Objective Functions to Determine the Number of Topics for Topic Modeling

Silvio Peikert
Fraunhofer FOKUS
Berlin, Germany
silvio.peikert@fokus.fraunhofer.de

Clemens Kubach
Fraunhofer FOKUS
Berlin, Germany
clemens.kubach@fokus.fraunhofer.de

Jamal Al Qundus
Faculty of Engineering and Information Technology, Middle East University
Amman, Jordan
jalqundus@meu.edu.jo

Le Duyen Sandra Vu*
Fraunhofer FOKUS
Berlin, Germany
le.duyen.sandra.vu@fokus.fraunhofer.de

Adrian Paschke
Fraunhofer FOKUS
Berlin, Germany
adrian.paschke@fokus.fraunhofer.de

ABSTRACT

Topic modeling is a well-known task in unsupervised machine learning, where clustering algorithms are used to find latent topics. Several algorithms are presented in the literature, but the best known of them suffer from the drawback of requiring a lot of hyperparameter tuning to achieve good results. Especially, the number of latent topics or clusters (k) needs to be known in advance. In view of this situation, this paper analyses objective functions that help to evaluate the models in order to determine optimal hyperparameters. An empirical qualitative study was conducted using the NMF algorithm on different datasets to experimentally determine numerical properties of topic models which indicate an optimal k. Based on this study, we propose objective functions to select optimal topic models and discuss their results on different datasets.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

topic model evaluation, topic modeling, hyperparameter tuning, topic model coherence, non-negative matrix factorization, latent dirichlet allocation

ACM Reference Format:
Silvio Peikert, Clemens Kubach, Jamal Al Qundus, Le Duyen Sandra Vu, and Adrian Paschke. 2021. Objective Functions to Determine the Number of Topics for Topic Modeling . In *The 23rd International Conference on Information Integration and Web Intelligence (iiWAS2021), November 29-December 1, 2021, Linz, Austria.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487664.3487710

iiWAS2021, November 29-December 1, 2021, Linz, Austria
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-9556-4/21/11…$15.00
https://doi.org/10.1145/3487664.3487710

1 INTRODUCTION

Intuitively, we are constantly seeking knowledge by examining data to achieve various goals, e.g., content curation and system assessment [17, 22], content quality [2, 16], trusted sources [1, 3], disaster detection [10], pandemic prevention [21]. We are surrounded by data, which is produced in increasing quantities. This imposes challenges to document archive management, which require new tools for automatic organization, exploration, indexing, and search [5] to be solved. A technique called topic modeling has been developed for document management e.g. [20, 25]. It comes from the research areas of machine learning and statistics, and is thus based on the basic idea of recognizing combinations of patterns - called features - and their frequencies in order to compute probabilities to identify latent topics in a corpus and assign documents to these topics. In this way, topic models lead to a dimensionality reduction of large collections of unstructured documents and thus make document archives accessible.

Topic models rely on algorithms to cluster documents in order to discover topics included in their content. One of the most popular algorithms is the Latent Dirichlet Allocation (LDA) proposed by Blei et al. [7]. Another effective algorithm is the Non-negative Matrix Factorization (NMF) proposed by Paatero and Tapper [19]. These algorithms assume prior knowledge about the data, which need to be specified as hyperparameters. In particular, the exact number of latent topics needs to be known in advance. "However, finding the right number of latent topics in a given corpus has remained an open ended question" [6]. Moreover, there is "no easy way to choose the proper number of topics in a model beyond a major iterative approach" and "no standard method for choosing the number of topics" noted [26] and [27], respectively. Nevertheless, best practice is the distributional evaluation of topics generated by topic models. For this purpose, statistical measures such as perplexity and topic coherence have been widely used in the literature. Perplexity is a statistical measurement used in information theory often as a metric to evaluate the performance of topic/language models [13], while it has inverse proportional relation to model generalisation. Topic coherence measures how the words of a topic are related to each other and is based on Pointwise Mutual Information (PMI). PMI measures the relatedness between topic words and context features and is used to determine coherence scores by measuring the distance between vectors [4]. The topic coherence is fundamental to our approach to evaluate the distribution of words across generated topics within a selected number of topics because of its advantages in interpretability and understandability for humans.

