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 SDGs. To realize the 17 goals of the SDGs, online media has a crucial role for the success of the SDGs in Indonesia. Information published in online media related to SDGs is an important consideration for the government, society and all elements. Where we can get information and know the progress. The news presented by online media on their news sites can be used as a modeling topic to find out what topics are widely reported by the media. This topic modeling method uses Latent Dirichlet Allocation (LDA), where the words in the topic of each news document are represented as terms. Then it can be seen how many times the word appears in each document.**

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

I. INTRODUCTION

Development is a process that has continuity between various dimensions, economic dimensions, social and environmental which has the aim of the welfare of society. The development is not perfect yet, there are lots of exploitation resources carried out arbitrarily, without paying attention to environmental aspects. As a result, the damage to the environment occurs more and more often, which can interfere with life.

Attention to the environmental pollution case has been carried out for a long time, this is a concern of the world and other countries. Formulation of Millennium Development Goals (MDGs) holding a High-Level Conference (KTT) in Stockholm, Sweden in 1972. [1] In 2000 started the implementation of Millennium Development Goals (MDGs) who have formulated the world development program up to 2015, all countries attended the meeting committed to integrating MDGs as part of a national development program, to deal with resolving issues regarding the fulfillment of human rights and freedoms, peace, security, and development. [2] Indonesia has achieved most of the MDGs targets, from 67 which 49 indicators MDGs have achieved, there are still several indicators that must be continued in their implementation. [3]

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

Indonesia is one of the developing countries which are the members of the United Nations (UN) and supports the SDGs program, which carries the theme *"Transforming Our World: the 2030 Agenda for Sustainable Development"* with 17 sustainable development goals of the 17 goals which divided into 169 targets. In order to succeed the Indonesian government's program for the realization of SDGs program, every element of society has an important role to participate in making it a success. [4][5]

The goals of SDGs are divided into 3 pillars which must be carried out in a harmonious and integrated manner, namely economical, social and environmental.[6] as for the 17 goals of SDGs as shown in Fig. 1, i.e., 1) No proverty; 2) Zero hunger; 3) Good health and well-being; 4) Quality education; 5) Gender equality; 6) Clean water and sanitation; 7) A ffordable 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] Fig.1 shows 17 goals from SDGs achievement.

[1,2,3] *Department of Electrical Engineering and Information Technology, Gadjah Mada University, Grafika street No.2, Yogyakarta 55281 INDONESIA (phone: 0274-552-301; fax: 0274-547-506; e-mail: hidayatulfitri@mail.ugm.ac.id, widyawan@ugm.ac.id, indahsoesanti@ugm.ac.id)*

Fig. 1 Sustainable Development Goals (SDGs) 17 Goals

The study related to text mining especially on text classification has been done by many previous researchers. In research [8] aiming to classify news into five categories using the Latent Dirichlet Allocation (LDA) method. Overall, the result shows the best accuracy is 70%. Researcher also tried to compare with Naïve Bayes at 5-fold with the highest performance. LDA gives better result than Naïve Bayes. Nalini [9] perform an identification on key terms from each word on tweet by applying the two methods of LDA and Naïve Bayes. His research shows that Naïve Bayes is the most appropriate classification algorithm compared to J48 and KNN.

Vu Bui [10] combines algorithm by comparing the effects of eight distance measures for document clustering using LDA+K-Means. Experiment was carried out on two datasets. LDA+K-Means can increase the effect of clustering result when selecting the appropriate value of the number of topics for LDA and probabilistic based distance measures for the K-Means.

The study [11] using 10 news sites are used to distinguish between news article and advertisement. Produce good accuracy using Naïve Bayes algorithm. Similarly with [12] in categorizing news article, Naïve Bayes shows a better performance than other algorithms.

II. THEORETICAL REVIEW

A. Classification

Classification is a significant cognitive way for people to know the unknown world. For example, depending on their different characteristics, people classify creatures into plants and animals and further into phylum, class, order, family, genus, and species. People with dissimilar faiths are classified into different parties, sects, etc. in sociology. When similar things are place into one category in actual production and life and things with obvious differences are put into different categories, dealing with them will be much easier.

B. Text Pre-processing

Text pre-processing is a process that is done to prepare the dataset dianalysis. The purpose of doing *text pre-processing* so that in the *text-processing* process have several steps that can be done, which are:

1) Case Folding: Case folding is a process to convert sentences in the text to lowercase.[13]

2) Filtering: stages to retrieve useful words in the text.

3) Stopwords: : where the *stopwords* process is to eliminate words that are not important and have no meaning. So it does not affect the meaning when removing the word such as, “ada”, “dalam”, “yang”, “dan”, “ini”.[14]

4) Tokenizing: this process is to sort out the words in a sentence into words called token.

