IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.17, No.4, October 2023, pp. 441~450
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.86025

Identify Reviews of Pedulilindungi Applications using Topic Modeling with Latent Dirichlet Allocation Method

Layli Hardiyanti*[1], Dina Anggraini[2], Ana Kurniawati[3]
[1,2,3]Department Information Systems; Universitas Gunadarma, Jawa Barat, Indonesia
e-mail: *[1]**hardiyantilayli@gmail.com**, [2] dina_anggraini@staff.gunadarma.ac.id,
[3]ana@staff.gunadarma.ac.id

Abstrak

Kemunculan covid-19 sejak desember 2019 telah mengubah tatanan kehidupan di seluruh dunia termasuk Indonesia. Beragam upaya dilakukan pemerintah guna mengendalikan situasi pandemi ini. Salah satunya dengan membuat aplikasi bernama PeduliLindungi. Aplikasi ini diharapkan menjadi sesuatu yang dapat diandalkan pemerintah dan digunakan seluruh masyarakat di masa pandemik. Menjadi aturan baru, penggunaan aplikasi PeduliLindungi memunculkan banyak ulasan dalam penilaian kualitas dan kinerja dari aplikasi tersebut. Dengan muncul dan berkembangnya aplikasi ini, menghasilkan banyak topik yang sering diulas dikalangan masyarakat dan menjadi trend. Hal tersebut dapat diidentifikasi melalui ulasan pengguna aplikasi PeduliLindungi menggunakan algoritma Latent Dirichlet Allocation (LDA). Data diperoleh dari Google Play Store dan melalui tahap pre-processing sebanyak 15.522 data. Pada Tahap processing dilakukan pembuatan dictionary dan corpus, penentuan jumlah topik dan pemodelan topik algoritma LDA. Data berupa frekuensi kemunculan kata menjadi parameter kemudian dimodelkan menggunakan metode LDA. Proses pemodelan topik menghasilkan 10 topik terbaik. Hasil pemodelan topik divisualisasikan ke dalam bentuk wordcloud dan pesebaran topik bahasan yang mewakili kumpulan ulasan masyarakat tentang aplikasi PeduliLindungi yang paling banyak dibicarakan oleh pengguna. Topik-topik dikatakan baik karena pada setiap topiknya tidak memiliki keterkaitan atau kemiripan.

Kata kunci*— Topic Modeling, COVID-19, PeduliLindungi, Latent Dirichlet Allocation.*

Abstract

The emergence of Covid-19 in December 2019 has disrupted life worldwide, including Indonesia. The government has made various efforts to control the pandemic, one of which is the development of an application called PeduliLindungi. This app aims to be a reliable tool for the government and the entire community during the pandemic. As a new regulation, the use of PeduliLindungi has prompted numerous reviews assessing its quality and performance. With the app's emergence and growth, various topics have emerged and become trending among the public. These topics were identified through user reviews of the PeduliLindungi app, using the Latent Dirichlet Allocation (LDA) algorithm. The data, consisting of 15,522 reviews, was collected from the Google Play Store and underwent pre-processing, including dictionary and corpus creation, determining the number of topics, and modeling with LDA. The resulting topic modeling process generated the ten most prominent topics. The outcomes were visualized using word clouds and topic distribution graphs, representing the most discussed aspects of the PeduliLindungi app among users. These topics are considered diverse since each issue has no relation or similarity to one another.

Keywords— *Topic Modeling, COVID-19, PeduliLindungi, Latent Dirichlet Allocation.*

Received June 21th,2023; Revised July 27th, 2023; Accepted October 30th, 2023

1. INTRODUCTION

Technology, information, and communication developments significantly affect the government's and the public's awareness of information. Indirectly, this encourages researchers to develop the latest technology so that the delivery and processing of information between humans can be done more quickly and accurately. Advances in technology, knowledge, and communication (ICT) are also accompanied by advances in mobile application devices or mobile apps; as a result, the process of delivering information, communication, and other activities becomes faster, more efficient, and more economical. The government is pushing for digitizing services and how to achieve connectedness so that integration can be created in the administration of public services [1]. Advances in technology, information, and communication today drive this. Not only in terms of public services but the government's adaptation process in terms of technology is also carried out for various purposes, one of which is to track and tackle outbreaks of dangerous diseases (pandemics) that have hit the country so that they can be overcome quickly and accurately.

