JOIN (Jurnal Online Informatika)
Volume 6 No.1 | June 2021: 17-24
DOI: 10.15575/join.v6i1.636

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Discovering Computer Science Research Topic Trends using Latent Dirichlet Allocation

Kartika Rizqi Nastiti [1], Ahmad Fathan Hidayatullah[2], Ahmad R. Pratama[3]
[1,2,3]Department of Informatics, Universitas Islam Indonesia, Indonesia

Article Info

Article history:

Received September 12, 2020
Revised December 07, 2020
Accepted January 11, 2021
Published June 15, 2021

Keywords:

Coherence value
Computer science
Latent Dirichlet Allocation
Topic discovery
Topic modeling

ABSTRACT

Before conducting a research project, researchers must find the trends and state of the art in their research field. However, that is not necessarily an easy job for researchers, partly due to the lack of specific tools to filter the required information by time range. This study aims to provide a solution to that problem by performing a topic modeling approach to the scraped data from Google Scholar between 2010 and 2019. We utilized Latent Dirichlet Allocation (LDA) combined with Term Frequency-Indexed Document Frequency (TF-IDF) to build topic models and employed the coherence score method to determine how many different topics there are for each year's data. We also provided a visualization of the topic interpretation and word distribution for each topic as well as its relevance using word cloud and PyLDAvis. In the future, we expect to add more features to show the relevance and interconnections between each topic to make it even easier for researchers to use this tool in their research projects.

Corresponding Author:

Ahmad Fathan Hidayatullah,
Department of Informatics,
Universitas Islam Indonesia,
Jl. Kaliurang km 14.5 Sleman Yogyakarta, Indonesia
Email: fathan@uii.ac.id

1. INTRODUCTION

The need for research, development, and applications in computer science and information technology throughout the world continues to grow over time [1]. In the last decade, Indonesia has been trying to catch up with other countries regarding research and research publications. Based on the higher education database in Indonesia as of 2019, there are at least 774 majors of informatics and computer science degrees, ranging from bachelor to doctorate degrees. That high number is undoubtedly followed by the needs for research among academia in higher education.

Before conducting a research project, researchers need to find information about the research trends in a particular field and topic over time. The research can be performed based on the data, case studies, methods, and other different variables from published papers in related journals or conference proceedings. Journals proceedings and conference papers are notable sources to gain research trends from various fields of science [2]. Often, that process can also be performed through some web search engines that index scholarly literature across various publishing formats and disciplines, like Google Scholar, based on keywords and published time. However, finding related research through search engine websites could not provide the research trends in a specific time period automatically. In addition, researchers must classify and identify manually from many scientific research papers to obtain the research trends in a particular topic.

Based on those problems, the development of some tools that can be used as a service and help researchers to obtain research trends in a particular field and topic is necessary. Therefore, a topic modeling approach can be proposed to reveal and extract topics from some given scientific articles. Topic modeling has been used by researchers as an approach to find research trends.

Lamba and Madhusudhan [3] presented a topic modeling using LDA to help librarians identify the trending themes in the library and information science field area. Hamzah et al. [4] applied the Latent Dirichlet Allocation method to report a map on mobile learning research topics trends. They examined a total of 146 papers from ScienceDirect and Scopus published in 2007-2018 by determining 5 to 50 topics. To find the most appropriate number of topics, they used coherence value. Finally, they found that there were 25 most relevant topics to their research goal.

Xu et al. [5] proposed a sensitive word weighted-Latent Dirichlet Allocation (LDA) model to recognize topics about network sensitive information. They constructed a vocabulary containing sensitive terms using word2vec. The proposed method has successfully improved the quantity of topic recognition and quality of sensitive information as well. Liu et al. [6] conducted a study to extract research topics in clinical psychology articles. They performed two different studies to extract the topics from clinical psychology journals using Latent Dirichlet Allocation. The first study extracted topics between 1981 and 2018. As for the second study, they employed a dynamic variant LDA to help recognize the development of the topics from 2007 until 2018.