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

⇩

	Setelah melakukan proses t*ext pre-processing*
Kalimat 1 :	penghormatan hak asasi manusia pencapaian tujuan pembangunan berkelanjutan
Kalimat 2 :	tanoto foundation unicef luncurkan metode pengukuran status tumbuh kembang anak usia

Fig. 2 Proses *text pre-processing*.

Fig. 2 shows the process from the result of doing *text pre-processing* at the case *folding stage* by reducing all the letters in sentences such as "UNICEF" after the *case folding* process into "unicef". *Filtering* and *stopwords* are eliminating words that do not affect its meaning as in sentence 1 and sentence 2 the words "dalam", "dan" are removed.

Fig. 3 shows the tokenization process of sentence 1 and sentence 2, by sorting out word by word after the *case folding, filtering* and *stopwords* process. Word by word shown in Fig.3 is called token.

Kalimat 1
penghormatan
hak
asasi
manusia
pencapaian
tujuan
pembangunan
berkelanjutan

Kalimat 2
tanoto
foundation
unicef
luncurkan
metode
pengukuran
status
tumbuh
kembang
anak
usia

Fig. 3 Process Tokenizing.

C. Term Frequency – Inverse Document Frequency (TF-IDF)

Term frequency Inverse Document Frequency (TF-IDF) is an algorithmic method used to calculate the weight of each word by finding how far the relationship between word or term with document [15]. The TF-IDF method is efficient and has an accurate results. The process in this method is done by calculating the value of Term Frequency (TF) and Inverse Document Frequency (IDF) for each token in each document in the corpus. Term Frequency (TF) is the frequency of occurrence of a term in the document, the greater the number of occurrences, the greater the weight. While the Inverse Document Frequency (IDF) is a calculation of the terms distributed throughout the document and the IDF shows the availability relationship of a term in the entire documents. The fewer the number of documents which contain the term, the higher IDF value is.

The word weight value in TF-IDF method is obtained by multiplying between the value of Term Frequency (TF) and Inverse Document Frequency (IDF). Word weight value is used as a reference value in evaluating how important a word in a set of documents. The TF-IDF method is often used in making search engines because this method is able to 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 often applied in making summarization systems, text classification systems and more.

D. Topic Modelling

The topic modelling method LDA is an unsupervised, probabilistic modelling method which extracts topics from a collection of papers. A topic is defined as a distribution over a fixed vocabulary. LDA analyses the words in each paper and cal-culates the joint probability distribution between the observed (words in the paper) and the unobserved (the hidden structure of topics). The method uses a 'Bag of Words' approach where the semantics and meaning of sentences are not evaluated. Rather, the method evaluates the frequency of words. It is therefore assumed that the most frequent words within a topic will present an aboutness of the topic.

A key aspect of LDA is to group papers into a fixed number of topics, which must be given as a parameter when executing LDA. To estimate the number of topics, a cross-validation method is used to calculate the perplexity, as used in information theory. [16]

III. Methodology

A. Crawling Data

The data obtained in this study through data *crawling* on *online media* website. The *crawling* process is done to capture the headlines that have been categorized by *online media* as SDGs news. However, the *crawling* process using *ScrapeStorm*, one of *open source software* is by entering the URL of the destination media, the page from the URL is displayed and the *crawling* process is carried out. At this stage *filter* on the column can be performed that used for research

There are four news sources used in this study, namely okezone.com, tribunnews.com, kompas.com, detik.com, the selection of the four news sources based on the top sites in Indonesia on alexa.com ranking accessed on April 15, 2021. The headline obtained 874 news that have been categorized based on the type of SDGs news.

B. Dataset Preparation

The preparation process of dataset after the *crawling* data stage is cleaning the columns that have no effect on the research process. Columns are used as an attribute. The dataset preparation process required for data readiness before entering the text pre-processing process. Maturely, dataset is ready to be processed to the next stage.

C. Tools and Materials

Tool used in this research is a laptop with the following specifications:

- Operating System : Windows 10
- Processor : Intel Core i7-10750H 2.60GHz
- RAM : 16,0 GB
- Type System : 64-bit Operating system

IV. Result And Discussion

The preparation of this study was conducted by preparing a dataset to be processed. The dataset that has been crawling needs to perform pre-processing stage or known as text pre-processing. The used dataset is 874 headline news related to SDGs news. Dataset is then carried out in the pre-processing stage to clean up the data

A. *Text Pre-processing*

There are several stages in the *text pre-processing* process i.e. *case folding, tokenizing, filtering* and *stopwords*.

