Text Mining to Analyse Publication Topics of COVID-19 using HDP and LDA Methods

Rakhmah Wahyu Mayasari[1], Kartika Fithriasari[2], Dedy Dwi Prastyo[3]
{rakhmah13@mhs.statistika.its.ac.id[1], kartika_f@statistika.its.ac.id[2], dedy-dp@statistika.its.ac.id[3]}

[1]Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia[1,2,3]

Abstract. COVID-19 is a disease caused by the novel coronavirus, which almost all countries are affected. This worldwide effect has led many researchers to conduct research related to COVID-19. It is wanted to know what topics have been carried out from all the studies published by researchers in various countries. This research analyzes the data crawled from full abstracts of publications related to COVID-19 start January 2020 to August 2020. The abstract's text was crawled and then preprocessed by eliminating punctuation, lowering text, lemmatizer, and stopword. Furthermore, the clean data is ready for analysis using the text mining method to allocate topics and use as future research information. The methods used are the Hierarchical Dirichlet Process (HDP) and Latent Dirichlet Allocation (LDA) approaches. It also found that the LDA method has a coherence score of 42% higher than the HDP method, which means LDA is more appropriate in this case.

Keywords: COVID-19, Hierarchical Dirichlet Process, Latent Dirichlet Allocation,Text Mining

1 Introduction

At the end of 2019, a type of virus emerged by the novel coronavirus, namely Coronavirus Disease 2019 (COVID-19). COVID-19 is a new type of disease that has never been previously identified in humans, where the virus that causes COVID-19 is called Sars-CoV-2. On December 31, 2019, the WHO China Country Office reported pneumonia of unknown etiology in Wuhan, Hubei, China. Until one month later, on January 30, 2020, the WHO has announced that this disease is a Public Health Emergency of International Concern (PHEIC). The transmission of COVID-19 is fast and has spread to other countries. As of March 25, 2020, cases were reported in 192 countries exposed to COVID-19. Scientists in almost all countries have researched COVID-19. Many studies have been published so that the public can obtain information from the results of these studies. Most of the research results are stored in the Science Direct database.

Science Direct is a database that contains a collection of quality full-text documents checked by Elsevier reviewers. More than 1.2 million articles have been collected, accessed freely through the Science Direct database. The Science Direct database also contains thousands of research related to COVID-19. From these thousands of studies, an idea emerged to find out

AECon 2020, December 19-20, Purwokerto, Indonesia

DOI 10.4108/eai.19-12-2020.2309174

what topics are about. The appropriate analysis to use is text mining analysis, which is part of data mining analysis.

Text mining is an analysis where the data uses text data. This analysis is a branch of data mining science conducted to obtain quality information from a series of texts in a document [1-3]. In-text mining, the topic modeling technique is very appropriate to do in this analysis. The topic modeling technique is used to get topics that match a set of documents. The methods in topic modeling techniques that are still being developed to date include Latent Semantic Indexing (LSI), probabilistic Latent Semantic Indexing (pLSI), Latent Dirichlet Allocation (LDA), and Hierarchical Dirichlet Process (HDP). Research has been conducted by [4], which concluded that the Hierarchical Dirichlet Process (HDP) has better sensitivity and accuracy than the C-word and Cocitation methods. Another study concludes that LDA is a popular text mining analysis method in research from 2000 to 2017 [5]. Research on topic modeling was also carried out by [6] to determine the research subject for each document (publication) and see research trends regarding libraries, archives, and museums. In this study, the topic modeling technique will carry out on the data publication topics of COVID-19 by employing the HDP and LDA methods. The best approach is selected based on the coherence score.

1.1 Text Mining

Text mining is an analysis where the data uses text data. This analysis is a branch of data mining science conducted to obtain quality information from a series of texts in a document [1-3]. Text mining's primary focus is on how many documents can be divided into several groups based on the document type. These extensive documents must be prepared with data to be processed before grouping analysis. So before analyzing or grouping the data, it is necessary to prepare data, commonly called preprocessing data. Data preprocessing is done to prepare the text data ready for text mining analysis [7].