Since the emergence of Covid-19 in early December 2019, various efforts have been made by the government to control this pandemic situation, one of which is by creating an application called PeduliLindungi. The PeduliLindungi application is expected to become a platform that the government relies on and is used by the community amid a pandemic. Initially, the PeduliLindungi application appeared due to an emergency, where this application was able to track community activities (tracking), tracing, and restrictions (fencing) amid the co-19 pandemic [2]. Over time, this application has also developed its uses, including as a warning and surveillance, to find out the results of Covid-19 tests that have been carried out, proof of access to public services, and downloading, informing zoning and crowd notifications, health checks, and accessing vaccine certificates.

With the emergence and development of the PeduliLindungi application, many topics often discussed among the public have emerged and have become a trend. Therefore, this research was conducted to identify user reviews about the PeduliLindungi application. Topic Modeling is a statistical example for determining core topics based on a set of documents [3] and using the identification of topic modeling to find out the current issues of the case as people's views regarding the present PeduliLindungi application to form more concise information with a broader scope.

The LDA method in research conducted by Alif (2020) [3] regarding the implementation of topic modeling using the LDA method on the thesis abstract data of the English Literature study program UIN Sunan Ampel Surabaya using 584 theory abstract data obtained the result that the words in cluster topics according to topic division according to concentration in the English Literature Study Program (UINSA). Another study by Bagus Wicaksana (2020) [4] was used to analyze the topic of Twitter user data regarding the application "Ruang Guru" using the latent Dirichlet allocation topic model to obtain the result that the LDA clustering of the Ruangguru application succeeded in grouping Twitter data into 28 topics with the most frequently discussed topics being Ruangguru discount. Another study by Bobi (2020) [5] regarding topic modeling in the article data of researchers and PDUPT emergency fund recipients using gensim obtained the results of grouping them into 18 topics with a coherence value of 0.56. In this study, the data used were research titles/articles on the Google Scholar website from researchers belonging to the PDUPT research in the last three years, with 181,326 titles and 3,894 researchers.

The LDA method in this study is chosen because it is a form of text mining to determine a pattern in a document. Besides being able to summarize, group, and connect, this LDA method can process substantial amounts of data. The LDA method is also a topic modeling algorithm with a generative probability model for document collections.

2. METHODS

2.1 Topic Modeling

Topic modeling is a technique used to find keyword representations of documents. The discovered keywords are then utilized for indexing and retrieving documents according to the users' needs [6]. Topic modeling is a form of statistical modeling employed to identify abstract topics or objects within a document. In the creation of clustering models, several processes are undertaken to obtain the best model. The selection of the topic modeling model is accomplished using the Latent Dirichlet Allocation (LDA) method. An example of a topic model utilized for text classification within a specific topic in a document is Latent Dirichlet Allocation (LDA) [7]. The process of topic modeling aims to acquire the word distribution that forms a topic and documents with specific topics, which are then used to create the topic model based on predetermined variables in the previous stage. Input parameters, namely the number of topics, the number of words in a topic, and the number of iterations, are required for constructing the topic model. When determining the number of topics, coherence values are calculated from the text data obtained from the dictionary and corpus to obtain the optimal number of topics

2.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is one approach or method that can be used to analyze large-sized documents. LDA is utilized for clustering, document summarization, and handling large-scale data. This is because LDA produces a list of topics with weights for each document [8]. The Dirichlet distribution is employed to obtain the topic distribution per document. In the generation phase, the output obtained from the Dirichlet distribution is used to allocate several words in the document to different topics. In LDA, documents are observed objects, and the topic distribution of each document and the classification of words for each document's topics are hidden structures [9]. According to Blei (2003), LDA is a generative probabilistic model of texts commonly known as a corpus. The fundamental idea behind the LDA approach is that each document is represented as a random mixture of hidden topics, where each topic has its own selected characteristics based on the distribution of words within it [10]. Blei describes the LDA method as a probabilistic model, as shown in Figure 1 below.