Sun and Yin [7] applied Latent Dirichlet Allocation (LDA) on article abstract in transportation research and revealed 50 critical topics from 22 leading transportation journals from 1990 to 2015. They also performed temporal analysis for each journal. The study found that special issues on particular topics could be identified from temporal analysis. In addition, by measuring the temporal trends at the regional level, they could clearly find different research patterns from each country. Amado et al. [8] generated topics through Latent Dirichlet Allocation modeling. Their study aimed to identify the trends in big data marketing research by analyzing a total of 1560 articles published between 2010-2015. In that study, the authors revealed two main points from the topic analysis, such as cross-domain topics for big data and marketing and significant insights on authors' affiliations regarding the geographical area.

Another study by Zou [9] also applied Latent Dirichlet Allocation (LDA) to obtain research trends associated with drug safety based on titles and abstracts from the MEDLINE index from 2007 until 2016. The results were the popular research topics from year to year, topics distribution over words, and clusterization of topics distribution using dendrogram. The work by Waluya [10] retrieved relevant information using queries or keywords inputted by the users. Probabilistic Latent Semantic Analysis (PLSA) is applied to build topic models because it was considered capable of retrieving the information about similarity values between documents so that it can categorize each document into the relevant topics. Wu et al. [11] compared three topic modeling methods using China's web-based news portal. The research applied scenarios towards cosine value and compared it using K-center clusterization and used F-measure to calculate the result of clusterization itself. They found that better topic classes can be obtained by using complex methods, such as PLSA2 and LDA, than by using simpler topic modeling methods.

In this study, we propose an implementation of topic modeling using Latent Dirichlet Allocation (LDA) for computer science and information technology research in Indonesia. Based on the previous studies, LDA is the most utilized unsupervised learning method in topic modeling to help determine topics from text [12]. LDA is the most popular probabilistic topic model to be applied in analyzing scientific publications in many areas, such as content comparison, citation network analysis, and time gaps [13]–[15].

Moreover, the topic modeling task will be conducted based on titles and the year of publication. To obtain better topic modeling results, we also propose TF-IDF (Term Frequency-Inverse Document Frequency) and phrase identification to be applied before the topic modeling task [16], [17]. The objective of this study is to facilitate academicians in obtaining the trends and illustration of research topics in the field of computer science and information technology in Indonesia. Furthermore, the results of topic modeling can be used as references and considerations before doing any research project.

2. METHOD

The research pipeline of our study is illustrated in Figure 1. In the first stage, we perform data collection. Secondly, preprocessing is carried out to remove unnecessary parts from the text. After preprocessing, we applied phrase identification to identify phrases from texts [17]. The next step is applying term weighting for each token using TF-IDF (Term Frequency-Inverse Document Frequency). Topic modeling and coherence value calculation are performed to obtain the topics and the best number of topics from the corpus. Finally, we conduct data visualization to illustrate our topic modeling result. A detailed explanation of each stage conducted in this study will be described respectively in the next section.

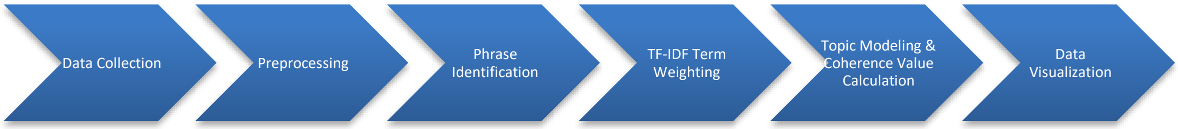


Figure 1. Research Pipeline

2.1. Data Collection

The data utilized in this study are collected using Harzing's Publish or Perish[1] desktop-based application that specifically scrapes the information of journal researches from scientific journal search engines, such as Google Scholar, Scopus, and Crossref. We collected the title of scientific papers published from 2010 until 2019 from Google Scholar. The titles of scientific papers are gained by using keywords related to computer science, informatics and information technology in Bahasa Indonesia, for example, *rekayasa perangkat lunak* (software engineering), *sistem cerdas* (intelligent system), *sistem informasi* (information system) and *jaringan komputer* (computer network).