1.2 Pre-processing Data

Preprocessing text is the text data preparation stage that needs to be done before proceeding to the analysis. This step needs to be done first because the raw text data obtained is unstructured, and the text mining process cannot be carried out. The stages of data preprocessing are as follows.

1) Case Folding. This stage is used to change all text characters to lowercase letters (not capital letters) and eliminate numbers and punctuation marks. The procedure in the case folding step is processing the letters of the alphabet "a" to "z" so that non-alphabet characters be removed as well as punctuation and numbers [8].
2) Tokenizing. It is the stage of deciding word for word in a sentence. This stage aims to break the sentence into word pieces to break the strings' sequence into pieces of their constituent words [9].
3) Lemmatizer. It is the stage for obtaining essential words. The procedure of this lemmatizer stage is to remove affixes in words. For example, the term "drugged," "drugs," "drugging" has the same root word, namely "drug," so that at this steaming stage, the word is replaced with the word "drug" [10].
4) Stopword. It is the stage of removing vocabulary that is not a unique word or does not convey any message significantly to the text. The vocabulary referred to is such as conjunctions and adverbs such as "and," "our," "from," and so on [11].
5) Topic Modeling. Topic modeling is a technique in machine learning that is classified as an unsupervised method. This technique is used to get topics from a set of documents, where this document contains text [12]. From this set of documents, statistical modeling will be

carried out to obtain the topic by first obtaining the document patterns [13]. There are several methods of topic modeling. Starting from the Latent Semantic Indexing (LSI) method in 1990, then developed into a probabilistic Latent Semantic Indexing (pLSI) in 1999. Then it was developed again until 2003 introduced Bayesian from pLSI, namely Latent Dirichlet Allocation (LDA). In contrast with the LDA, there is a method of topic modeling that also adheres to the principle of Bayesian Hierarchical Dirichlet Process (HDP) in 2005.

1.3 Hierarchical Dirichlet Process (HDP)

Hierarchical Dirichlet Process (HDP) is a method of topic modeling that uses a mixture model in its component division. In this HDP method, it is assumed that a group j will is formed, where the number of components in each group is n_k . It is also assumed that each group's data points are easy to exchange, and from that data, the point will be set up with a mixture model. Each model that is formed will have a particular proportion forming a different group. This step is needed because each group has different characteristics resulting in a different combination of proportions.

$$\begin{aligned} G_0 \mid \gamma, H &\sim DP(\gamma, H) \\ G_j \mid \delta, G_0 &\sim DP(\delta_0, G_0) \\ \varphi_{ki} \mid G_k &\sim G_k \\ x_{ki} \mid \varphi_{ki} &\sim F(\varphi_{ki}). \end{aligned} \tag{1}$$

Where G_0 is the overall probability measure, and G_k is the probability size in each group k . G_0 contains the $DP(\gamma, H)$ distribution where H is the baseline measure and γ is the concentration parameter. Whereas each group G_k is conditionally independent, containing the $DP(\delta_0, G_0)$ distribution. Also known, each x_{ki} has a factor φ_{ki} with $F(\varphi_{ki})$ the distribution. An illustration of the distribution G_0 as in Figure 1 below [14].

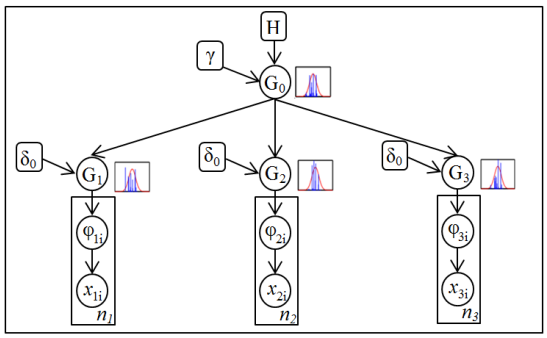

Fig. 1. HDP model illustration with three topic output

Figure 1 shows the distribution content at each point, and also illustrates the position and parameters used. It is also illustrated that the output (topic) formed is 3.