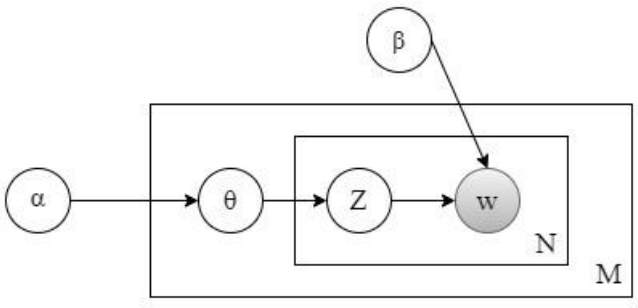

Figure 1 LDA Model Representation

When:

β : Dirichlet parameter to calculate word distribution on topic
α : Dirichlet parameter to calculate the distribution of topics to documents
w : Word in a document
N : A group of words
θ : Distribution of topics in a corpus
Z : topic of a particular word contained in a document

In Figure 1, as shown by Blei (2003), the representation of the LDA model consists of multiple levels. The parameter θ, which represents the topic distribution, is at the corpus level, which encompasses the collection of documents M. The parameter α determines the distribution of topics within a document, and a higher value indicates a larger number of mixed topics discussed in a document.

On the other hand, the parameter β governs the distribution of words within a topic, where a higher value leads to a larger collection of words in a topic. Conversely, a smaller value results in fewer words, making the topic more specific. The variable θ, which operates at the document level (M), represents the topic distribution within that document. A higher value signifies the presence of more topics in the document, while a lower value indicates a more detailed or specific topic. In the variables z and w, which operate at the word level (N), z represents a specific topic within a document, and w represents a word associated with a specific topic in a document. Based on the previous explanation, the generative process in LDA allows for correspondence between the joint distribution of hidden variables and observed variables. The following equation represents the probability or likelihood of a corpus based on the aforementioned notation.

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

In equation 1 above, the variables represent the following:

- θ represents the topic variable, where each θ is a distribution over a set of words.
- The variable α represents the document-level variable with one sample per document, representing the topic proportions for each document to variable d.
- The variables Z and w represent the word-level variables, with one sample for each word in the document. The variable z represents the specific topic assignment for each word, while w represents the actual word.

In summary, the equation captures the distributions of topics and their assignments at different levels: topic-level, document-level, and word-level.

2.2 Topic Coherence

Topic modeling or topic modeling explains the collection of words from a document or corpus. Based on a term in the paper used, the retrieval of topic relations is carried out with the assumption that a document contains a small set of summarized topics, in which the case requires correlation with human interpretation [11].

A set of words generated on a model topic with a value based on the level of coherence or in human translation or interpretation with each level of ease is referred to as Topic coherence. Having a measure of the value of a topic, Topic coherence estimates the status or rank of semantic equality between words in a case [12].

Such calculations or measurements help distinguish between a subject that can be interpreted semantically and a statistically related topic. Topic coherence is a measure or benchmark that can be used to evaluate topic modeling where if the topic coherence score is high, then the resulting model will be good [13]. Wisdom (2017) says topic coherence can be assumed to have a much better interpretation ability from topic modeling than perplexity. However, the matrix result of confusion does not have a reasonable correlation with the interpretation of the human model [7].

2.3 Data Samples

The data collection process was carried out on reviews of the PeduliLindungi software application found on the Google Play website by applying the Python programming language using the google_play_scraper package on Google Collab as many as 400,000 review data. Furthermore, review filtering was carried out so that the data generated was only from user reviews of the relevant PeduliLindungi application.

2.4 Research methods

The research method employed in this study focuses on topic modeling of user reviews of the PeduliLindungi application to determine the frequently discussed topics among users. The analysis of topic modeling in this research utilizes the Latent Dirichlet Allocation (LDA) method. The research flow can be observed in Figure 2.

