Topic Modeling in the News Document on Sustainable Development Goals

Hidayatul Fitri[1], Widyawan[2], Indah Soesanti[3]

Abstract—**Indonesia is a developing country and supports the program of the Sustainable Development Goals (SDGs) which consist of 17 goals. SDGs is not only the government's duty, but a shared duty from any elements. Online media has a crucial role in implementing goals of Indonesia's SDG. Information published in online news related to the SDGs is an important consideration for the government, society, and all elements. Categorizing news manually to find out news topics is very time-consuming and done by the ability of news editors. News presented by online media on the news site can be used as topic modeling, where hidden topics can be found in the news on online media. Topic modeling will classify data based on a particular topic and determine the relationship between text. Latent Dirichlet allocation (LDA) is one of the methods on topic modeling to find out the trend of topics of SDGs news. Based on the result of this research, the implementation of LDA is the right choice for finding topics in a document. The result of topic modeling with k = 17 obtained the highest coherence score of 0.5405 on topic 8. Topic 8 discussed news related to the eighth SDGs goals, namely decent work and economic growth. This categorization was based on words formed after the LDA process. Then, topic 5 discussed the news on the 17th SDGs goals, namely partnerships for the goals. Topic 6 discussed the news of the first SDGs, namely no poverty.**

Keywords—**Topic Modeling, LDA, SDGs, News, Media Online.**

I. Introduction

Development is a continual process that occurs in various dimensions, economic, social, and environmental dimensions. The goal is to promote welfare of the community. Even though the development is not yet perfect, it forfeits a lot of natural resources exploitation that are carried out arbitrarily, without paying attention to aspects of the environment. As a result, damage to the environment which can disrupt life occurs more frequently.

Environmental pollution cases have long been the concern of the world and other countries. The formulation of the Millennium Development Goals (MDGs) was proposed at the High-Level Conference 1972 in Stockholm, Sweden [1]. In 2000, the MDGs formulated the world development agenda which was effective until 2015. The countries presented at the conference were committed to integrating the MDGs as part of national development programs, as efforts to address the resolution of issues regarding the fulfillment of human rights and freedoms, peace, security, and development [2]. Indonesia has achieved most of the MDGs targets by achieving 49 out of 67 indicators of the MDGs, however there are still several indicators that must be implemented continuously [3].

Before the end of the MDGs, the UN Summit on MDGs 2010 formulated a world development agenda 2015. It was agreed at the general assembly of the United Nations (UN) in September 2015. Sustainable Development Goals (SDGs) agenda is ratified as a global agreement and effective from 2016 to 2030 [3].

Indonesia is one of the developing countries which becomes member of the UN and supports the SDGs program, which carries the theme "Transforming Our World: the 2030 Agenda for Sustainable Development" with 17 sustainable development goals from which are divided into 169 targets. In the context of the success of the Indonesian government's program to realize SDGs program, every element of society has an essential role to participate in making it a success [4], [5].

Online media has a crucial role in implementing the goals of the SDGs, in which news related to sustainable development in Indonesia can be a source in helping the government, society, and other elements in the development of the SDGs in Indonesia. News presented on online media related to the SDGs can be categorized based on the objectives of the SDGs program.

The goals of the SDGs are divided into three pillars which must be implemented in a harmonious and integrated manner, namely economic, social, and environmental [6]. The 17 goals of the SDGs are as shown in Fig. 1, i.e., 1) no poverty; 2) zero hunger; 3) good health and well-being; 4) quality education; 5) gender equality; 6) clean water and sanitation; 7) affordable and clean energy; 8) decent work and economic growth; 9) industry, innovation, and infrastructure; 10) reduced inequalities; 11) sustainable cities and communities; 12) responsible consumption and production; 13) climate action; 14) life below water; 15) life on land; 16) peace, justice, and strong institutions; and 17) partnerships for the goals [7].

Categorizing news manually to find out news is very time-consuming and is done by the ability of news editors. This problem can be solved with topic modeling which will group text data based on a certain topic and find the relationship between one text and another text from a corpus. The data used in the topic modeling is data that does not have a label and is included in unsupervised learning [8].

Topic modeling looks for what is contained in a large corpus and collects documents that have the same meaning and latent variables, namely knowing the thematic structure of a large corpus. The topic modeling consists of entities, namely "word", "document", and "corpora". The word is the basic unit of discrete data in a document, it is an item that is assigned an index for each unique word in the document. The document is an arrangement of N words. Corpus is collection M document, while corpora is plural form of corpus [9].

By implementing latent Dirichlet allocation (LDA), it is possible to know the trend of topics. The topic on SDGs news in online media from the news documents is obtained. LDA is one of the topic modeling methods that is used to determine the

[1,2,3] *Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Jln. Grafika No.2, Kampus UGM, Yogyakarta 55281 INDONESIA (phone: 0274-552305; e-mail: [1]hidayatulfitri@mail.ugm.ac.id, [2]widyawan@ugm.ac.id, [3]indahsoesanti@ugm.ac.id)*

1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

Fig. 1 Seventeen Sustainable Development Goals (SDGs).