2.2. Preprocessing

Preprocessing is an important step that can influence the results of topic modeling. An appropriate application of the steps in preprocessing will improve the quality of the resulting topics modeling. The preprocessing steps carried out in this study are adjusted to the conditions of the research title data that had been obtained previously. The preprocessing steps are as the following:

1. Omitting punctuation and symbols (e.g. :., - () & / '": +?! * # $%).
2. Omitting numbers, but not the numbers in the middle or directly after or before the alphabet.
3. Case folding by changing all letters into lowercase letters.
4. Removing stop words that lack meaning and do not represent the content of a particular sentence (e.g. *"dan"* (and), *"atau"* (or), *"yang"* (which)).

2.3. Phrase Identification

The phrase identification process is performed to obtain phrases from the text. It is essential to identify phrases from the corpus before performing topic modeling to avoid loss of context and meaning from topic results. For example, the phrase *"rumah sakit"* (hospital) will have a different context and meaning compared with the respective words, *"rumah"* (house) and *"sakit"* (sick). Therefore, in this study, we identify phrases by creating bigram and trigram models from our corpus. The phrase identification is carried out by counting the occurrences of two or three words that appear together. If the consecutive words appear together at least five times, they would be considered as a phrase by adding underscore between the consecutive words [17].

2.4. TF-IDF (Term Frequency-Inverse Document Frequency) Term Weighting

The purpose of TF-IDF term weighting is that the process of word analysis can be done more effectively by reducing the amount of vocabulary and eliminating noise or unnecessary words [18][19]. The TF-IDF calculation process used logarithmic notation as can be seen in equation (1).

$$\boldsymbol{w_{i,j} = tf_{i,j} \times log\ log\ (\frac{N}{df_i})} \tag{1}$$

For a term i in document j, the $tf_{i,j}$ represents the number of occurrences of i in j. The N notation represents the total number of documents. The df_i represents the number of documents containing i. Words with high TF-IDF values are words that rarely appear in many documents. The low TF-IDF values are for the words that appear the most in many documents. Any word with an obtained TF-IDF value of 0 means that that word is not in the document.

2.5. Topic Modeling with Latent Dirichlet Allocation (LDA) and Coherence Value Calculation

Latent Dirichlet Allocation (LDA) is an unsupervised learning method in machine learning techniques. LDA is a generative probabilistic model that is applied to a set of discrete data such as text data [10]. The visual model representation of LDA can be seen in Figure 2 [20].

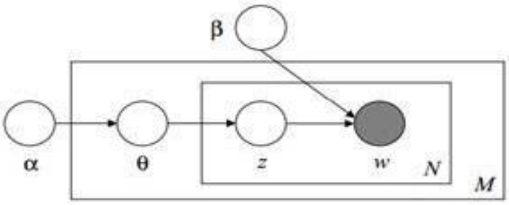


Figure 2. LDA model representation

Figure 2 illustrates the three levels of LDA representation, where M represents the document and N represents the number of words in a document. The first level is the corpus level parameter represented by α and β. This corpus level parameter is assumed to be sampled once in the corpus generation process, then the

[1] https://harzing.com/resources/publish-or-perish

document level variable (θ) will be sampled once for each document. Furthermore, word-level variables are symbolized by z and w. Both variables will be sampled once for each word in each document. Based on the notations, the generative process on LDA will correspond to the joint distribution of latent variables and observed variables. The calculation of the probability of a corpus is shown by the equation (2).

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d \tag{2}$$

The β notation describes the topic, where each β is a distribution of several of words. The variable θ_d is the document level variable with one sample per document representing the proportion of topics for documents to d. The z_{dn} and w_{dn} notations represent variables at the word level with one sample for each word in each document.

During the topic modeling task using LDA, we calculate a coherence score to obtain the most appropriate number of topics [17]. Topic coherence calculates a particular topic by computing the degree of semantic closeness between high-scoring terms in the topic [21]. We determined the best number of topics based on the highest coherence score [17]. In this study, we employed the C_v measure as a method for measuring coherence. The C_v measurement is based on a sliding window that utilizes normalized pointwise mutual information (NPMI) and the cosine similarity [22].