2 RELATED WORK

The literature search revealed a number of methods for predicting an optimal number of topics without prior knowledge of the data under consideration. Some are tractable, others are challenging to interpret. Perplexity [26], rate of perplexity change (RPC) [26], coherence [18], average cosine distance [8], symmetric Kullback-Leibler (KL) divergence [6], a simple heuristic based on user queries [12], Normalized Absolute Coherence (NAC) and Normalized Absolute Perplexity (NAP) are methods for such estimations that aim to achieve both high accuracy and low processing time for LDA [14]. Among these works are interesting approaches, however they almost all [6, 14, 15, 26, 27] agree that there is no simple heuristic to determine an optimal number of latent topics in advance without performing an iteration process, which involves to actually compute the topic models.

Zhao et al. [26] use a cross-validation method to calculate the rate of perplexity change (RPC) as a function in order to determine an appropriate number of topics. The RPC function is calculated with P_i as the average perplexity for the i-th candidate of the r topic numbers $t_1, .., t_i, ..t_r$. These are subsequently formulated as a rate of change in $RPC(i) = |\frac{P_i - P_{i-1}}{t_i - t_{i-1}}|$. To select an appropriate number of topics with this function, the first change-point where the value of the candidate with a lower topic number is smaller than the following candidate with the next higher number of topics is to be chosen. The stability and effectiveness of the approach were evaluated on three different types and sizes of datasets. To evaluate stability, the RPC-based method and the perplexity-based method were compared. Whereas the hierarchical clustering algorithm and k-means clustering were applied to labeled datasets. Cluster analysis was applied on the LDA model output to evaluate the effectiveness of the proposed method against a true label.

Considering LDA as a matrix factorization method, Arun et al. [6] propose a measure based on the symmetric KL-Divergence and calculating the decomposition distributions of (1) the value of the topic-word matrix and (2) the value of the document topic resulting from the binary factorization of the document-word matrix. The calculated values are comparable and can be used to determine an optimal number of topics. For this purpose, a k out of the small range at a corresponding dip in the resulting curve has to be selected. Empirical results show that the number is proportional to the size of the dataset and as the size of the dataset increases, their optimal number is any number in the small range. The effectiveness of the approach was tested on multiple datasets consisting of texts and images.

Hasan et al. [14] propose two methods called Normalized Absolute Coherence (NAC) and Normalized Absolute Perplexity (NAP) to determine the optimal number of topics. They use K-fold cross-validation to construct LDA models. An ordinate represents a range of values for candidate numbers of the optimal number of topics. LDA models were trained K times for each candidate value of a number of topics, and perplexity, coherence, and F-measure were calculated. The average values of all metrics values are used in the calculation of NAP and NAC to determine the optimal number of topics. They use datasets from biology and medicine labeled with topics tagged by domain experts. The accuracy of a LDA model is measured by clustering efficiency, which was evaluated by comparing predicted clusters generated by the LDA model with the true classes identified by experts.

Newman et al. [18] used different scoring methods based on WordNet, Wikipedia, and the Google search engine to estimate a coherence within a topic. Based on the work of Chang et al. [9], the drawback of perplexity-based measurements in aspects of human interpretability is considered. Therefore, focusing on the intrinsic semantic quality of the topics is fundamental for all of their methods. Their results are evaluated by comparing them to human ratings (good, neutral, or bad) for a set of top-10 words representing the topics. The evaluation showed that the proposed method based on co-occurrence of words with Pointwise Mutual Information (PMI) has a high correlation with human ratings.

