1—5. (CH) (2018).
7. de Arruda, H.F., Costa, L.F., Amancio, D.R.: Topic segmentation via community detection in complex networks. Chaos, 26, 1–10 (2015).
8. Greene, D., Cunningham, P.: Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In: Proc. 23rd International Conference on Machine learning (ICML'06), pp. 377–384. ACM Press, New York (2006).
9. Hamm, A., Odrowski, S.: Term-Community-Based Topic Detection with Variable Resolution. Information, 12, 221–252 (2021).
10. Hubert, L., Arabie, P.: Comparing partitions. J. Classif., 2, 193—218 (1985).
11. Kim, M., Sayama, H.: The power of communities: A text classification model with automated labeling process using network community detection. In: International Conference on Network Science, pp. 231–243. Springer, Berlin (2020).
12. Lancichinetti, A., Sirer, M.I., Wang, J.X., Acuna, D., K öording, K., Amaral, L.A.N.: High-reproducibility and high-accuracy method for automated topic classification. Phys. Rev. X., 5, 1-–11 (2015).
13. Newman M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E, 74, 1–2 (2006).
14. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, 814–818 (2005).
15. Salerno, M.D., Tataru, C.A., Mallory, M.R.: Word Community Allocation: Discovering Latent Topics via Word Co-Occurrence Network Structure. (2015). http://snap.stanford.edu/class/cs224w-2015/projects_2015/Word_Community_Allocation.pdf
16. Sayyadi, H., Raschid, L.: A graph analytical approach for topic detection. ACM Trans. Internet Technol., 1–23 (2013).
17. Uysal, A.K., Gunal, S.: The impact of preprocessing on text classification. IInf. Process. Manage., 50, 104-112 (2014).
18. Usai, A., Pironti, M., Mital, M., Mejri, C.A.: Knowledge discovery out of text data: a systematic review via text mining. J. Knowl. Manag., 22, 1471–1488 (2018).
19. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput. Surv., 45, 1–35 (2013).