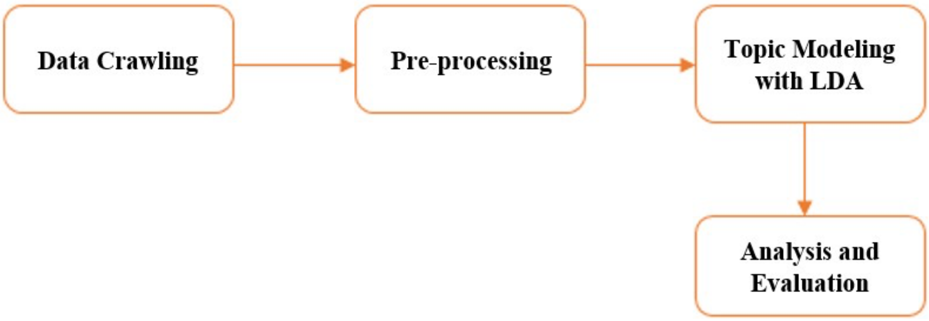


Fig. 2 Workflow of the research.

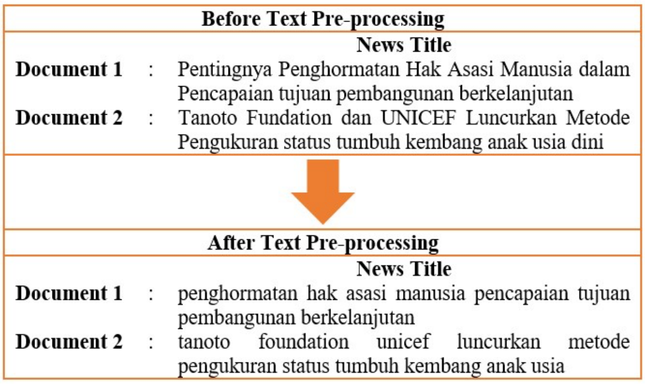
pattern in a document that produces a topic [10]. By applying topic modeling using the LDA method, the results of topic trends can be analyzed and visualized. The results of this topic modeling can be used as a reference for online media, the government, and the public in viewing the development of the SDGs news topic.

II. Topic Modeling

Research related to text mining, especially on text classification and modeling topics, has been carried out by previous researchers. Reference [11] aimed to classify news into five categories using the LDA method. The result showed that the best overall accuracy was 70%. Researchers also tried to compare with Naïve Bayes at 5-fold with the highest performance. LDA gives better results than Naïve Bayes. Reference [12] identified the key terms of each word in the tweet by applying two methods, namely LDA and Naïve Bayes. This research results indicate that Naïve Bayes is the most appropriate classification algorithm compared to J48 and KNN.

Reference [13] combined algorithms by comparing the effects of eight distance measures for document clustering using LDA+K-Means. Experiments were carried out on two datasets. The experiments showed that LDA+K-Means could increase the effect of clustering results by selecting the appropriate value of the number of topics for LDA and probabilistic-based distance measures for K-Means. Reference [14] applied modeling topics with LDA techniques from news portals related to large-scale social restrictions (PSBB) policies sourced from 9 online media. Determining the number of topics to be evaluated using the coherence and prevalence values of each topic. The results showed that news could be grouped into four major groups.

Reference [15] visualized the results of modeling the topic of the research field to identify the LDA method in conducting topic modeling analysis on the research title and finding the coherence value of each topic. As a result, 94.1% of respondents stated that the topic modeling was very good and 5.9% stated that it was good. In contrast to [10], two topic modeling experiments were applied on Wikipedia articles and Twitter users [15]. Building a topic model as a topic perspective solution in finding, exploring and recommending articles, and preparing user topic models. The aim of this research can be useful for computational tool for social and business research. In this study, offline Wikipedia articles and the process of



Before Text Pre-processing

		News Title
Document 1	:	Pentingnya Penghormatan Hak Asasi Manusia dalam Pencapaian tujuan pembangunan berkelanjutan
Document 2	:	Tanoto Fundation dan UNICEF Luncurkan Metode Pengukuran status tumbuh kembang anak usia dini

After Text Pre-processing

		News Title
Document 1	:	penghormatan hak asasi manusia pencapaian tujuan pembangunan berkelanjutan
Document 2	:	tanoto foundation unicef luncurkan metode pengukuran status tumbuh kembang anak usia

Fig. 3 Text pre-processing.

TABLE I
TOKENIZING PROCESS

Document 1	Document 2
penghormatan hak asasi manusia pencapaian tujuan pembangunan berkelanjutan	tanoto foundation unicef luncurkan metode pengukuran status tumbuh kembang anak usia

modeling topics on Twitter were not implemented and images posted by users were ignored.

In this study, the trend of news topics related to the SDGs was analyzed and visualized. Furthermore, the highest coherence value of the topic was found out.

III. DATA CRAWLING

The methods used in this study were data crawling, pre-processing, topic modeling with LDA, analysis, and evaluation. Fig. 2 shows the stages carried out in this study.