1.4 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a development method of pLSI developed by Blei et al. In 2003. This method is one of the models that has been proven effective in making models on the topics that are formed, and this method is the most frequently used method in text mining related analysis in 2000-2017 [5]. The LDA method is very representative in making topics in every document, where the topic contains a multinomial distribution of a set of words in each document. If illustrated in a diagram, it will be as in Figure 2 below

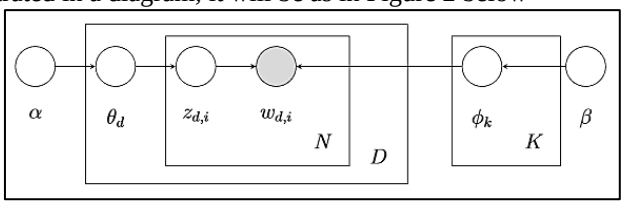

Fig. 2. Illustration of the LDA Method

Based on Figure 2, the equation for the LDA method can be written as follows.

$$
\begin{aligned}
p(w,z,\theta|\alpha,\beta) &= p(\theta|\alpha)\,p(z|\theta)\,p(\phi|\beta)\,p(w|z,\phi) \\
&= \left(\prod_{d=1}^{D}\frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k-1}\right)\left(\prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{N_{d,k,.}}\right)\times \\
&\left(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k,.})}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}-1}\right)\left(\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{N_{d,.,k,v}}\right) \\
&= \left(\prod_{d=1}^{D}\frac{\Gamma(\alpha.)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k+N_{d,k,.}-1}\right)\left(\prod_{k=1}^{K}\frac{\Gamma(\beta_{k,.})}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}+N_{d,.,k,v}-1}\right)
\end{aligned}
\tag{2}
$$

In the LDA method, it is assumed that the characteristics of a topic are determined by the distribution of the words in it. Where a collection of words referred to a document, and a set of documents called the corpus. While a collection of terms/words in the corpus, is called vocabulary. In the LDA method, the concept of the process starts from determining the number of topics, then initializing the topics randomly to the words in it. This applies to every document in the corpus. Furthermore, the calculation of the probability value of the topic in the document, as well as the probability of words on the topic. This aims to see the accuracy of the topic in the document and the probability of words on the topic.

1.5 Score Coherence

The grouping results need to be evaluated to know the consistency of the model in a grouping. The topic evaluation used is the calculated coherence score, where coherence is suitable for assessments related to the topic's quality [15]. There are two methods of calculating the coherence value, namely measurement using UCI and UMass. UCI measurement is done by calculating Pointwise Mutual Information (PMI) between two words. Equation 3 describes the calculation of the score for the UCI measurement as follows.

$$
score(v_i, v_j, \varepsilon) = \log\frac{p(v_i, v_j)+\varepsilon}{p(v_i)\,p(v_j)} \tag{3}
$$

Meanwhile, the UMass measurement is done based on the appearance of words in the document being modeled. The score calculation for UMass measurement is as in Equation (4) below.

$$score\left(v_i, v_j, \varepsilon\right) = \log \frac{D\left(v_i, v_j\right) + \varepsilon}{D\left(v_j\right)} \quad , \tag{4}$$

Where v_i is the word (*i*) and v_j is the word (*j*), $I \neq j$. $p\left(v_i\right)$ is the probability of the word, which is obtained from the frequency of occurrence of the word (*i*) in the full document. $p\left(v_j\right)$ The word probability is obtained from the frequency of occurrence of the word (*j*) in the full document [16]. Meanwhile, for the UMass measurement, it is known that $D\left(v_j\right)$ it is the number of documents that contain at least one word (*i*) and $D\left(v_i, v_j\right)$ is the number of documents containing at least one word (*i*) and one word (*j*) [17]. ε It is a smoothing factor that aims to form a real number. Some sources use one as a substitute ε . The calculation of the coherence value is explained as in Equation 5 below.

