© Jurnal Nasional Teknik Elektro dan Teknologi Informasi
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License
DOI: 10.22146/jnteti.v13i2.9464

Modeling Climate Change Issues in Indonesia Based on Media Headlines

Anang Kunaefi[1], Aris Fanani[2]

[1] Information System Study Program, Faculty of Science and Technology, UIN Sunan Ampel Surabaya, Surabaya, Jawa Timur 60294, Indonesia
[2] Mathematics Study Program, Faculty of Science and Technology, UIN Sunan Ampel Surabaya, Surabaya, Jawa Timur 60294, Indonesia

[Received: 1 November 2023, Revised: 19 January 2024, Accepted: 25 April 2024]
Corresponding Author: Anang Kunaefi (email: akunaefi@uinsa.ac.id)

ABSTRACT — Climate change has become a global issue affecting all countries in the last decades. This phenomenon poses a concern to Indonesia as it is one of the climate change's epicenters. Various studies have shown that climate change can harm multiple community activities, such as unstable agricultural production, decreased people's health, and global warming. This study tried to model and analyze climate change topics discussed in the media. Finding hidden topics from texts can provide clues and information regarding public conversation surrounding climate change, such as public thoughts, perceptions, and readiness to mitigate the possible adverse effects of climate change. In order to identify hidden subjects from the corpus, this work modeled climate change issues in Indonesia using the latent Dirichlet allocation (LDA) algorithm to analyze texts from Indonesian media headlines. As many as 7,000 headline data from five online media were collected from 2017 to 2021 using web scraping techniques. The proposed approach produced eight topics related to climate change, which were determined by the highest coherence value of 0.560. Those topics were renewable energy, carbon emissions, environmental management, development economics, international cooperation, policy/regulation, rehabilitation, and disaster. Based on the results, the model could sufficiently describe the theme of discussion in society and photograph public thoughts and the government's readiness in the form of policies and regulations in dealing with the climate change phenomenon.

KEYWORDS — Climate Change, LDA, Text Analysis, Text Mining, Topic Modeling.

I. INTRODUCTION

Changes in climate patterns caused by greenhouse gas emissions and other human activities that disregard the natural order are referred to as climate change. These changes have caused various natural phenomena that are detrimental to humans, such as natural disasters and global warming. In general, there is an increase in gas emissions and temperatures of 1 °C–1.5 °C worldwide. In 2018, there were 315 cases of natural disasters due to climate change, which hit more than 68.5 million people, with material losses reaching US$31.7 billion [1]. In addition, food security, water, health, and ecosystems are vulnerable to climate change.

Numerous nations have acknowledged that climate change has negative effects. To encourage all nations to have the same knowledge when establishing policies on climate change, the World Meteorological Organization (WMO) and the United Nations founded the Intergovernmental Panel on Climate Change (IPCC) in 1988 (IPCC 2013). The creation of the United Nations Framework Convention on Climate Change (UNFCCC) in 1992 was the most crucial step. This convention has been approved by 169 nations worldwide. Conference of the Parties (COP) conferences are held annually by representatives of UNFCCC member nations to discuss advancements in the fight against climate change. Berlin, Germany, hosted the first UNFCCC COP in 1995.

Indonesia, the third largest emitter of greenhouse gases in the world, has been seen as a vulnerable country to climate change [2]. Various studies have shown that climate change has negative impacts, such as decrease in agricultural production [3], [4]; food security vulnerability [5]; deforestation; and threat to environment biodiversity [6].

Therefore, a systematic effort is needed from all stakeholders, especially the government, to face the challenges of climate change. To what extent the attention and awareness of stakeholders in Indonesia towards the phenomenon of climate change can be seen and evaluated from the news coverage in the media that capture various conversation in society on a daily basis.

This study tried to model and analyze climate change topics discussed in the media to portray public thoughts and perceptions to tackle the previously mentioned threats. In addition, this study sought to determine climate change challenges and whether public discourse have been anticipated, particularly by Indonesian policymakers.