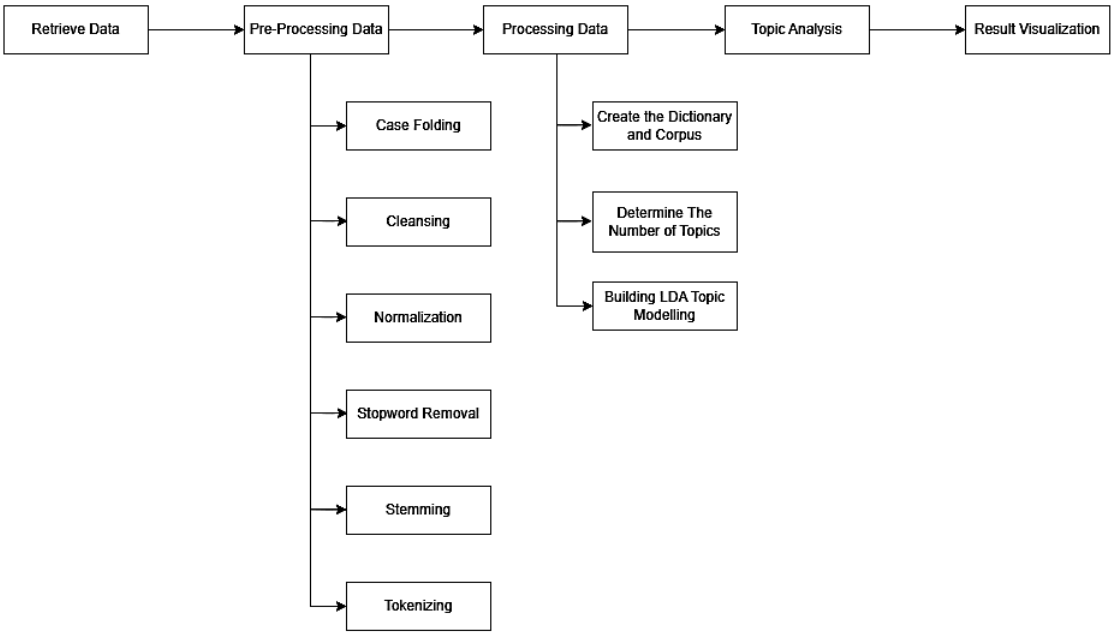

Figure 2 Research Flow

Retrieval of PeduliLindungi application review data on Google Play using the google_play_scraper package provided by Python. Furthermore, the pre-processing stage is carried out by several data processing processes such as case folding, cleaning, normalization, removing stopword, stemming, and tokenizing. Then the data processing stage, where the dictionaries and corpus are needed to model the topics, determine the number of issues, and build the LDA topic modeling. After that, the analysis of topics follows, where the topics based on the output data from the previous stage are interpreted. In this process, descriptive and informative descriptions of each topic are generated, representing the content within each topic. The final stage is the visualization of the results. This process involves creating visual representations, such as word clouds, based on the topic modeling outcomes for each topic.

3. RESULTS AND DISCUSSION

Application review data from the PeduliLindungi application on Google Play which has been taken using the google_play_scraper package provided by Python as many as 15,522 data have passed the data preprocessing stage, namely case folding, cleaning, normalization,

stopwords removal, stemming, and tokenizing, will continue at the data processing stage where, the process inside is making the dictionary and corpus needed for topic modeling, determining the number of topics, and building topic modeling.

3.1 Creating Dictionary and Corpus

Create a dictionary and corpus to turn words into numbers. Furthermore, each title will be converted into a numerical representation based on the existing dictionary in a corpus. Each sentence will be converted into a list of word IDs and the number of occurrences of the word. Furthermore, the dictionary is changed to a bag-of-words reference object to count the number of events of each word in the term or matrix of these words.

3.2 Determining the Number of Topics and measuring the value of coherence

The process of determining the number of topics is carried out using topic coherence score measurements to find the number of issues with the highest coherence values by evaluating topic modeling. Measuring the coherence value uses a limit of 20 points starting from 1 with step 1. The results obtained from assessing the coherence value are in the form of a value chart which will determine the number of topics to be used. The following graph of coherence values can be seen in Figure 3.

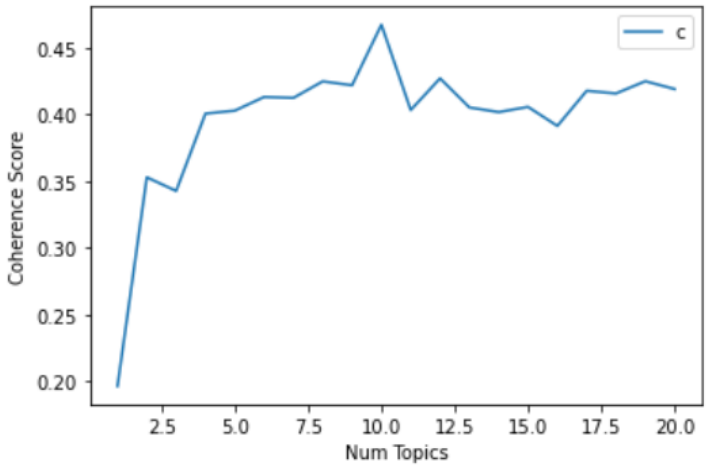

Figure 3 Coherence Value Graph

Figure 3 is a graphical visualization of table 1 below with different coherence values for each topic.