A. Data Crawling

The data was obtained in this study through data crawling on online media websites. The crawling process was carried out to retrieve news titles that had been categorized by online media as SDGs news. This process used ScrapeStorm, one of the open-source software. By entering the URL of the destination media, the page from that URL is displayed and the crawling process is carried out. At this stage, filters can be performed on the columns used for research.

There were four news sources used in this study, namely *okezone.com*, *tribunnews.com*, *kompas.com*, and *detik.com*. The selection of the four news sources was based on top sites in Indonesia on *alexa.com* ranking which was accessed on April 15, 2021. The headlines obtained were 874 news stories that had been categorized based on the type of news on the SDGs.

B. Dataset Preparation

The preparation process is cleaning columns that do not affect the research process. Columns are used as attributes. This dataset preparation process is needed for data readiness before entering the text pre-processing process. After that, the dataset is ready to be processed to the subsequent stage.

C. Text Pre-Processing

Text pre-processing is a process carried out to prepare the dataset to be analyzed. The purpose of conducting the text pre-processing is so that in the text pre-processing process several stages can be done, namely:

1) Case Folding: Case folding is a process to convert sentences in text to lowercase [16].

2) Filtering: It is a stage to retrieve useful words in the text.

3) Stop Words Removal: The stop words process is conducted to remove words that are not important and have no meaning. Removing the word such as “there”, “in”, “which”, “and”, and “this” will not affect the meaning [17].

4) Tokenizing: This process is conducted to sort the words in a sentence into words called tokens.

Fig. 3 shows the results of the process of conducting text pre-processing stage. Case folding was conducted to convert all the uppercase letters in a sentence such as “UNICEF” into lowercase letters such as “unicef”. Filtering and stop words were conducted by removing words that did not affect their meaning, such as words “dalam” and “dan” in document 1 and document 2, respectively.

Table I shows the tokenization process of document 1 and document 2. This process was conducted by sorting word by word after the case folding, filtering, and stop words process. Word by word shown in Table I is called a token.

D. Word Weighting

Term frequency-inverse document frequency (TF-IDF) is an algorithmic method which is used to calculate the weight of each word by looking for how far the relationship between words or terms with documents [18]. The TF-IDF method is efficient and has an accurate result. The process carried out in this method was to calculate the value of term frequency (TF) and inverse document frequency (IDF) for each token in each document in the corpus. TF is the frequency of occurrence of a term in the document. The greater the number of occurrences, the greater the weight. On the other hand, IDF is a calculation of the terms distributed throughout the document. It shows the relationship between the availability of a term in all documents. The fewer the number of documents containing the term, the higher the IDF value.

The word weight value in the TF-IDF method is obtained from the product of the TF and IDF value. The word weight value is used as a reference value in evaluating how important a word is in a set of documents. The TF-IDF method is often used in making search engines since this method can sort documents based on the relevance of existing features in documents based on keywords. Besides being implemented in a document sorting system, the TF-IDF method is also often applied in making a summarization system, text classification system, and others.

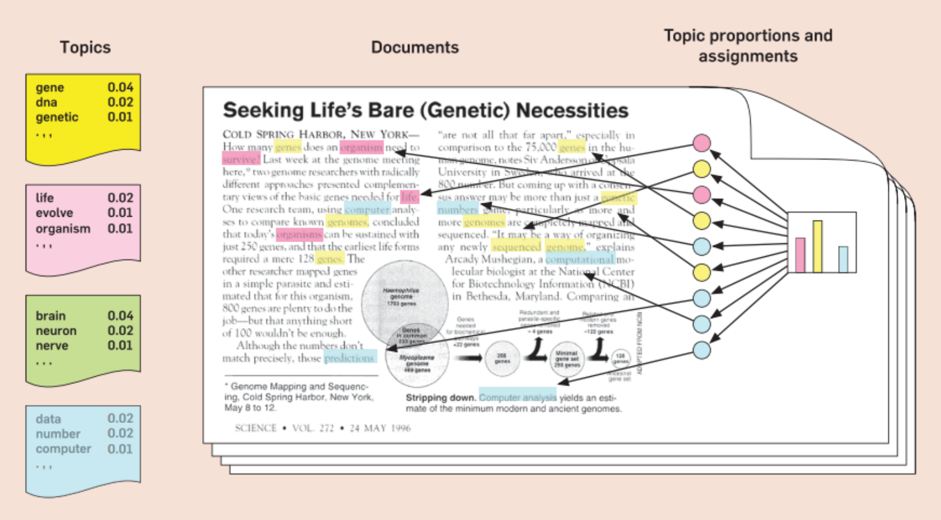


Fig. 4 How LDA works.