Previous studies used a text mining approach to model climate change themes from media texts. One study modeled climate change based on news media coverage in India [7], while another study analyzed the policy regarding climate change using a natural language processing algorithm [8]. In this study, however, the approach to modeling climate change issues was proposed along with a machine learning pipeline and results were discussed.

In this study, the dataset was built by scraping media headline texts from several media websites. Then, the dataset was cleaned, preprocessed, and stored for clustering algorithm. Next, the latent Dirichlet allocations (LDA) algorithm was used to analyze the texts to derive hidden knowledge from the corpus. The rationale behind employing LDA instead of alternative algorithms is its exceptional fit for short text corpora, like headline titles and user reviews.

Analyzing media content can provide clues and information about topics being discussed in the society regarding climate

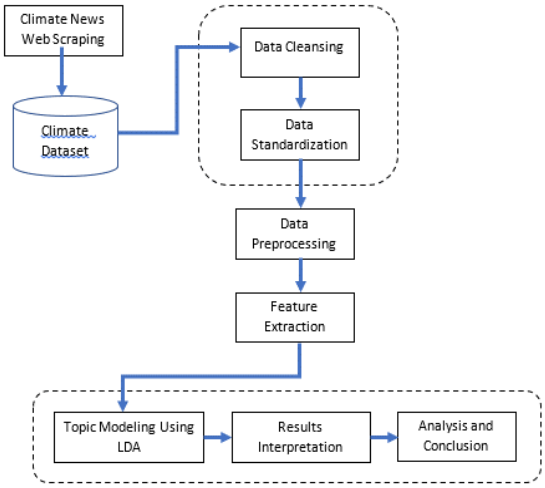

Figure 1. Research method.

change. Furthermore, the emerging issue based on the text mining model can portray Indonesia's readiness to face climate change challenges.

This paper makes two contributions. First, topic modeling algorithm was implemented based on collected Indonesian media headlines dataset. Second, the findings were examined, explained, and compared to prior research.

This paper is structured as follows: Section I provides a general overview of this research. In Section II, the methodology is thoroughly explained, along with the dataset and the experimental design. The experiment's empirical results are presented in Section III, and then a discussion of the results is presented in Section IV. In Section V, suggestions for additional research are added.

II. METHODOLOGY

In this study, the proposed research method consists of four main steps: a) data collection, b) data preprocessing, c) topic modeling, and d) topic analysis and discussion, as depicted in Figure 1. In the subsection that follows, each step of the method is explained.

A. DATA COLLECTION

The initial step in the workflow is collecting datasets, i.e., article headlines, from several national media sources. For the sake of simplicity, information headlines from online newspaper were used. Nevertheless, to guarantee the validity and integrity of the data, some criteria were applied. First, the selected online media should have a paper-based version of the newspaper, or the online media should be publicly well-known and have an excellent reputation recognized nationally. Second, the selected online media must provide access to articles published from 2017 to 2021 according to the time frame of this research. Based on these criteria, the eligible online media are Kompas, Detik, Merdeka, SindoNews, and AntaraNews, as listed in Table I.

Next, the web scraping technique is used, which is a technique for automatically searching data on websites to get the headline on websites that meet certain filters. The phrase "climate change" was used as a keyword to grasp the headline text. The total headline data collected was 7,000 records.

B. DATA PREPROCESSING

At this stage, there are two processing phases, namely data cleansing and linguistic preprocessing, each of which will be explained as follows.