$$coherence\left(V\right) = \sum_{\left(v_i, v_j\right) \in V} score\left(v_i, v_j, \varepsilon\right) \tag{5}$$

Equation (5) states that the coherence score is the sum of the scores on each formed topic.

2 Method

The data used in this study is a set of abstracts obtained from research journals that have been published in the Science Direct database. Data collection in the Science Direct database was carried out in journals published between January 2020 and August 2020. Data was collected by scraping it on the official Science Direct website. Not all publishers were included in this analysis, but only 35 publishers because only specific articles contain abstracts that can be used as data in this study. Apart from the type of article, the site score is also a consideration. Only publishers with a cite score of at least one are selected. The keywords used in the determination of the journal are "COVID-19", "2019-nCoV", "SARS-CoV-2", or "SARS-COV-2". From the journals that have been obtained, the abstract collection is done by scraping each publication from the selected publisher. The abstract scraping results are then carried out with preprocessing in several stages, i.e., case folding, tokenizing, lemmatizer, and stopword. After obtaining the appropriate data, topic modeling analysis was carried out using the HDP and LDA methods. An investigation is carried out until it is found that the optimal coherence score between the two methods and the optimal number of topics.

3 Result and Discussion

The data used is data from scraping on publications related to COVID-19. The scraping results found that several published journals do not contain abstracts such that they were excluded from the analysis. There are 4451 publications obtained from scrapping, but only 2264 journal publications contain abstracts. These 2264 documents will be analyzed further.

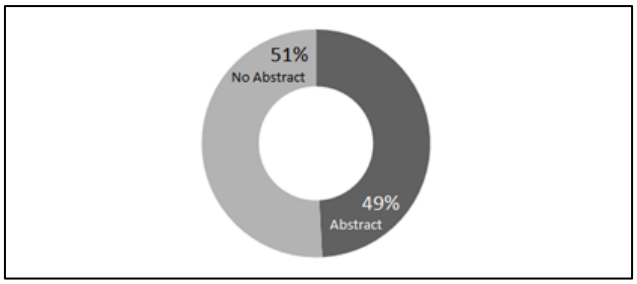

Fig. 3. The presence of abstract from scrapped documents

The document containing abstracts were preprocessed with the case folding, tokenizing, lemmatizer, and stopword stages. This preprocessing stage is carried out by changing all letters to non-capital letters, then removing punctuation marks. Furthermore, the sentence breakdown is carried out into words to form word fragments in each sentence. Next, transform the words into essential words. Having obtained the essential words, then remove some terms such as "and," "our," "from," etc. Any other words that are not needed are omitted, i.e., 'covid-19', '2019-novel', 'sars-ncov-2', etc. were also removed from the data.

The data has been through preprocessed cleansing so that the data is in the appropriate state. The data is ready to be analyzed to perform topics modeling using the HDP method, resulting in a coherence score as the graph in Figure 2.

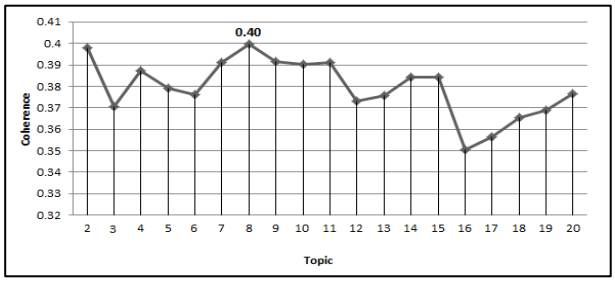

Fig. 4. Coherence Score for HDP Method

Figure 4 shows the coherence score for each number of topics from 2 groups up to 20 groups. From all the groups tested, it can be seen that the peak in the number of groups is 8. Using the HDP method, the best number of groups is eight groups, indicated by the highest coherence score. Furthermore, LDA method, coherence score results are known as in Figure 3.