Ultimately, we also perform document counting to find out the number of documents that belong to a particular topic. The term ‘documents’ at this stage refers to the research title. The purpose of this process is to find out the number of documents performing research on a particular topic so that we can obtain the research trend in a particular period of time. For each research title, there would be one topic with the highest contribution to it. In this stage, we assign corpus and TF-IDF values of each word as the input. The output of this stage is the most dominant topic for each research title.

2.6. Visualization

This study utilizes two different visualization tools to visualize the topic models, pyLDAvis and word cloud. PyLDAvis can be utilized to see the word distribution for each relevant topic with its level of relevance [23]. Moreover, PyLDAvis is also used to visualize the closeness between one topic and another using a Cartesian diagram. As for the word cloud, it shows the word visualization based on the word frequency [24]. Word cloud visualization will display as the representation of the word distribution for each topic. The word cloud results are presented using a dashboard page where users could specify the year or year range to look for research trends.

3. RESULTS AND DISCUSSION

We have collected about 81,516 titles of the scientific papers. We then apply preprocessing tasks to our dataset. After that, bigram and trigram language models are built to identify the phrase by using Gensim’s Phrases model. The formation of bags of words is obtained by calculating the value of TF-IDF using the Gensim library that is already provided in Python.

The next step is to make an LDA model for each year. To build an LDA model, this study utilizes the Gensim library with the *LdaModel()* function, while to get the coherence value, the *CoherenceModel()* function is used. For the *LdaModel()* function, there are four parameters that must be fulfilled, namely corpus, id2word, num_topics, and iterations. Corpus is a parameter that refers to the dataset of our study. The id2word parameter refers to the dictionary that is mapping from word to IDS to words. The id2word parameter is utilized to identify the vocabulary size of our corpus. Num_topics parameter is assigned by the number of topics that we want to obtain from the corpus. Finally, the iterations are used to assign the maximum number of iterations through the data when concluding the topic distribution from the corpus.

As for the coherence value, it was obtained after the model had been generated. The *CoherenceModel()* function was applied to evaluate the topic model result. The number of topics with the highest coherence value will be chosen as the most appropriate number of topics. In our experiment, we assigned the range number of topics from 2 to 20 topics to obtain the best number of topics. The example results of calculating the coherence value in the year 2019 can be seen in Figure 3, and the details of the coherence value for the diagram can be seen in Figure 4. According to the chart in Figure 3, it is clear that the highest coherence value is achieved when the number of topics is equal to 16. Therefore, we can conclude that the most suitable number of topics in 2019 is 16 topics.

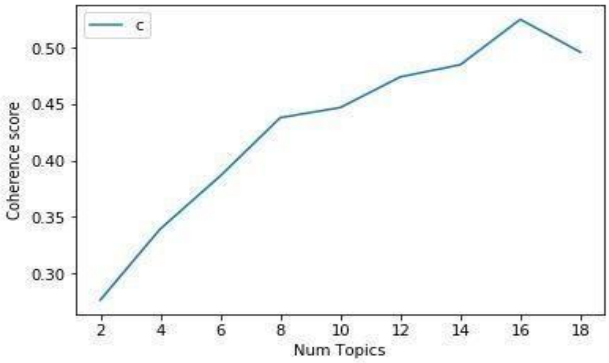
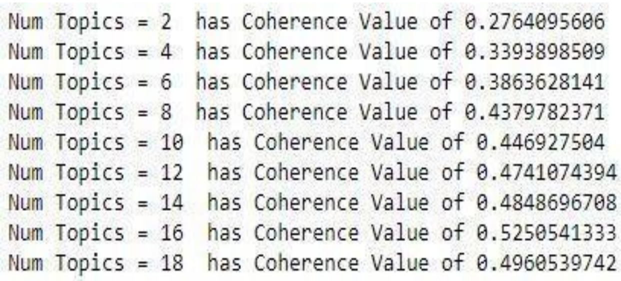


Figure 3. Coherence score chart in the year 2019



```
Num Topics = 2  has Coherence Value of 0.2764095606
Num Topics = 4  has Coherence Value of 0.3393898509
Num Topics = 6  has Coherence Value of 0.3863628141
Num Topics = 8  has Coherence Value of 0.4379782371
Num Topics = 10  has Coherence Value of 0.446927504
Num Topics = 12  has Coherence Value of 0.4741074394
Num Topics = 14  has Coherence Value of 0.4848696708
Num Topics = 16  has Coherence Value of 0.5250541333
Num Topics = 18  has Coherence Value of 0.4960539742
```