TABLE I
ONLINE MEDIA DATA SOURCES

Media Name	URL	Number of Headlines
Kompas	www.kompas.com	290
Detik	www.detik.com	109
Merdeka	www.merdeka.com	50
SindoNews	www.sindonews.com	762
AntaraNews	www.antaranews.com	6,000

TABLE II
EXAMPLE OF THE DATA PREPROCESSING RESULTS

No.	Headline Text	Result
1.	"Tanam Mangrove, Pemerintah Pulihkan Lingkungan Sekitar Pantai"	"tanam," "mangrove," "pemerintah," "pulihkan," "lingkungan," "sekitar," "pantai"
2.	"BMKG Tingkatkan Akurasi Informasi Cuaca Dengan Metode Baru"	"bmkg," "tingkatkan," "akurasi," "informasi," "cuaca," "metode," "baru"
3.	"Perubahan Iklim Perbesar Kerugian Ekonomi Para Petani"	"perubahan," "iklim," "perbesar," "kerugian," "ekonomi," "petani"
4.	"Relawan Muda Untuk Bumi Yang Sehat Lakukan Penanaman Ribuan Pohon"	"relawan," "muda," "bumi," "sehat," "lakukan," "penanaman," "ribuan," "pohon"

1) DATA CLEANSING

The data obtained from the web scraping procedure frequently contain much noise. The reason is that the text data collected comes from many sources that have various formats [9]. Therefore, in this phase, the data were cleaned so that the data could be used for further processing. The data cleaning process includes three steps, namely, eliminating blank text, removing symbol characters, and standardization of data.

In terms of blank text elimination, data retrieved from the web often contains empty text due to application error of other technical problems. If it occurs, the record containing blank text will be removed to ensure data integrity.

In automatic data collection, it is common for text to contain symbolic characters that do not contribute to the meaning of the text. Some examples of symbols are emoticons, punctuation marks, or other symbols caused by writing errors or differences in text format. This symbol must be removed to avoid interference with the text mining process. However, the remaining text is preserved for the next step.

In general, the data standardization process aims to standardize data formats using only one form to avoid ambiguity in the data. The data standardization process is crucial because text data comes from various sources with different forms. For example, date data on one website uses a short date format (numbers); while in another website, it is written in text (letters).

2) LINGUISTIC PREPROCESSING

The linguistic preprocessing stage aims to prepare data according to the needs of the text mining algorithm. There are three processing stages: case folding, tokenization, and stop word removal. Each step will be explained as follows.

The first stage is case folding. This stage aims to change all uppercase letters in the text to lowercase. The text mining

algorithm relies on the probability of a term which depends on the frequency of occurrence of a term or word. For example, "Safe" will be converted into "safe" to indicate that these terms are equal.

The next step is stop words removal. Stop words, such as conjunction or object pronouns, are nonspecific words that appear frequently in sentences. Due to their nonspecific nature, stop words must be removed because they do not contain significant meaning in the text mining process. Some examples of words that are categorized as stop words are "and," "is," and "despite." Indonesian stop words using the Natural Language Toolkit (NLTK) library were utilized.

The final step is tokenization. The tokenization process is the process of changing sentences into tokens (i.e., words). This step aims to prepare the data for the topic modeling algorithm since the algorithm is based on the list of words for each document. Table II gives some examples of the transformation from the original headline text to the result of data preprocessing steps.

C. TOPIC MODELING

Topic modeling is a prominent technique in text analytics field and machine learning that helps in discovering the hidden thematic structures within a collection of documents. LDA is one of the approaches for topic modeling that is most frequently employed [10].

The foundation of LDA is the idea that each document in a corpus is composed of a variety of topics, and that each word is assigned to one of these topics. It explains that there is an underlying set of topics across the entire corpus and that documents are generated probabilistically based on these topics. The model used a Dirichlet distribution to model the distribution of topics within documents and the distribution of words within topics. Through a generative process, LDA assigns topics to documents and words to topics iteratively until it converges to a coherent set of topics and their associated word distributions.

The LDA model has found numerous applications in text analysis, including document clustering, and short reviews clustering [11], [12]. Its versatility, interpretability, and ability to discover hidden themes make it a fundamental tool for researchers and practitioners in fields such as information retrieval, social sciences, and natural language processing. The intuitive representation of topics as probability distributions over words has made LDA an indispensable tool for understanding and extracting meaningful information from large text corpora.