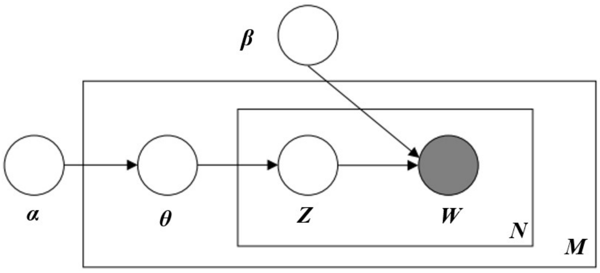


Fig. 5 Representation of LDA model.

E. *Topic Modeling with LDA*

The LDA topic modeling method is an unsupervised probabilistic model. It extracts topics from a collection of documents. A topic is defined as the distribution of vocabulary. LDA analyzes the words of each document and calculates the probability distribution between the observed and unobserved words. The LDA method uses the bag-of-words (BoW) where the semantics and meaning of sentences are not evaluated. Instead, this method evaluates the frequency of words. Therefore, it is assumed that the words that appear most often in a topic will indicate that topic.

The process of grouping documents into a certain number of topics is given a parameter when running LDA, namely by estimating the optimal number of topics. In estimating the number of topics, the cross-validation method is used to calculate perplexity and is used in information theory [19]. Fig. 4 is the workflow of the LDA process [20], while Fig. 5 is the representation of the LDA model. In Fig. 5, variable β is Dirichlet parameter for word distribution to topic, M is a document set, N is a word set, W is a word, Z is the assignment index topic, θ is a document, and α is the Dirichlet parameter for topic distribution to document.

Based on the model representation in Fig. 5, LDA has three levels which are topic distribution parameters, namely a collection of M documents at the corpus level. The higher the value, the more mix of topics in a document. The higher the value, the more words in the topic. The smaller the value, the fewer words in the topic and contains more specific words. The variable represents the distribution of topics in the document. The higher the value, the more topics found in the document [9].

F. *Support Tools*

The tool used in this research was a laptop with the following specifications:

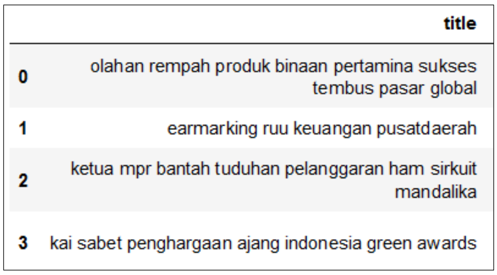
- Operating System : Windows 10
- Processor : Intel Core i7-10750H 2.60 GHz
- RAM : 16 GB
- System Type : 64-bit operating system.

IV. Result and Discussion

The preparation of this research was done by preparing the dataset to be processed. Datasets that were crawled must undergo a pre-processing stage or known as text pre-processing. The dataset used was 874 news titles related to SDGs news. This dataset was then pre-processed to clean the data.

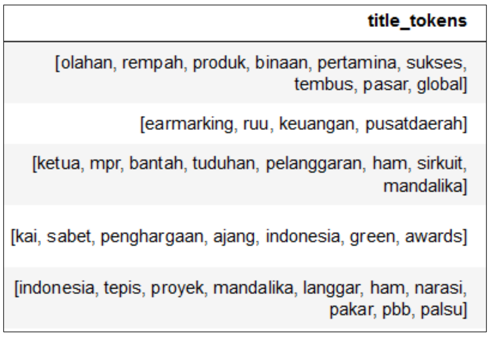
A. *Text Pre-processing*

Several stages were carried out in the text pre-processing process, namely case folding, tokenizing, filtering, and stop words removal. The case folding process is transforming all the uppercase words in a sentence to lowercase. Tokenizing is the division of the text into words or phrases called tokens. Filtering is filtering words that have no meaning or words that do not affect the research process. Stop words have the same



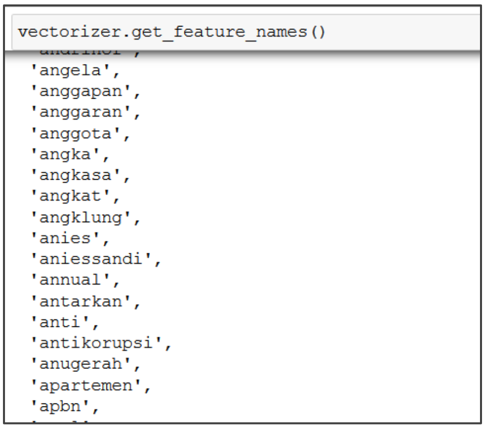
	title
0	olahan rempah produk binaan pertamina sukses tembus pasar global
1	earmarking ruu keuangan pusatdaerah
2	ketua mpr bantah tuduhan pelanggaran ham sirkuit mandalika
3	kai sabet penghargaan ajang indonesia green awards

Fig. 6 After text pre-processing.



title_tokens
[olahan, rempah, produk, binaan, pertamina, sukses, tembus, pasar, global]
[earmarking, ruu, keuangan, pusatdaerah]
[ketua, mpr, bantah, tuduhan, pelanggaran, ham, sirkuit, mandalika]
[kai, sabet, penghargaan, ajang, indonesia, green, awards]
[indonesia, tepis, proyek, mandalika, langgar, ham, narasi, pakar, pbb, palsu]