Table 1 Coherence Value

Num_Topics	Coherence Score	Num_Topics	Coherence Score
1	0.20	11	0.40
2	0.35	12	0.43
3	0.34	13	0.41
4	0.40	14	0.40
5	0.40	15	0.41
6	0.41	16	0.39
7	0.41	17	0.42
8	0.43	18	0.42
9	0.42	19	0.43
10	0.47	20	0.42

Based on table 1 shows the topic with the highest coherence value on the 10th topic. This topic has a coherence value of 0.47. This indicates that the case has the highest chance of getting the best topic model. So that it can be determined the use of the parameter, the number of topics used is 10.

3.3 Topic Modeling

Application of topic modeling using corpus, dictionary, and some topics obtained in the previous process. This process will display results in table 2 below.

Table 2 LDA Topic Modeling Results

Topic	Result
1	0.306*"vaksin" + 0.221*"sertifikat" + 0.084*"muncul" + 0.028*"cek" + 0.021*"lihat" + 0.019*"status" + 0.016*"vaksinasi" + 0.012*"merah" + 0.012*"dosis" + 0.011*"kartu"
2	0.156*"gabung" + 0.132*"daftar" + 0.085*"nomor" + 0.064*"kode" + 0.056*"verifikasi" + 0.051*"habis" + 0.049*"email" + 0.046*"akun" + 0.032*"otp" + 0.030*"kirim"
3	0.101*"tuju" + 0.050*"klik" + 0.042*"tombol" + 0.039*"negara" + 0.039*"warga" + 0.035*"ponsel" + 0.035*"pilih" + 0.034*"centang" + 0.034*"layar" + 0.033*"tampil"
4	0.138*"guna" + 0.123*"syarat" + 0.096*"bijak" + 0.076*"privasi" + 0.029*"tuju" + 0.028*"centang" + 0.025*"tentu" + 0.024*"mentok" + 0.024*"tahan" + 0.015*"halaman"
5	0.112*"eror" + 0.111*"cek" + 0.061*"scan" + 0.041*"lokasi" + 0.033*"gagal" + 0.032*"mall" + 0.025*"qr" + 0.025*"lambat" + 0.025*"paksa" + 0.020*"buat"
6	0.132*"tanggal" + 0.074*"lahir" + 0.063*"mudah" + 0.060*"sulit" + 0.060*"ribet" + 0.052*"isi" + 0.037*"klaim" + 0.028*"captcha" + 0.019*"sistem" + 0.018*"salah"
7	0.141*"bantu" + 0.082*"bagus" + 0.045*"hasil" + 0.041*"terima" + 0.030*"mudah" + 0.029*"masyarakat" + 0.029*"moga" + 0.026*"covid"+ 0.024*"hilang" + 0.023*"manfaat"
8	0.152*"data" + 0.117*"nik" + 0.076*"nama" + 0.055*"sesuai" + 0.040*"salah" + 0.037*"ktp" + 0.036*"lengkap" + 0.036*"isi" + 0.034*"payah" + 0.020*"temu"
9	0.258*"aplikasi" + 0.089*"masuk" + 0.062*"buka" + 0.047*"pedulilindungi" + 0.042*"baik" + 0.038*"baharu" + 0.037*"ponsel" + 0.033*"mohon" + 0.033*"unduh" + 0.033*"pakai"
10	0.165*"susah" + 0.078*"orang" + 0.050*"catat" + 0.048*"kerja" + 0.044*"barcode" + 0.038*"perintah" + 0.029*"scan" + 0.029*"parah" + 0.028*"bug" + 0.026*"buruk"

In Topic 1 it is represented as 0.306*"vaksin" + 0.221*"sertifikat" + 0.084*"muncul" + 0.028*"cek" + 0.021*"lihat" + 0.019*"status" + 0.016*"vaksinasi" + 0.012*"merah" + 0.012*"dosis" + 0.011*"kartu". This means the top 10 keywords contributing to this topic are: ‘vaksin, ‘sertfikat, ‘muncul, ‘cek’, ‘lihat’, ‘status’, ‘vaksinasi’, ‘merah’, ‘dosis’ dan ‘kartu’ dan bobot ‘vaksin’ and the weight of 0.306. The weight reflects how meaningful the keyword is for that topic.