Röder et al. [23] compared and evaluated several measurement methods for topic coherence with the correlation to human ratings. For this purpose, they considered various measures based on different configurations proposed in literature, e.g., PMI as well as normalized PMI (NPMI). They found that NPMI was closer to human ratings than the others. From their "systematic study of the configuration space of the coherence measures", a new measure (C_v) was identified as best performing and is formulated in more detail by Syed and Spruit [24].

The configuration space proposed by Röder et al. [23] is subdivided in four components: segmentation, probability estimation, confirmation measure and aggregation, each of which can be described separately by various approaches. Segmentation refers to the set of words W of a given topic and depends on how the words should be compared to each other. It results in a set of pairs $S = (W', W^*)$ of words or word subsets. For the later coherence calculation, the pairs are needed to determine how strongly the respective subsets support each other. For the probability estimation, the documents of a reference corpus, which can also be the same corpus used for the topic modeling process, have to be used with a chosen estimation method for calculating the probabilities P. With the use of a boolean sliding window of length l, each document is divided into virtual documents of a word set of size l beforehand. Furthermore, they distinguish between direct and indirect confirmation measures that compute for each segmented pair S and the corresponding probabilities P "how strong the conditioning word set W^* supports W'". The direct variant includes measures such as PMI and NPMI. Indirect measures are composed of a direct measure and a method for calculating a vector similarity. Finally, the individual confirmation values are aggregated to a single value for the given topic defined by its top words. C_v is based on an indirect cosine confirmation measure with NPMI and a boolean sliding window. According to their results, a topic model can be evaluated by applying this measure to each topic of a model and calculating the arithmetic mean of the resulting values.

3 METHODOLOGY

Our approach uses the advantages of a topic coherence measurement proposed by Röder et al. [23] for internal topic evaluation and combines it with external topic evaluation using measures based on topic similarity. Subsequently, we can join them together for a comparison between entire models with an extended focus on interpretability by humans.

For our experimental setup, we consider two datasets that contain documents of the short text type. Typically, short texts contain a dominant topic compared to regular texts. We chose the well-known data collection from the 20-Newsgroups[1] dataset as objective reference. It contains 18,846 documents that are labeled and structured into 20 groups. These represent true classes that can be used as a gold standard for evaluating topic modeling approaches. As Topic Modeling is typically applied to unstructured copora, we also created a more natural dataset: The Tagesschau Corpus with over 14,000 articles of the German TV news[2]. We have collected the news articles from the period 03.02.2020 to 03.02.2021. The corpus contains a wide range of different documents on topics present in the news during the mentioned period.

To generate the topic models, we focused on the NMF algorithm[3]. According to Chen et al. [11], NMF performs better than LDA for short text datasets on the same configuration. Another advantage of NMF over LDA is that it produces better results with less parameter tuning, as we observed in our own experiments. In particular, the focus of our metric is on the influence of the number of topics parameter for a given corpus, which is common for most algorithms. Therefore, NMF is a good reference for this demonstration, however the applicability of our method does not depend on the algorithm used.

The high-level procedure is shown in figure 1. When the topic models for a set of various k are created, we have to evaluate them. Firstly, topic coherence is computed using the C_v metric, with the corresponding dataset as the reference corpus, that describes the semantic quality of the respective topic labels consisting of N top words -see Section 2-. In this way, we obtain an independent intrinsic score for each topic of a model, which highly correlates with human interpretability, as shown in [18], [23].

In addition to topic coherence as an intrinsic score, we also want to look in more depth at the interrelationship between the topics of a model and integrate this into scoring. As Cao et al. [8] already found for LDA and we also found for NMF that coherent topics are split in most cases when k is increased. In this case, a part of a previous topic is factorized into new similar topics, which resemble the previous topic to some extent. The corresponding topics are therefore very similar to each other. With an increasing amount of highly similar topics in a model, the information gain with increasing k is reduced resulting in worse model generalisation properties. We aim to determine these split topics by mapping each topic in a model to the most similar topic. For this purpose, we consider two variants for a mapping $\overrightarrow{m}$.