The *case folding* process change all the words in a sentence from uppercase letters to lowercase words. *Tokenizing* is a division of text into words, phrases called token. *Filtering* is filtering out the words that have no meaning or a word that has no effect on the research process. *Stopwords* have the same function as filters which remove words that included in *stopwords* such as the word "yang" , "di", "dan", "ke", "dalam" and others [17]. In addition, the *text pre-processing* process is also the removal of delimiter which removes punctuation, marks, characters such as (,), (.), (:), (;), (-), (?), (!) and more.

Fig. 4 is a result after the *text pre-processing* process was carried out on the dataset, by reducing uppercase to lowercase and eliminating *delimiter, stopwords* and *filtering*. Fig. 5 is a process after tokenization of headline news, separated word by word into token.

	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. 4 After process 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. 5 After process tokenizing

Cleaning the text on this dataset is necessary for the research process, so that the text is ready to be analyzed and to reduce redundant data.

B. *Word Weighting*

Term Frequency-Inverse Document Frequency (TF-IDF) is a feature weighting method in a document where the frequency value of the word occurrence in the document shows how important the word is and 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 is carried out after the pre-processing stage, where at the time of assigning weights to each term, the words have been selected and have meaning and affect the contents of each document.

The term weighting is done using the library from *Sklearn* by importing the *TfidfVectorizer* from *feature_extraction.text* can be seen in Fig. 6.

Fig. 7 retrieves the vectored words in each document in the dataset. Fig. 8 is the result of term weighting, and is declared the value of each word in the document.

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer()
response = vectorizer.fit_transform(data['title'])

print(response)
```

Fig. 6 Import TfidfVectorizer.

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

Fig. 7 Word vector in the document.

```
(0, 592)        0.2653975837104881
(0, 1334)       0.3249196594838025
(0, 1903)       0.3591644984064433
(0, 1839)       0.3317817768551522
(0, 1500)       0.20972885647055806
(0, 255)        0.3081591121282697
(0, 1579)       0.3249196594838025
(0, 1645)       0.412548201019983
(0, 1274)       0.412548201019983
(1, 1605)       0.5538534450271237
(1, 938)        0.36993829471315365
(1, 1684)       0.4996382501049107
(1, 497)        0.5538534450271237
(2, 1088)       0.35816334870056726
(2, 1789)       0.379917571332621
(2, 623)        0.32097425947649794
(2, 1357)       0.379917571332621
(2, 1991)       0.379917571332621
(2, 149)        0.379917571332621
(2, 1224)       0.3427284821085517
(2, 936)        0.2718129576079279
(3, 117)        0.4317717944871137
(3, 605)        0.3698380345716125
(3, 699)        0.19699038170093258
(3, 26)         0.37695954377642393
:       :
(870, 620)      0.3744610576850718
(870, 1394)     0.3346886488587038
(870, 1367)     0.21714582944704425
```

Fig. 8 Result term weighting

C. *Modeling Section*

After carrying out the word weighting process with the TF-IDF method, the next step is to model the datasets that have passed the pre-processing and word weighting stages.

This stage is done by running topic modeling to model the topic with the Latent Dirichlet Allocation (LDA) model. The dataset that will be modeled is in the form of text, so the first step is to convert each text into a vector by using the *CountVectorizer* that has been provided in the *Sklearn* library as follows:

```
from sklearn.feature_extraction.text import CountVectorizer
```

After the text is converted into a vector, then it is entered into modeling using the LDA model.

Fig. 9 Terms that often appear

Fig. 9 results from wordcloud that show the most frequently occurring words in the entire document. The word 'Indonesia' looks bigger and clearer than other words. The word 'indonesia' appears in 436 documents, the word 'sdgs' appears in 97 documents and the word 'development' appears in 78 documents. Description of the number of terms that appear in the entire document can be seen in Fig. 10.

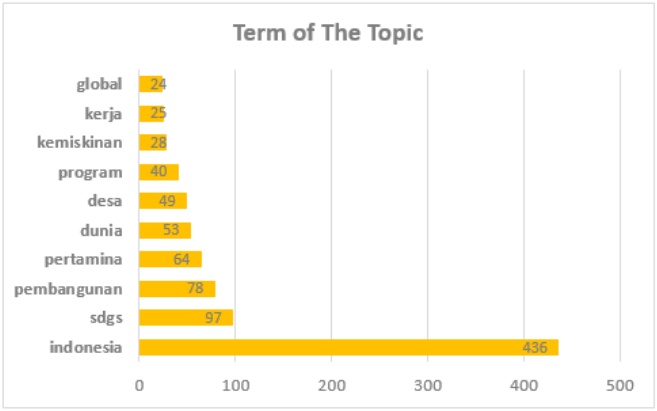

Fig. 10 Term of the topic.

By applying the LDA model, it can be seen the visualization results of each document related to the topic raised. fig. 11 is one of the results of the visualization of a document after the LDA model is applied.

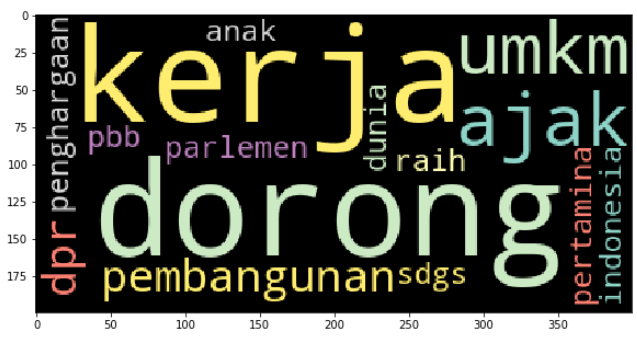

Fig. 11 Topic in document from LDA model.

V. CONCLUSION

After doing several stages, namely the data crawling stage, the pre-processing stage, the word weighting stage and the modeling stage with topic modeling. Latent Dirichlet Allocation (LDA) is the right choice for finding topics in a document. Like online media can use this modeling to classify news according to its category. Reports related to the Sustainable Development Goals (SDGs) in several online media in Indonesia, have not discussed in detail and thoroughly related to sustainable development issues in Indonesia. It can be seen from the results that the topics that often arise are related to Indonesia, SDGs, development, Pertamina and the world.

This modeling topic can be applied in finding topics that are often discussed or often appear in a document. The results will be classified.

ACKNOWLEDGEMENT

Acknowledgement for the IJITEE team for taking the time to create this template.

REFERENCE

[1] W. Wahyuningsih, "Millenium Develompent Goals (Mdgs) Dan Sustainable Development Goals (Sdgs) Dalam Kesejahteraan Sosial," *Bisma*, vol. 11, no. 3, p. 390, 2018.

[2] kompasiana, "Apa itu MDGs_ - Kompasiana.pdf," 2015. [Online]. Available: https://www.kompasiana.com/annisadewikusumawardani/5528a3dff17e61fa6f8b4570/apa-itu-mdgs. [Accessed: 04-Apr-2021].

[3] kementerian P. Bappenas, "SDGs_bappenas.pdf." [Online]. Available: http://sdgs.bappenas.go.id/tentang-3/. [Accessed: 04-Apr-2021].

[4] "sdg2030indonesia.pdf." [Online]. Available: https://www.sdg2030indonesia.org/page/8-apa-itu. [Accessed: 04-Apr-2021].

[5] R. Zaki, "Arti Penting 'Sustainable Development Goals' Bagi Indonesia," *Binus University*, 2016. .

[6] P. Seminar, S. Issn, and V. Iii, "Prosiding Seminar STIAMI ISSN 2355-2883 Volume III, No. 01, Februari 2016," vol. III, no. 01, 2016.

[7] http://sdgs.bappenas.go.id, "Sekilas SDGs," *dashboard SDGs*, 2016. [Online]. Available: http://sdgs.bappenas.go.id/sekilas-sdgs/. [Accessed: 04-Apr-2021].

[8] R. Kusumaningrum, M. I. A. Wiedjayanto, S. Adhy, and Suryono, "Classification of Indonesian news articles based on Latent Dirichlet Allocation," *Proc. 2016 Int. Conf. Data Softw. Eng. ICoDSE 2016*, pp. 1–5, 2017.

[9] K. Nalini and L. Jaba Sheela, "Classification using Latent Dirichlet Allocation with Naive Bayes Classifier to detect Cyber Bullying in Twitter," *Indian J. Sci. Technol.*, vol. 9, no. 28, pp. 3–7, 2016.

[10] Q. Vu Bui, K. Sayadi, S. Ben Amor, and M. Bui, "Combining Latent Dirichlet Allocation and K-means for Documents Clustering: Effect of Probabilistic Based Distance Measures," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10191 LNAI, pp. 248–257, 2017.

[11] S. K. Dwivedi and C. Arya, "News web page classification using url content and structure attributes," *Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016*, no. October, pp. 317–322, 2017.

[12] A. N. Chy, M. H. Seddiqui, and S. Das, "Bangla news classification using naive Bayes classifier," *16th Int'l Conf. Comput. Inf. Technol. ICCIT 2013*, no. March, pp. 366–371, 2014.

[13] J. Sains *et al.*, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," vol. 4, no. 2, 2015.

[14] L. D. Utami, "Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes," vol. 1, no. 2, pp. 120–126, 2015.

[15] A. Deolika, Kusrini, and E. T. Luthfi, "ANALISIS PEMBOBOTAN KATA PADA KLASIFIKASI TEXT MINING," vol. 3, no. 2, pp. 179–184, 2019.

[16] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *J. Big Data*, vol. 6, no. 1, 2019.

[17] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014.