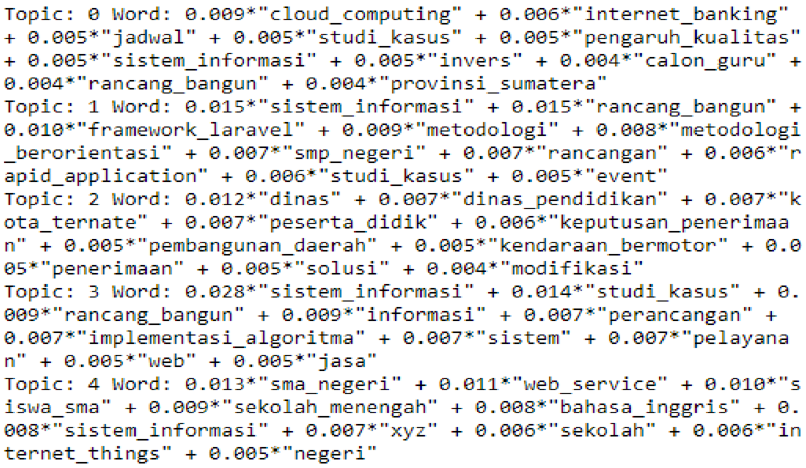
Figure 4. The list of coherence values of 2019

Table 1 shows more detail about the coherence values result and the most appropriate number of topics from 2010 until 2019. In addition, we also make aggregate for several different periods of the year. For instance, we merge the scientific research papers in the year 2018-2019, 2017-2019, 2016-2019, until 2010-2019.

Table 1. Coherence Values

Years	Number of Topics	Coherence Value
2010	18	0.4814599141
2011	16	0.5495813255
2012	18	0.4934268752
2013	18	0.5096695948
2014	12	0.4985809264
2015	6	0.5329307382
2016	18	0.54215305
2017	16	0.5160163434
2018	8	0.5169211186
2019	10	0.6078580128
2018–2019	14	0.5183195359
2017–2019	16	0.4987779732
2016–2019	18	0.4776285773
2015–2019	16	0.4994886329
2014–2019	14	0.5045297658
2013–2019	16	0.5417630578
2012–2019	14	0.4588848188
2011–2019	16	0.4579770423
2010–2019	18	0.4505668673

Figure 5 shows the sample of topic model results in the year of 2019. The model in the figure shows a list of keywords and their weightage or importance of each keyword from each topic. For example, we can conclude that cloud computing is the most discussed topic in the topic 0 (zero) with the weight of 0.009 because the phrase "cloud computing" has the highest importance. To help understand the topic better, we utilize word cloud visualization. Word cloud visualization is a visualization tool that provides the relative prevalence of words that have been ranked by their importance to illustrate the main themes of a collection of texts [25]. In word cloud, the size of the words represents the importance of the words from a collection of texts. The bigger the size of a word/term/phrase indicates that it appears more frequently. Figure 6 shows an example of word clouds that we have generated from the topic in the year of 2019. From the example in Figure 6, we can see two examples of word clouds. Based on the size of the term, the word cloud at the top discusses cloud computing topic while the word cloud at the bottom represents information system topic.



```
Topic: 0 Word: 0.009*"cloud_computing" + 0.006*"internet_banking"
+ 0.005*"jadwal" + 0.005*"studi_kasus" + 0.005*"pengaruh_kualitas"
+ 0.005*"sistem_informasi" + 0.005*"invers" + 0.004*"calon_guru" +
0.004*"rancang_bangun" + 0.004*"provinsi_sumatera"
Topic: 1 Word: 0.015*"sistem_informasi" + 0.015*"rancang_bangun" +
0.010*"framework_laravel" + 0.009*"metodologi" + 0.008*"metodologi
_berorientasi" + 0.007*"smp_negeri" + 0.007*"rancangan" + 0.006*"r
apid_application" + 0.006*"studi_kasus" + 0.005*"event"
Topic: 2 Word: 0.012*"dinas" + 0.007*"dinas_pendidikan" + 0.007*"k
ota_ternate" + 0.007*"peserta_didik" + 0.006*"keputusan_penerimaa
n" + 0.005*"pembangunan_daerah" + 0.005*"kendaraan_bermotor" + 0.0
05*"penerimaan" + 0.005*"solusi" + 0.004*"modifikasi"
Topic: 3 Word: 0.028*"sistem_informasi" + 0.014*"studi_kasus" + 0.
009*"rancang_bangun" + 0.009*"informasi" + 0.007*"perancangan" +
0.007*"implementasi_algoritma" + 0.007*"sistem" + 0.007*"pelayana
n" + 0.005*"web" + 0.005*"jasa"
Topic: 4 Word: 0.013*"sma_negeri" + 0.011*"web_service" + 0.010*"s
iswa_sma" + 0.009*"sekolah_menengah" + 0.008*"bahasa_inggris" + 0.
008*"sistem_informasi" + 0.007*"xyz" + 0.006*"sekolah" + 0.006*"in
ternet_things" + 0.005*"negeri"
```