The first one is based on the normalized number of word overlaps of the respective topic labels. In this respect, it does not require any additional information about the weights of the concerned words. This procedure results in the mapping $\overrightarrow{m_{ove}}$ shown in equation 1.

In the second variant, we use cosine similarity $cos_{t_i,t_j}(\theta)$ as a measure of similarity between topics. For this variant, we use the entire topic vector t with the corresponding word weights extracted from the model matrix. Since cosine similarity already yields values between 0 and 1 due to the positive weights, no additional normalization with the number of top words is required. This procedure results in the mapping $\overrightarrow{m_{cos}}$ shown in equation 2.

$$\overrightarrow{m_{ove}} := \begin{pmatrix} max\{|t_1 \cap t|/2N|\forall t \in T \setminus \{t_1\}\} \\ \vdots \\ max\{|t_K \cap t|/2N|\forall t \in T \setminus \{t_K\}\} \end{pmatrix} \tag{1}$$

$$\overrightarrow{m_{cos}} := \begin{pmatrix} max\{cos_{t_1,t}(\theta)|\forall t \in T \setminus \{t_1\}\} \\ \vdots \\ max\{cos_{t_K,t}(\theta)|\forall t \in T \setminus \{t_K\}\} \end{pmatrix} \tag{2}$$

$$\overrightarrow{\varphi}_{C_{Cv}} := \overrightarrow{C_v} \odot (\overrightarrow{1} - \overrightarrow{m}) \tag{3}$$

$$C_{Cv} := avg(\overrightarrow{\varphi}_{C_{Cv}}) \tag{4}$$

Equation 3 represents the combination of the topic coherence values and a similarity measure in $\overrightarrow{m}$, where $\overrightarrow{m}$ can be one of the two mapping variants, either $\overrightarrow{m_{ove}}$ or $\overrightarrow{m_{cos}}$. The vector $\overrightarrow{C_v}$ of topic coherence values is component-wise multiplied by the vector calculated from 1 minus the corresponding mapping results in a vector $\overrightarrow{\varphi}_{C_{Cv}}$ with the values of the combined topic coherence.

By this measure in both variants, topics with high coherence but lower unique features relative to the other topics in the model are scored lower. The arithmetic mean value over these scores results in the relevant combined coherence C_{Cv} of the whole model (equation 4). The values of this metric is equal or lower than the score produced by the plain C_v metric. To validate our scoring function, we exhaustively search over a wide range of k to obtain a curve. A useful scoring function should have a significant maximum at or at least near the true k of the corpus.

4 RESULTS AND DISCUSSION

Figures 2 and 3 illustrate the measured values of C_v and both C_{Cv} variants for the considered corpora. In general, the curves of the coherences initially rise at low k. After a maximum, a downtrend follows with increasing k. The model with the corresponding number of topics with the maximum coherence score is an indicator for the best model.

For the 20 newsgroups corpus in figure 2, we can observe a downward trend for all three metrics from a certain maximum. Even the plain C_v coherence performs well in that aspect, however the difference between its maximum value and its neighbouring samples is smaller and thus the location of the maximum is more prone to shifting cause by noise than the results of both variants of the combined coherence. In particular, C_{Cv} with cosine similarity offers a more significant difference between its maximum and other close candidates and is therefore more stable, while it suggests the same optimal model like C_v at k=17, which is slightly below the assumed 20 groups of the labeled dataset. The fact that even

[1] The data in its original format can also be found at http://qwone.com/~jason/20Newsgroups/.

[2] They publish their written content of the last 365 days at https://www.tagesschau.de/archiv/.