Fig. 7 After tokenizing process.



```
vectorizer.get_feature_names()
 'angela',
 'anggapan',
 'anggaran',
 'anggota',
 'angka',
 'angkasa',
 'angkat',
 'angklung',
 'anies',
 'aniessandi',
 'annual',
 'antarkan',
 'anti',
 'antikorupsi',
 'anugerah',
 'apartemen',
 'apbn',
```

Fig. 8 Word vector in the document.

function as filters. Words that are included in stop words such as the words “yang”, “di”, “dan”, “ke”, “dalam”, and others are removed [21]. In addition, the process of text pre-processing is also the removal of the delimiter which removes punctuation such as (,), (.), (:), (;), (-), (?), (!), and others.

Fig. 6 shows the results after the text pre-processing process carried out on the dataset, namely by reducing uppercase words to lowercase letters and eliminating delimiters, stop words, and filtering. Fig. 7 shows the results after tokenization process of news titles, which is separating word by word into tokens. Cleaning the text in this dataset is essential for the research process in order to obtain the text which is ready for analysis and to reduce redundant data.

B. *Word Weighting*

TF-IDF is a feature weighting method in a document where the frequency value of the word occurs in the document shows how important the word is, while the frequency value of the data containing the word shows how common the word is. The TF-IDF method is used in text classification to determine the features that affect the document to be classified.

The word or term weighting process was carried out after the pre-processing stage, where at the time of assigning weights to each term, the words had been selected, had meaning, and affected the contents of each document. Term weighting was done by using a library of sklearn by importing TfidfVectorizer of feature_extraction.text.

Fig. 8 shows the results of retrieving the vectored words in each document in the dataset. Table II shows the result of the weighting of terms and declares the value of each word in the document.

TABLE II
RESULT OF TERM WEIGHTING

Term	Term Weighting
(0, 592)	0.265397584
(0, 1334)	0.324919659
(0, 1903)	0.359164498
(0, 1839)	0.331781777
(0, 1500)	0.209728856
(0, 255)	0.308159112
(0, 1579)	0.324919659
(0, 1645)	0.412548201
(0, 1274)	0.412548201
(1, 1605)	0.553853445
(1, 938)	0.369938295
(1, 1684)	0.499638250
(1, 497)	0.553853445
(2, 1088)	0.358163349
(2, 1789)	0.379917571
(2, 623)	0.320974259
(2, 1357)	0.379917571
(2, 1991)	0.379917571
(2, 149)	0.379917571
(2, 1224)	0.342728482
(2, 936)	0.271812958
(3, 117)	0.431771794
(3, 605)	0.369838035
(3, 699)	0.196990382
(3, 26)	0.376959544

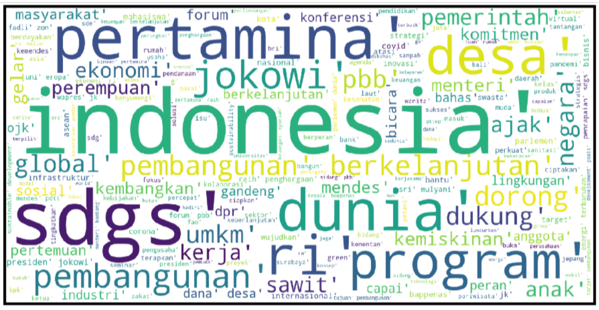


Fig. 9 Terms that often appear.

C. *Modeling*

After carrying out the word weighting process using the TF-IDF method, the next step was to model the datasets passing the pre-processing and word-weighting stages.

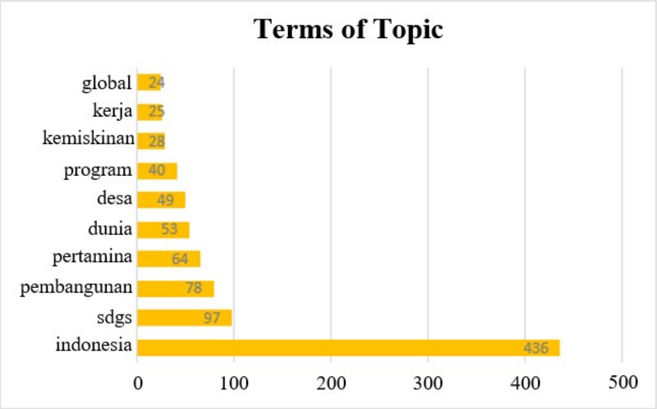


Fig. 10 Terms of topic.

This stage was done by running topic modeling to model the topic with the LDA model dataset that was modeled in the form of text. This stage was followed by the initial step that converted each text into a vector from the corpus based on tokenization with the feature extraction used, namely CountVectorizer. The occurrence of the words appeared most frequently in the SDGs news data. The data obtained is showed in the word cloud in Fig. 9. The bigger the size of letters, the more dominant and influential that words are.