Figure 5. Sample of topic modeling result

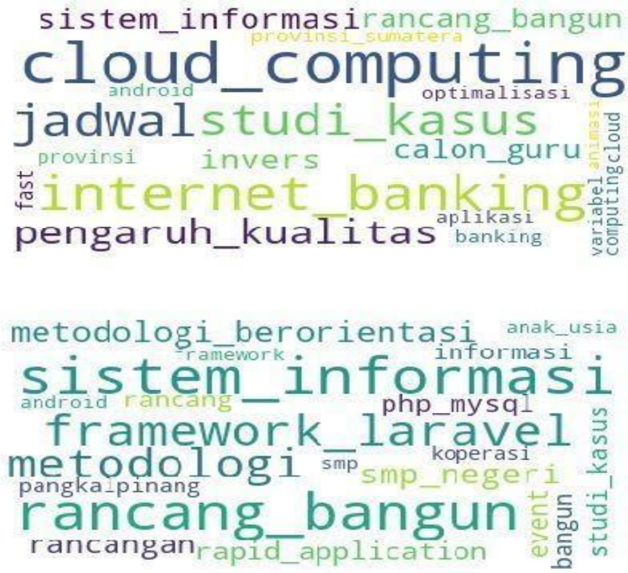


Figure 6. Example of word clouds for topics in 2019

In this work, we also provide another topic modeling visualization using PyLDAvis library. PyLDAvis is able to display the distribution of words of each topic along with the level of relevance of each word for the topic. The left panel on PyLDAvis shows a general perspective of the topics [8]. On the left panel, the areas of the circles are corresponding to the relative prevalences of the topics within the corpus. In addition, the left panel also provides information about inter-topic distances. The right panel provides information about the overall term frequency and approximation term frequency within the selected topic. An example of pyLDAvis is shown in Figure 7 where we can see that the term *sistem_information* (information system) dominates the overall term frequency. The same term also dominates in topic number one, which can be inferred that the most discussed topic for topic number one is information system.