[3] We used the NMF implementation provided in the Python library Gensim https://pypi.org/project/robics/

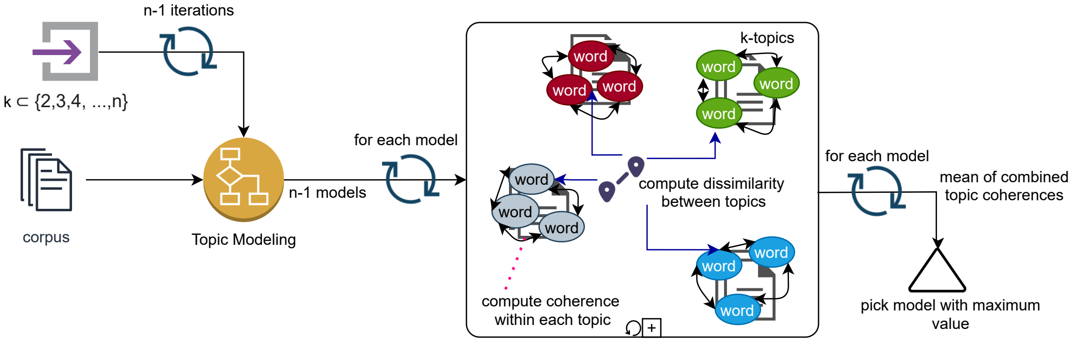
n-1 iterations
k ⊂ {2,3,4, ...,n}
corpus
Topic Modeling
n-1 models
for each model
k-topics
word
compute dissimilarity between topics
compute coherence within each topic
for each model
mean of combined topic coherences
pick model with maximum value

Figure 1: illustrates the workflow proposed for topic model selection

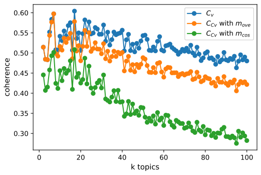

Figure 2: 20 Newsgroups corpus

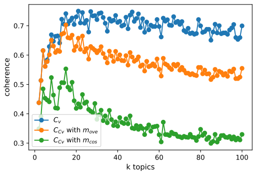

Figure 3: Tagesschau corpus

the plain C_v coherence gives a good result in this case could be caused by the strictly grouped characteristics of the corpus. Models with more than the 20 pre-categorized groups force articles from different topics to be mixed as a consequence, which is reflected in the intrinsic topic coherence. Importantly, even with these characteristics of the corpus, our combined versions could preserve the local maximum of the C_v metric.

Looking at the results of the more natural collection from the Tagesschau corpus in figure 3, a downtrend and a clear maximum can only be estimated very weakly with C_v. The increasing similarity between topics is not reflected in the coherence values. In contrast, with both versions of our C_{Cv} metric, we are still able to select the best topic model with k=15 and models with higher k were correctly scored lower because of increasing intersections between topics.

The idea of evaluating pairwise similarities between topics while keeping the advantages of topic coherences was successfully combined and demonstrated with our metric. Models with an increasing number of similar topics tend to lose their score. Especially, the version based on cosine similarity in mapping $\overrightarrow{m_{cos}}$ can enable a balance between human interpretability and good generalisation properties by reducing the contribution of split topics to the computed score. According to our experiments, C_{Cv} with $\overrightarrow{m_{cos}}$ is the best suited objective function for determining an optimal k for both corpora.

5 CONCLUSION AND FURTHER WORK

In principle, topic modeling for a complex corpus and without domain knowledge about the contents it contains is a difficult task. For the best known algorithms, an optimal number of topics k needs to be identified. To evaluate resulting models, the measurement of the semantic quality of the generated topics with topic coherence is insufficient. Also the aspect of similarity between different topics has an essential meaning. With the two variants of our objective function, which combines the intrinsic and extrinsic aspect, a measure to evaluate model quality was presented and its effectiveness was demonstrated. Using this metric an optimal model can be obtained using parameter search methods.

Further studies should investigate a comparison with methodologies of related work, including comprehensive experiments over various corpora. For this purpose, the resulting topics of the best models should be evaluated by considering their generalised coverage of all information contained in the datasets. This can be achieved through a survey of domain experts or by performing an analysis of results on synthetic corpora.