LDA assumes that there are K topics in the entire corpus, where K is a user-specified parameter. Each document is seen as a mixture of these K topics, and the proportions of these topics within a document are governed by a probability distribution, typically a Dirichlet distribution. Each topic is described by a distribution of words. This distribution reflects the likelihood of observing a particular word given the topic.

Based on the LDA graphical model in Figure 2, when generating a corpus, the Dirichlet prior parameters (α and β) are sampled once at the corpus level, and parameter θ is sampled once at the document level for each document. Let D be the number of documents in the corpus, W be the number of unique words in the vocabulary, and N be the number of words in a document. Then, θ is a $\boldsymbol{D}\times\boldsymbol{K}$ matrix representing the topic distribution for each document. θ_{dk} represents the probability of topic k in document d.

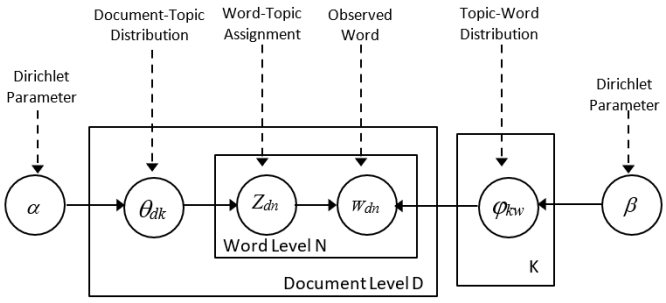

Figure 2. Graphical model of LDA.

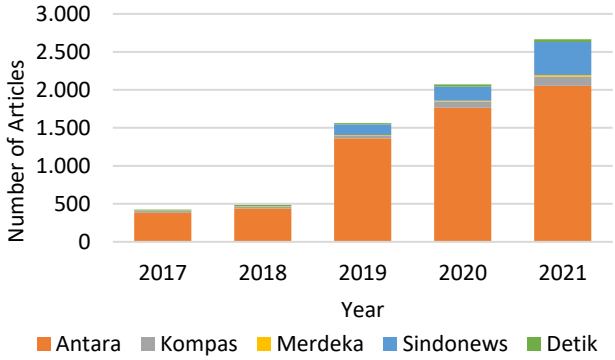

Figure 3. Distribution of the data based on frequency and year.

φ is a $\boldsymbol{K}\times\boldsymbol{W}$ matrix representing the word distribution for each topic. φ_{kw} represents the probability of word w in topic k. z is a $\boldsymbol{D}\times\boldsymbol{N}$ matrix representing the topic assignments for each word in each document. z_{dn} represents the topic assignment for the nth word in document d. While w is a $\boldsymbol{D}\times\boldsymbol{N}$ matrix representing the words in the documents. w_{dn} represents the nth word in document d.

III. RESULTS

A. EXPLORATORY DATA ANALYSIS

Exploratory data analysis was conducted on the data to get a better preliminary understanding of the dataset. Figure 3 shows the number of year distributions on the data. Year of 2021 had the largest dataset, with 2,666 total records; meanwhile, 2017 had the smallest dataset, with 425 records. Antara was the leader in data dissemination, followed by Sindonews, Kompas, Merdeka, and Detik, making up extremely small percentages of the total. The policies governing the quantity of data accessible on each of these media websites varied, accounting for the notable variations in data distribution.

Then, word clouds were generated from the data to get the most frequent words each year. Wordcloud is a helpful analysis tool for seeing what words frequently appear in the dataset. The bigger the display of the terms in the word cloud, the bigger the frequency of those words in the dataset.

In 2018, the most frequently occurring issues were represented by the words "pemerintah" (government), "dunia" (world), "lingkungan" (environment), and "negara" (country) (see Figure 4(a)). Figure 4(a) shows the government's frequent media coverage of environmental issues.