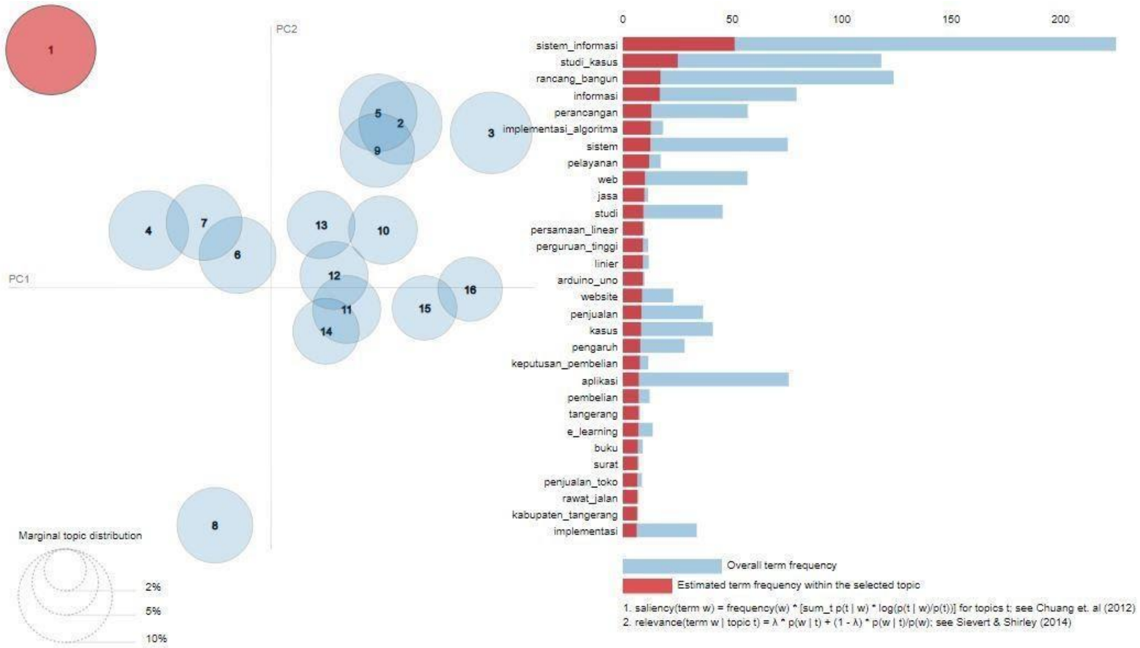


Figure 7. Topic Visualization for 2019 using PyLDAvis

4. CONCLUSION AND FUTURE WORK

In this study, we have successfully conducted a topic modeling of scholarly literature written in Bahasa Indonesia in the field of computer science and information technology. We performed the topic modeling by utilizing the Latent Dirichlet Allocation (LDA) method to discover topics and research trends in the field of computer science and information technology in Indonesia between 2010 and 2019. We combined the LDA method with Term Frequency-Indexed Document Frequency (TF-IDF) to build the topic models. The topic modeling result has provided good topic results regarding research trends in the field of computer science and information technology within a period of ten years. In addition, we also provided a visualization of the topic interpretation and word distribution for each topic along with its relevance using word cloud and PyLDAvis.

While this study has discovered and extracted some research trends from Google Scholar, it is still difficult to determine the appropriateness of some topics because there were terms that are still unevenly spread across several topics. Also, the word distribution for each topic in general tends to be less related to one another. Therefore, further research can be pursued to give a better analysis of why this is the case. With a deeper and further analysis, we can expect to add more features to show the relevance and interconnections between one topic to another. This new feature would make it even easier and more convenience for researchers to use this tool in exploring topics and discovering trends for their research projects.

5. REFERENCES

[1] Nastiti, Kartika Rizqi, “Pemodelan Topik untuk Penelitian di Bidang Informatika Menggunakan Metode Latent Dirichlet Allocation,” Undergraduate Thesis, Universitas Islam Indonesia, 2019.

[2] S. Das, K. Dixon, X. Sun, A. Dutta, and M. Zupancich, “Trends in Transportation Research: Exploring Content Analysis in Topics,” *Transportation Research Record,* vol. 2614, no. 1, pp. 27–38, Jan. 2017.

[3] M. Lamba and M. Madhusdhuan, “Application Of Topic Mining And Prediction Modeling Tools For Library And Information Science Journals,” *Zenodo,* Jan. 2018, DOI: 10.5281/zenodo.1298739.

[4] A. Hamzah, A. F. Hidayatullah, and A. G. Persada, “Discovering Trends of Mobile Learning Research Using Topic Modelling Approach,” *International Journal of Interactive Mobile Technologies (iJIM),* vol. 14, no. 9, pp. 1–11, 2020.

[5] G. Xu, X. Wu, H. Yao, F. Li, and Z. Yu, “Research on Topic Recognition of Network Sensitive Information Based on SW-LDA Model,” *IEEE Access*, vol. 7, pp. 21527–21538, Feb. 2019.

[6] S. Liu, R. Y. Zhang, and T. Kishimoto, “Analysis and prospect of clinical psychology based on topic models: hot research topics and scientific trends in the latest decades,” *Psychology, Health & Medicine*, pp. 1–13, 2020.