ACKNOWLEDGMENTS

This work has been partially supported by the "Wachstumskern Qurator – Corporate Smart Insights" project (03WKDA1F) funded by the German Federal Ministry of Education and Research (BMBF)

REFERENCES

[1] Jamal Al Qundus and Adrian Paschke. 2018. Investigating the effect of attributes on user trust in social media. In *International conference on database and expert systems applications*. Springer, 278–288.
[2] Jamal Al Qundus, Adrian Paschke, Shivam Gupta, Ahmad M Alzouby, and Malik Yousef. 2020. Exploring the impact of short-text complexity and structure on its quality in social media. *Journal of Enterprise Information Management* (2020).
[3] Jamal Al Qundus, Adrian Paschke, Sameer Kumar, and Shivam Gupta. 2019. Calculating trust in domain analysis: Theoretical trust model. *International Journal of Information Management* 48 (2019), 1–11.
[4] Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. 13–22.
[5] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6, 1 (2015).
[6] Rajkumar Arun, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 391–402.
[7] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.
[8] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing* 72, 7-9 (2009), 1775–1781.
[9] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*. 288–296.

[10] Salim Chehida, Abdelhakim Baouya, Saddek Bensalem, and Marius Bozga. 2020. Applied statistical model checking for a sensor behavior analysis. In *International Conference on the Quality of Information and Communications Technology*. Springer, 399–411.

[11] Yong Chen, Hui Zhang, Rui Liu, Zhiwen Ye, and Jianying Lin. 2019. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems* 163 (2019), 1–13.

[12] Romain Deveaud, Eric SanJuan, and Patrice Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17, 1 (2014), 61–84.

[13] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101, suppl 1 (2004), 5228–5235.

[14] Mahedi Hasan, Anichur Rahman, Md Razaul Karim, Md Saikat Islam Khan, and Md Jahidul Islam. 2021. Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA). In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*. Springer, 341–354.

[15] Thomas Jacobs and Robin Tschötschel. 2019. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology* 22, 5 (2019), 469–485.

[16] Matheus Marinho, Danilo Arruda, Fernando Wanderley, and Anthony Lins. 2018. A systematic approach of dataset definition for a supervised machine learning using NFR framework. In *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*. IEEE, 110–118.

[17] Isao Namba, Rieko Yamamoto, and Mikio Aoyama. 2020. Towards Guidelines for Assessing Qualities of Machine Learning Systems. In *Quality of Information and Communications Technology: 13th International Conference, QUATIC 2020, Faro, Portugal, September 9-11, 2020, Proceedings*, Vol. 1266. Springer Nature, 17.

[18] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. 100–108.

[19] Pentti Paatero and Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 2 (1994), 111–126.

[20] Jamal Al Qundus, Silvio Peikert, and Adrian Paschke. 2021. AI supported topic modeling using KNIME-workflows. *arXiv preprint arXiv:2104.09428* (2021).

[21] Jamal Al Qundus, Ralph Schäfermeier, Naouel Karam, Silvio Peikert, and Adrian Paschke. 2021. ROC: An Ontology for Country Responses towards COVID-19. *arXiv preprint arXiv:2104.07345* (2021).

[22] Georg Rehm, Peter Bourgonje, Stefanie Hegele, Florian Kintzel, Julián Moreno Schneider, Malte Ostendorff, Karolina Zaczynska, Armin Berger, Stefan Grill, Sören Räuchle, et al. 2020. QURATOR: innovative technologies for content and data curation. *arXiv preprint arXiv:2004.12195* (2020).

[23] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.

[24] Shaheen Syed and Marco Spruit. 2017. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, 165–174.

[25] Malik Yousef, Jamal Al Qundus, Silvio Peikert, and Adrian Paschke. 2020. TopicsRanksDC: Distance-Based Topic Ranking Applied on Two-Class Data. In *International Conference on Database and Expert Systems Applications*. Springer, 11–21.

[26] Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics*, Vol. 16. Springer, 1–10.

[27] Chen Zou. 2018. Analyzing research trends on drug safety using topic modeling. *Expert opinion on drug safety* 17, 6 (2018), 629–636.