In 2019, shown in Figure 4(b), the issues in the media had not changed much, where the words that frequently appeared were "pemerintah" (government), "lingkungan" (environment), "klhk" (the Ministry of Environment and Forestry), and "bmkg"

Figure 4. Word clouds based on year (a) 2018, (b) 2019, (c) 2020, and (d) 2021.

(Meteorological, Climatological, and Geophysical Agency). The illustration shows the government's active response to environmental issues so that the media quote it. The keyword "bmkg" also indicates a prominent weather phenomenon that year.

In 2020, the COVID-19 pandemic took place. Based on Figure 4(c), during this year, terms such as "pemerintah" (government), "masyarakat" (society), "dunia" (world), "ekonomi" (economy), and "covid" were frequently mentioned in the media. The figure shows that media coverage was dominated by news about the COVID-19 pandemic, which peaked in 2020. This issue also intersected with communities directly affected by mass layoffs and other financial difficulties.

In 2021, as shown in Figure 4(d), media coverage was dominated by the words "pemerintah" (government), "dunia" (world), "dampak" (impact), and "ekonomi" (economy). This illustration shows that the media often quotes government statements in pushing for economic recovery as a direct result of the pandemic.

B. TOPIC MODELING IMPLEMENTATION

At this point, topic modeling was utilized to analyze the corpus. The Gensim [13] and pyLDAvis [14] were used to implement the topic modeling algorithm. Gensim is a Python package that includes the LDA algorithm and provides a more convenient multicore implementation for increasing speed. It also provides more varied text processing facilities than other packages. The pyLDAvis, on the other hand, is a package designed to help users interpret the topics extracted from a fitted LDA topic model and present them in an interactive web-based visualization.

How to interact with high-dimensional data in a meaningful way that is understandable for humans is a challenge for many machine learning algorithms, including the LDA model. Many researchers suggest that the ideal strategy for solving this problem is to select hyperparameters heuristically and then fine-tune them using empirical tests. Following this strategy, three hyperparameters in the LDA algorithm were selected for the initial setting, and the value was adjusted accordingly based on the results. The adjusted parameters are as follows.

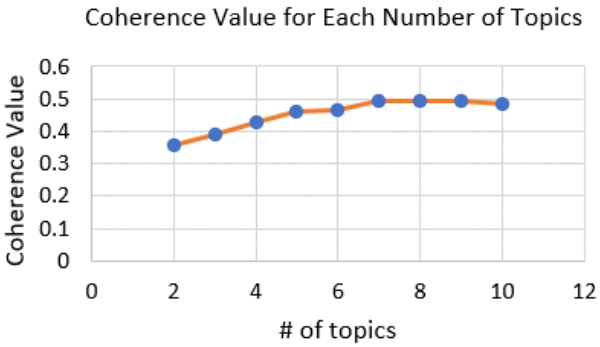

Figure 5. Coherence value plot for each number of topics.

TABLE III
THE COHERENCE VALUE FOR EACH SCENARIO

No.	# of Topics	Avg Coherence Value	Highest Coherence Value
1.	2	0.341	0.437
2.	3	0.398	0.450
3.	4	0.404	0.479
4.	5	0.438	0.524
5.	6	0.421	0.541
6.	7	0.447	0.559
7.	8	0.494	0.560
8.	9	0.486	0.556
9.	10	0.413	0.538

1) NUMBER OF TOPICS

The number of topics is frequently one of the most important hyperparameters in topic modeling. However, determining the ideal clusters or number of topics is often challenging. In this experiment, to determine the optimal number of topics, the algorithm was iterated in sequence over a range value (i.e., n = 2…10) of number of topics and evaluate the results based on the coherence value.