3.4 Topic Analysis

The analysis process is carried out by interpreting the results of topic modeling in the form of topics that have been obtained. Issues that do not have meaning are analyzed through the distribution of words obtained as a reference to produce a sentence containing some information.

The visualization of the topic modeling results displays word clouds for each topic, showcasing the words with the highest occurrence values. The visual representation of the topic modeling results for 10 topics in the form of a word cloud can be seen in Figure 4.

Figure 4 Wordcloud Visualization of 10 Topic Modeling Results

Visualization of topic distribution is used to see the distribution of words for each topic and the relationship between one case and another. Then the visualization results of the distribution of this topic are stored in the directory in HTML form, which can be seen in Figure 5.

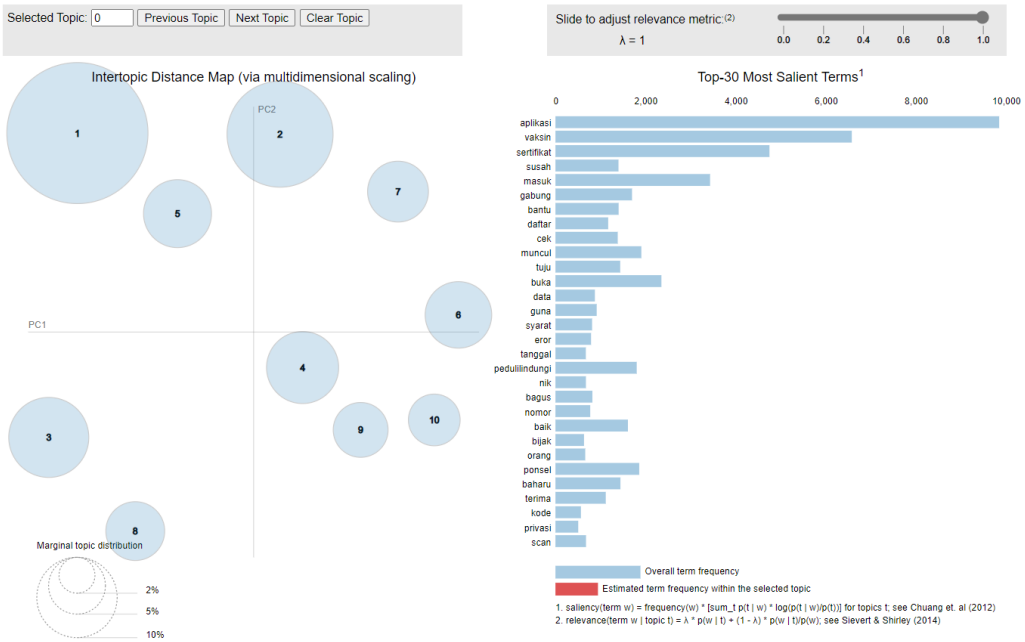

Figure 5 Visualization of Topic Distribution

In figure 5, there are each bubble represents a topic. The bigger the bubble, the higher the percentage of reviews in the corpus about that topic. Then, judging from the distance of the

bubbles, the farther apart the bubbles, the further the linkages between the topics, and vice versa. The closer the spaces of the bubbles, the closer the relations between the cases. Furthermore, the blue bar represents the number of each word in the corpus. A blue bar of the most used words will be displayed if no topics are selected. Red bars estimate the number of times a given topic generates a term.

Of the ten topics in Figure 5, you can see the bubble images for each case far apart, showing no connection or resemblance. So the resulting topic model is good. A good topic model will have giant bubbles and not overlap or overlap scattered all over the graph.

4. CONCLUSIONS

The implementation process of Topic Modeling using the Latent Dirichlet Allocation (LDA) method on the PeduliLindungi application reviews data begins with collecting data on the PeduliLindungi application reviews resulting in 15.522 relevant reviews. The data then undergoes preprocessing and processing stages to generate processed data, where the frequency of each word becomes the measure used in the LDA method for modeling.

In the LDA method, the number of topics is determined by selecting the value that yields the highest coherence score. The result of determining the optimal number of topics with the highest coherence score, which is 0.47, is obtained when using 10 topics.

Next, in the analysis process, accuracy is required in interpreting the generated topics because topic modeling with the LDA method is an unsupervised learning approach. According to the analysis results, the PeduliLindungi application user reviews data identified 10 topics, and the topic modeling results are considered good. These topics represent the most frequently discussed aspects of the PeduliLindungi application by users. The topics are considered good because each topic is distinct and unrelated to one another.