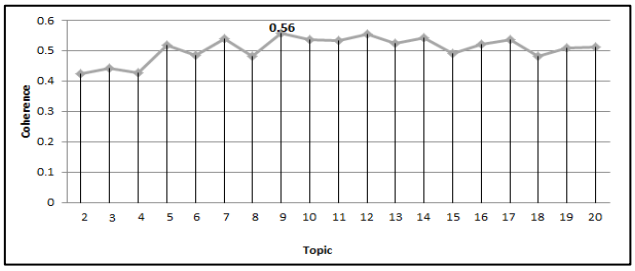

Fig. 5. Coherence Score for LDA Method

Figure 5 shows that using the LDA method can determine the best number of topics in nine topic groups, where the coherence score is 0.56. From the HDL and LDA methods that have been done, the two approaches can be compared to determine which method gives optimal results.

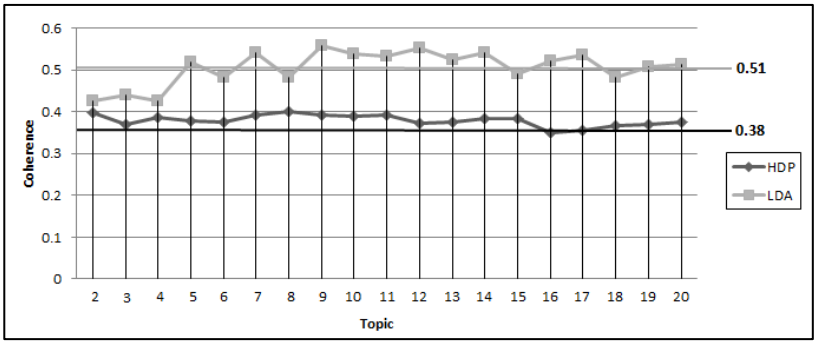

Fig. 6. Comparison Topic Modeling Method

Figure 6 shows the comparison of the score coherence between the HDP and LDA methods. From Figure 4, it can be seen that the average coherence score of the LDA method is higher than HDP. It can be seen that the average coherence score of the LDA method is 0.51, while the HDP score is 0.38. So that from the coherence score, it can be determined that the LDA method is more appropriate to use in the abstracts of journal publications related to COVID-19. The optimal number of topics is nine, where the coherence score on the number of nine topics with the HDP method is 0.39, while the LDA method is 0,56. So it can be seen that in the number of topics 9, the LDA method is 42% higher than the HDP method. Using the LDA method, we can find out the ten words most often appear in the document for each topic, which can be written in Table 1 below.

Table 1. Allocation Topic by LDA Method

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
el	data	case	data	material	death	infection	positive	immune
Los	disease	clinical	propose	surface	study	respiratory	result	ace2
SMD	care	severe	dataset	manufacturing	rate	drug	negative	receptor
pacientes	measure	respiratory	predict	force	number	disease	case	cytokine
growth	country	symptom	method	part	lockdown	may	detection	lung
GC	case	report	result	energy	country	potential	RT-PCR	response
se	impact	infection	process	manufacture	confirm	cause	antibody	increase
p	system	associate	analysis	tool	increase	human	assay	storm

Based on table 1, it can be seen that each topic has its characteristics. Ten words that often appear on each topic can be used as material to determine the topic's theme. As the first topic can be summarized as "growth," the second topic is "impact," and the following topics are "symptoms," "methods and analysis," "manufacturing tools," "the increasing number of confirmed deaths," "virus detection," and " body resistance. " In contrast to the LDA method, the HDP method is quite challenging to determine the theme of each topic. This result happened because the ten most frequently occurring words from the HDP method do not have their respective characteristics, so they are almost the same in each topic. This is a drawback of the HDP method, in addition to the relatively small coherence score. So that, indirectly, it can be seen that the advantages of the LDA method are the relatively high coherence score and the ease of knowing the characteristics of each topic that is formed.