[7] L. Sun and Y. Yin, “Discovering themes and trends in transportation research using topic modeling,” *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 49–66, Apr. 2017, DOI: 10.1016/j.trc.2017.01.013.

[8] A. Amado, P. Cortez, P. Rita, and S. Moro, “Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis,” *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1–7, Jan. 2018, DOI: 10.1016/j.iedeen.2017.06.002.

[9] C. Zou, “Analyzing research trends on drug safety using topic modeling,” *Expert Opinion on Drug Safety*, vol. 17, no. 6, pp. 629–636, 2018.

[10] O. K. Waluya, “Penerapan Information Retrieval Menggunakan Pemodelan Topik pada Dokumen Skripsi (Studi Kasus Ruang Baca Teknik Informatika UMG),” Undergraduate Thesis, Universitas Muhammadiyah Gresik, 2017.

[11] Y. Wu, Y. Ding, X. Wang, and J. Xu, “A comparative study of topic models for topic clustering of Chinese web news,” in *2010 3rd International Conference on Computer Science and Information Technology*, 2010, vol. 5, pp. 236–240.

[12] H. Chen, X. Wang, S. Pan, and F. Xiong, “Identify topic relations in scientific literature using topic modeling,” *IEEE Transactions on Engineering Management*, 2019.

[13] H. Chen, G. Zhang, D. Zhu, and J. Lu, “Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014,” *Technological Forecasting and Social Change*, vol. 119, pp. 39–52, Jun. 2017.

[14] Y. Ding, “Topic-based PageRank on author cocitation networks,” *Journal of the American Society for Information Science and Technology*, vol. 62, no. 3, pp. 449–466, 2011.

[15] A. Suominen and H. Toivanen, “Map of Science with Topic Modeling: Comparison of Unsupervised Learning and Human-Assigned Subject Classification,” *Journal of the Association for Information Science and Technology*, vol. 67, no. 10, pp. 2464–2476, 2016.

[16] G. Zhao, Y. Liu, W. Zhang, and Y. Wang, “TFIDF based Feature Words Extraction and Topic Modeling for Short Text,” in *Proceedings of the 2018 2nd International Conference on Management Engineering, Software Engineering and Service Sciences - ICMSS 2018*, Wuhan, China, 2018, pp. 188–191, DOI: 10.1145/3180374.3181354.

[17] A. F. Hidayatullah, W. Kurniawan, and C. I. Ratnasari, “Topic Modeling on Indonesian Online Shop Chat,” in *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval - NLPIR 2019*, Tokushima, Japan, 2019, pp. 121–126, DOI: 10.1145/3342827.3342831.

[18] B. Wang and S. Zhang, “A Novel Feature Selection Algorithm for Text Classification Based on TFIDF-Weight and KL-Divergence,” 2005, pp. 438–441.

[19] A. Danesh, B. Moshiri, and O. Fatemi, “Improve text classification accuracy based on classifier fusion methods,” in *2007 10th International Conference on Information Fusion*, Quebec City, QC, Canada, Jul. 2007, pp. 1–6, DOI: 10.1109/ICIF.2007.4408196.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, p. 30, 2003.

[21] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, “Exploring Topic Coherence over Many Models and Many Topics,” in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Islan, Korea, Jul. 2012, pp. 952–961.

[22] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM ’15*, Shanghai, China, 2015, pp. 399–408, DOI: 10.1145/2684822.2685324.

[23] C. Sievert and K. Shirley, “LDAvis: A method for visualizing and interpreting topics,” in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, Maryland, USA, 2014, pp. 63–70, DOI: 10.3115/v1/W14-3110.

[24] S. Jayashankar and R. Sridaran, “Superlative model using word cloud for short answers evaluation in eLearning,” *Education and Information Technologies*, vol. 22, no. 5, Oct. 2016.

[25] A. L. Uitdenbogerd, “World cloud: A prototype data choralification of text documents,” *Journal of New Music Research*, vol. 48, no. 3, pp. 253–263, May 2019, DOI: 10.1080/09298215.2019.1606255.