REFERENCES

[1] Doni. (29 Sep 2021). "Lewat Digitalisasi Pemerintah Integrasikan Pelayanan Publik Berbasis Elektronik," 29 Sept 2021 [Online]. Available: https://www.kominfo.go.id/content/ detail/37259/lewat-digitalisasi-pemerintah-integrasikan-pelayanan-publik-berbasiselektronik/0/ berita [Accessed: 20-Oct-2021]

[2] Indonesia.go.id. (19 Agustus 2021). "Kominfo dan Gojek Perluas Akses Pedulilindungi,". [Online]. Available: https: https://www.indonesia.go.id/ragam/komoditas/ekonomi/kominfo-gojek-perluas-aksespedulilindungi?lang=1?lang=1?lang=1?lang=1#:~:text=PeduliLindungi%20memiliki=PeduliLindungi%20memiliki%20fungsi%20tracing%20%28penelusuran%29%2C%20tracking%20%28pelacakan%29%2C%20dan,Terkait%20privasi%2C%20PeduliLindungi%20sangat%20memperhatikan%20kerahasiaan%20pribadi%20pengguna [Accessed:11-Nov-2021]

[3] Alfanzar, I, Khalid, and Rozas, I. "Topic Modelling Skripsi Menggunakan Metode Latent Dirichlet Allocation. Jurnal Sistem Informasi," JSil (*Jurnal Sistem Informasi).*, vol. 7, no. 1, pp. 7-13, Mar. 2020 [Online]. Available: DOI: 10.30656/jsii.v7i1.2036 [Accesed: 08-Nov-2022]

[4] Arianto, Wicaksano and Anuraga, Gangga, "Pemodelan Topik Pengguna Twitter Mengenai Aplikasi "Ruangguru"," *Jurnal ILMU DASAR*, vol. 21 no. 2, pp. 149-154, Juli. 2020 [Online]. Available: DOI: 10.30645/j-sakti.v4i1.188 [Accesed: 08-Nov-2022]

[5] Aditya, B. (2020). "Topic Modelling pada Data Artikel Peneliti Penerima Dana PDUPT Menggunakan *Gensim." ITS (Institue of sepuluh November*, pp. 1-20, Nov. 2020 [Online].

Available: https://123dok.com/document/zllr0wrz-topic-modelling-artikel-peneliti-penerima-pdupt-menggunakan-gensim.html [Accesed: 10-Nov-2022]

[6] Suhartono, D. "Probabilistic Latent Semantic Analysis (PLSA) untuk Klasifikasi Dokumen Teks Berbahasa Indonesia," Dec. 2014 [Online]. Available: http://arxiv.org/abs/1512.00576. [Accesed: 10-Nov-2022]

[7] Listari. "Topic Modeling Menggunakan Latent Dirchlect Allocation (Part 2): Topic Modeling with Gensim (Python)," 2019 [Online]. Availabe: https://medium.com/@listari.tari/topic-modeling-menggunakan-latent-dirchlect-allocation-part-2-topic-modeling-with-gensim-c9ffd196cb87. [Accesed: 15-Nov-2022]

[8] J. C. Campbell, A. Hindle and a. E. Stroulia, "Latent Dirichlet Allocation: Extracting Topics," 2014 [Online].

[9] D. M. Blei, "Probabilistic Topic Model," communications of the acm, vol. 55, 2012 [Online].

[10] D. M. Blei, "Latent Dirichlet Allocation," *Machine Learning Research 3*, pp. 933-1022, 2003 [Online].

[11] Putra, I. M., & Kusumawardani, R. P. (2017). Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA). *JURNAL TEKNIK ITS* Vol. 6, No. 2, A312, 2017 [Online]. Available: DOI: 10.12962/j23373539.v6i2.23205 [Accesed: 15-Nov-2022]

[12] H. S. Koh and M. Fienup, "Topic modeling as a tool for analyzing library chat transcripts," *Inf. Technol. Libr.*, vol. 40, no. 3, 2021 [Online]. Available: DOI: 10.6017/ital.v40i3.13333 [Accessed: 15-Nov-2022]

[13] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," *EMNLP 2011 - Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, no. 2, pp. 262–272, 201 [Online].