Fig. 9 shows the results from word cloud. It shows the words that frequently appear in the entire document. The word “Indonesia” looks bigger and clearer than other words since it appeared in 436 documents. The word “sdgs” appeared in 97 documents and the word “pembangunan” appeared in 78 documents. Information on the number of terms that appear in the entire document is shown in Fig. 10.

The LDA topic modeling was conducted by applying the gensim library. The first step was creating a dictionary from the corpus, then represented the text into vectors using the BoW. LDA used the representation to define the topics. The determination of the number of topics was based on the 17 goals of the SDGs and 17 topics were used. Table III displays 10 topics out of 17 generated topics.

Table III shows the frequency of words from each topic. The resulting model was evaluated to see how good the modeling results were. One way to evaluate the LDA model quantitatively was generally through a coherence score. The coherence score between words in a topic was generated and the LDA model’s coherence score measured the degree of semantic similarity between words in each topic. The greater the coherence score, the better the results obtained, which shows the level of similarity in the meaning of words in each topic.

Based on the results in Table III, there are three topics that have the highest score, namely topic 8, topic 5, and topic 6. It can be seen from the frequency that each word has great value. To help in understanding the model, Fig. 11 displays the results of the coherence score. Based on the graph, topic 8 has the highest coherence score of 0.5405. The result of this coherence score was the best result obtained from the 17 number topics specified. The results of word cloud to visualize the results of topics that have the highest coherence score among other topics is also presented. The topics that have a higher coherence score are easier to interpret based on the words in them.

TABLE III
FREQUENCY OF WORDS

Topic	Frequency of Words
Topic 0	0.018*"berkelanjutan" + 0.016*"pembangunan" + 0.014*"program" + 0.012*"indonesia" + 0.010*"swasta" + 0.010*"sdgs" + 0.008*"rumah" + 0.008*"pertamina" + 0.008*"daerah" + 0.008*"pemerintah"
Topic 1	0.034*"indonesia" + 0.016*"dunia" + 0.014*"ri" + 0.012*"global" + 0.008*"konferensi" + 0.006*"kerja" + 0.006*"mendes" + 0.006*"agen" + 0.006*"industri" + 0.006*"forum"
Topic 2	0.023*"indonesia" + 0.018*"pbb" + 0.012*"forum" + 0.012*"jokowi" + 0.011*"pertamina" + 0.010*"berkelanjutan" + 0.010*"pembangunan" + 0.009*"dunia" + 0.009*"umkm" + 0.008*"dorong"
Topic 3	0.020*"program" + 0.012*"kerja" + 0.012*"sdgs" + 0.010*"laut" + 0.010*"dorong" + 0.010*"pertamina" + 0.007*"dibahas" + 0.007*"sampah" + 0.007*"anggota" + 0.007*"pangan"
Topic 4	0.017*"pertamina" + 0.012*"gelar" + 0.009*"penghargaan" + 0.007*"kota" + 0.007*"forum" + 0.007*"siapkan" + 0.007*"indonesia" + 0.006*"syariah" + 0.006*"bahas" + 0.005*"juta"
Topic 5	0.032*"sdgs" + 0.020*"indonesia" + 0.015*"pbb" + 0.011*"program" + 0.011*"desa" + 0.011*"ri" + 0.011*"global" + 0.009*"dunia" + 0.007*"goals" + 0.007*"sustainable"
Topic 6	0.032*"indonesia" + 0.016*"sdgs" + 0.014*"dunia" + 0.009*"kemiskinan" + 0.008*"jokowi" + 0.008*"wapres" + 0.008*"dorong" + 0.006*"rumah" + 0.006*"pbb" + 0.006*"anak"
Topic 7	0.015*"pembangunan" + 0.010*"sdgs" + 0.008*"capai" + 0.008*"terapkan" + 0.008*"komitmen" + 0.008*"parlemen" + 0.006*"kerja" + 0.006*"gopac" + 0.006*"fadli" + 0.006*"emas"
Topic 8	0.030*"desa" + 0.024*"sdgs" + 0.017*"pembangunan" + 0.013*"nasional" + 0.013*"indonesia" + 0.011*"mendes" + 0.009*"pdtt" + 0.009*"perempuan" + 0.009*"dana" + 0.009*"ekonomi"
Topic 9	0.011*"desa" + 0.008*"strategi" + 0.008*"infrastruktur" + 0.008*"berkelanjutan" + 0.008*"ekonomi" + 0.008*"keuangan" + 0.008*"pbb" + 0.008*"negara" + 0.008*"dana" + 0.008*"rp"
Topic 10	0.025*"ri" + 0.015*"indonesia" + 0.014*"sdgs" + 0.009*"pembangunan" + 0.008*"ajang" + 0.007*"penghargaan" + 0.007*"kemiskinan" + 0.007*"lingkungan" + 0.007*"berkelanjutan" + 0.007*"pbb"

Topic 8 had the highest coherence score. The occurrence of words in topic 8 included, “village”, “SDGs”, “development”, “national”, “Indonesia”, “Mendes”, “PDTT”, “women”, “funds”, “economy”, “support”, “government”, “programs”, “RI”, “UMKM”, “achievement”, “environment”, “develop”, and “class”. It can be seen in Fig. 12. Based on the words that appear with high scores, topic 8 has been widely reported related to village development in the economic sector to realize the goals of the SDGs.