4 Conclusion

Based on the empirical results, this research concludes that in this case, the LDA method has a coherence score that is 42% higher than the HDP method. Besides, the LDA results are more comfortable determining the theme of each topic than the results from the HDP method. Also note that the best number of topics is nine topics, i.e., "growth," "impact," "symptoms," "methods and analysis," "manufacturing tools," "increase in confirmed death count," "virus detection," and "immune system." The theme of these topics can be used as input for researchers to focus on the further development of research related to COVID-19.

References

[1] Mayasari RW, Fithriasari K, Iriawan N, Winahju WS. Surabaya Government Performance Evaluation Using Tweet Analysis. Matematika 2020;36:31–42. https://doi.org/10.11113/matematika.v36.n1.1176.

[2] Qomariyah S, Iriawan N, Fithriasari K. Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. AIP Conf Proc 2019;2194. https://doi.org/10.1063/1.5139825.

[3] Kurniawan T. Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier Dan Support Vector Machine Media Mainstream Menggunakan Naïve Machine. IT J 2017;23:1.

[4] Liu X. Full-Text Citation Analysis : A New Method to Enhance. J Am Soc Inf Sci Technol 2013;64:1852–63. https://doi.org/10.1002/asi.

[5] Liu X. Full-Text Citation Analysis : A New Method to Enhance. J Am Soc Inf Sci Technol 2013;64:1852–63. https://doi.org/10.1002/asi.

[6] Modeling T. 박물관간의 연구 주제 분석 * n.d.;50:339–58.

[7] Feldman R and Sanger J 2007 The text mining handbook: advanced approaches in analyzing unstructured data (New York: Cambridge University Press)

[8] Made N, Lestari A, Gede IK, Putra D, Ketut AA, Cahyawan A. Personality Types Classification for Indonesian Text in Partners Searching Website Using Naive Bayes Methods. Int J Comput Sci Issues 2013;10:1–8.

[9] S V, R J. Text Mining: open Source Tokenization Tools – An Analysis. Adv Comput Intell An Int J 2016;3:37–47. https://doi.org/10.5121/acii.2016.3104.

[10] Of IJ. RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS A BRIEF SURVEY OF VARIOUS APPROACHES FOR FEATURE OF TEXT 2016;4:1–8.

[11] Dragut E, Fang F, Sistla P, Yu C, Meng W. Stop word and related problems in web interface integration. Proc VLDB Endow 2009;2:349–60. https://doi.org/10.14778/1687627.1687667.
[12] Dragut E, Fang F, Sistla P, Yu C, Meng W. Stop word and related problems in web interface integration. Proc VLDB Endow 2009;2:349–60. https://doi.org/10.14778/1687627.1687667.
[13] T Witter : T Opic M Odelling and U Ser 2019.
[14] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. J Am Stat Assoc 2006;101:1566–81. https://doi.org/10.1198/016214506000000302.
[15] Stevens K, Kegelmeyer P, Andrzejewski D, Buttler D. Exploring topic coherence over many models and many topics. EMNLP-CoNLL 2012 - 2012 Jt Conf Empir Methods Nat Lang Process Comput Nat Lang Learn Proc Conf 2012:952–61.
[16] Newman D, Noh Y, Talley E, Karimi S, Baldwin T. Evaluating Topic Models for Digital Libraries Categories and Subject Descriptors. Jcdl 2010:215–24.
[17] Mimno D, Wallach HM, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. EMNLP 2011 - Conf Empir Methods Nat Lang Process Proc Conf 2011:262–72.