The words that occurred in topic 5 included “SDGs”, “Indonesia”, “United Nations”, “program”, “village”, “RI”,

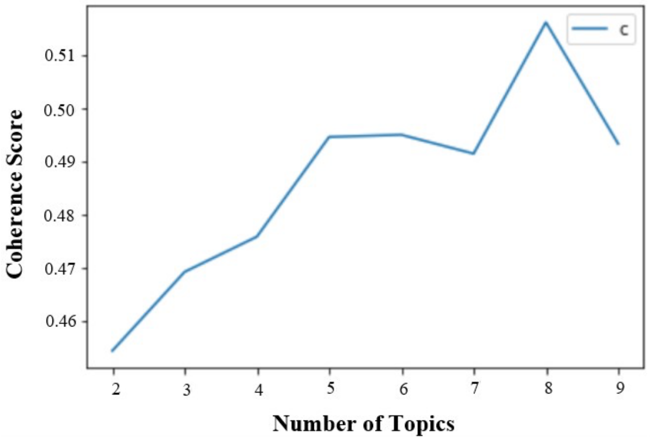


Fig. 11 Coherence score.

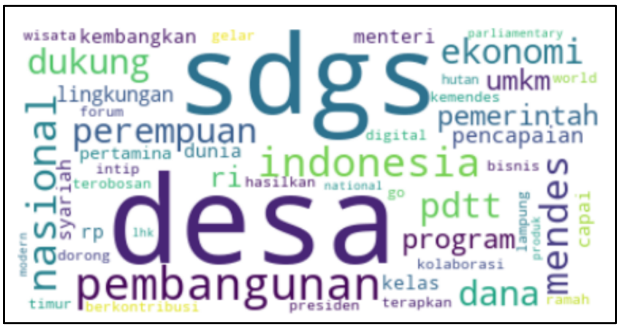


Fig. 12 Word cloud of topic 8.

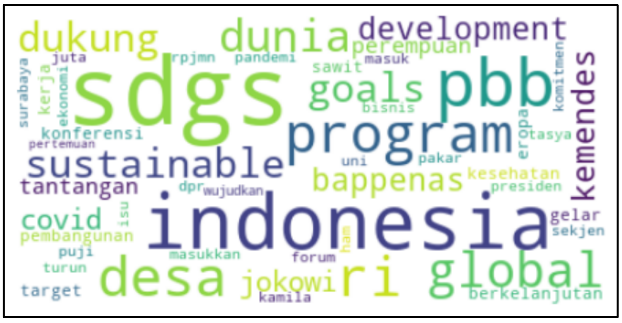


Fig. 13 Word cloud of topic 5.

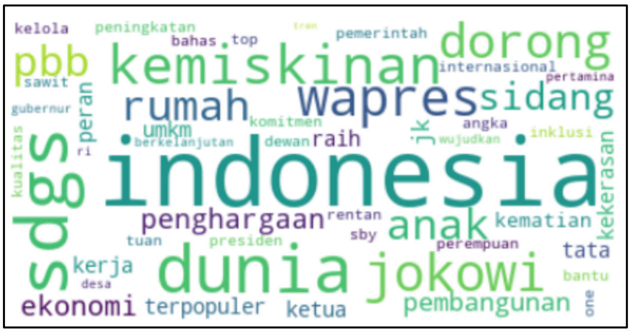


Fig. 14 Word cloud of topic 6.

“global”, “world”, “goals”, “sustainable”, “support”, “development”, “Bappenas”, “Ministry of Home Affairs”, “covid”, “Jokowi”, “challenges”, “women”, “sustainable”, and “European”. It can be seen in Fig. 13. The emergence of topic 5 is widely reported regarding the SDGs in Indonesia, which is a global UN program and is supported by various groups.

Words which appear in topic 6 included “Indonesia”, “SDGs”, “world”, “poverty”, “Jokowi”, “vice president”, “push”, “house”, “United Nations”, “children”, “session”, “awards”, “economy”, “development”, “jk”, “chairman”, “governance”, “popular”, “roles”, and “violence”. It can be seen in Fig. 14. Based on the emergence of the words, topic 6 discussed news related to poverty in Indonesia with several factors such as the economy, regional development, children, violence, and the role of the government.

V. Conclusion

On the topic modeling, LDA is the right choice for finding topics in a document. Online media can utilize this modeling to classify news according to its category. Reports related to the SDGs in several online media in Indonesia have not been discussed in detail and thoroughly.

Based on the results of topic modeling using 17 topics, topic 8 had the highest coherence score of 0.5405. Two other topics that had high scores are topic 5 and topic 6. This research could categorize news based on the goals of the SDGs including the following three topics, topic 8 discussed news related to the eighth SDGs goals, namely decent work, and economic growth. Topic 5 discussed the news on the 17th SDGs goals, namely partnerships for the goals. Topic 6 discussed the news of the first SDGs which is no poverty.

The modeling topic can be applied in finding topics that are often discussed or often appear in a document. In this study, the concept of LDA with TF-IDF was used to get the best coherence score. In the long run, it is hope that other study can apply other methods and combine them with LDA to get even more coherence scores and more datasets.

References

[1] Wahyuningsih, “Millenium Develompent Goals (MDGs) dan Sustainable Development Goals (SDGs) dalam Kesejahteraan Sosial,” *Bisma Jurnal Bisnis dan Manajemen*, Vol. 11, No. 3, pp. 390-399, Sep. 2017.

[2] A.D. Kusumawardani (2015) “Apa itu MDGs?” [Online], https://www.kompasiana.com/annisadewikusumawardani/5528a3dff17e61fa6f8b4570/apa-itu-mdgs, access date: Apr. 4, 2021.

[3] (2019) “Tentang SDGs: Apa itu SDGs?” [Online], http://sdgs.bappenas.go.id/tentang-3/, access date: Apr. 4, 2021.

[4] (2017) “Apa itu SDGs” [Online], https://www.sdg2030indonesia.org/page/8-apa-itu, access date: Apr. 4, 2021.

[5] R. Zaki (2016) “Arti Penting ‘Sustainable Development Goals’ bagi Indonesia,” [Online], https://business-law.binus.ac.id/2016/05/18/arti-penting-sustainable-development-goals-bagi-indonesia/, access date: Apr. 4, 2021.

[6] A.H. Rahardian, “Strategi Pembangunan Berkelanjutan,” *Prosiding Seminar STIAMI* , Vol. 3, No. 11, pp. 45-56, Feb. 2016.

[7] (2016) “Sekilas SDGs,” [Online], http://sdgs.bappenas.go.id/sekilas-sdgs/, access date: Apr. 4, 2021.

[8] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, “Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey,” *Multimedia Tools and Applications*, Vol. 78, No. 11, pp. 15169-15211, 2019.

[9] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research 3*, Vol. 3, pp. 993-1022, Jan. 2003.

[10] Z. Tong and H. Zhang, “A Text Mining Research Based on LDA Topic Modelling,” *The 6th International Conference on Computer Science, Engineering and Information Technology*, 2016, pp. 201-210.

[11] R. Kusumaningrum, M.I.A. Wiedjayanto, S. Adhy, and Suryono, “Classification of Indonesian News Articles Based on Latent Dirichlet Allocation,” *2016 International Conference on Data and Software Engineering (ICoDSE)*, 2016, pp. 1-5.

[12] K. Nalini and L.J. Sheela, “Classification Using Latent Dirichlet Allocation with Naive Bayes Classifier to Detect Cyber Bullying in Twitter,” *Indian Journal of Science and Technology*, Vol. 9, No. 28, pp. 3-7, 2016.

[13] Q.V. Bui, K. Sayadi, S.B. Amor, and M. Bui, “Combining Latent Dirichlet Allocation and K-Means for Documents Clustering: Effect of Probabilistic Based Distance Measures,” in *Intelligent Information and Database Systems. ACIIDS 2017. Lecture Notes in Computer Science*, Vol. 10191, N. Nguyen, S. Tojo, L. Nguyen, and B. Trawiński, Eds., Cham, Switzerland: Springer, 2017, pp. 248-257.

[14] Wahyudin, “Aplikasi Topic Modeling pada Pemberitaan Portal Berita Online selama Masa PSBB Pertama,” *Seminar Nasional Official Statistic*, 2019, pp. 309-318.

[15] Y. Sahria and D.H. Fudholi, “Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, Vol. 4, No. 2, pp. 336–344, 2020.

[16] D. Ariadi and K. Fithriasari, “Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer,” *Jurnal Sains dan Seni ITS*, Vol. 4, No. 2, pp. D.248-D.253, 2015.

[17] L.D. Utami and R.S. Wahono, “Integrasi Metode Information Gain untuk Seleksi Fitur dan Adaboost untuk Mengurangi Bias pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes,” *Journal of Intelligent Systems*, Vol. 1, No. 2, pp. 120-126, 2015.

[18] A. Deolika, Kusrini, and E.T. Luthfi, “Analisis Pembobotan Kata pada Klasifikasi Text Mining,” *Jurnal Teknologi Informasi*, Vol. 3, No. 2, pp. 179-184, 2019.

[19] C.B. Asmussen and C. Møller, “Smart Literature Review: A Practical Topic Modelling Approach to Exploratory Literature Review,” *Journal of Big Data*, Vol. 6, No. 1, pp. 1-18, 2019.

[20] D.M. Blei, “Probabilistic Topic Models,” *Communications of the ACM*, Vol. 55, No. 4, pp. 77-84, 2012.

[21] A.K. Uysal and S. Gunal, “The Impact of Preprocessing on Text Classification,” *Information Processing and Management*, Vol. 50, No. 1, pp. 104-112, 2014.