A statement can be considered coherent if each word in the statement supports the other. In topic modeling, coherence value can be obtained by calculating the similarity between frequent words in the corresponding topic. Based on the

TABLE IV
THE COHERENCE VALUE OF DIFFERENT α AND β

No.	Topics	Alpha (α)	Beta (β)	Coherence Value
1.	8	0.01	0.01	0.542
2.	8	0.01	0.3	0.488
3.	8	0.01	0.04	0.449
4.	8	0.01	0.91	0.413
5.	8	0.31	0.01	0.549
6.	8	0.31	0.31	0.490
7.	8	0.31	0.04	0.457
8.	8	0.31	0.91	0.476
9.	8	0.04	0.01	0.549
10.	8	0.04	0.04	0.449
11.	8	0.04	0.31	0.517
12.	8	0.04	0.91	0.452
13.	8	0.91	0.01	0.560
14.	8	0.91	0.31	0.487
15.	8	0.91	0.04	0.436
16	8	0.91	0.91	0.430

experiment, the optimal number of clusters was eight (see Figure 5), yielding an average coherence value of 0.494 and the highest coherence value of 0.560 (see Table III).

2) THE α AND β HYPERPARAMETERS

The α hyperparameter defines the document-topic density, whereas the β hyperparameter defines word-topic density. Documents with a higher α value contain more topics, while documents with a lower α value have fewer themes. Topics, on the other hand, are formed of a significant number of words in the corpus when β is high, and few numbers of words when β is low.

Using a setting range value (i.e., 0.01...1) of α and β hyperparameters, the algorithm was iterated with an increasing value of 0.01 for each parameter to obtain the optimal result for each hyperparameter. The coherence value results of different values of α and β are depicted in Table IV. Due to the large number of iterations and result values produced, not all of them can be displayed. Only a tiny portion, those with the highest value results, are displayed in Table IV due to practical considerations. As seen in the table, the optimal value of α hyperparameter is 0.91, and β is 0.01. This setting was used to build the final model, and the results will be discussed.

IV. DISCUSSION

Based on the value of number of topics equal to 8, the topic results of the model were listed in the following subsection. Then, topic interpretation based on the words that appeared in each cluster was discussed.

A. TOPIC INTERPRETATION

1) TOPIC 1 (RENEWABLE ENERGY)

Topic 1 was dominated by words such as "energi" (energy), "investasi" (investment), "esdm" (i.e., the ministry concerning mineral resources and energy), and "terbarukan" (renewable). This category can be interpreted with the topic of renewable energy. The distribution of words on the topic can be seen in Figure 6(a). The topic of energy is closely related to environmental issues, considering that most energy used today is fossil-based. The weakness of fossil-based energy lies in its limited availability, which is steadily depleting. Moreover, the production of fossil-based energy requires tens to hundreds of years.

The topic of renewable energy is also related to climate change. Increasing the earth's temperature, which is a global climate change phenomenon, has made researchers consider using solar-based energy (e.g., solar panels). This result is in concordance with [15] that state developing renewable energy, such as solar-based energy could have a positive impact on climate change.

2) TOPIC 2 (CARBON EMISSION)

In this topic, words that appeared were "emisi" (emissions), "karbon" (carbon), "industri" (industry), "kurang" (reduction), "konservasi" (conservation), "teknologi" (technology), "hijau" (green), and several other words. It can be interpreted that in this category, the model suggests the topic of reducing emissions or reducing carbon gases, especially those caused by industry.

It cannot be denied that the increase in carbon gas emissions which comes from industry or large-scale exploitation contributes to the cause of the climate change phenomenon. Therefore, there should be massive action to reduce the number of carbon emissions. The distribution of words in this group can be seen in Figure 6(b).

The topic of carbon emission could be captured by the LDA algorithm. The algorithm also grouped carbon emissions with words like "reduction" and "conservation" in the same category. This result suggests that there have been efforts to reduce emissions from various parties, one of which is through conservation and the increasingly widespread use of green technology. This result is aligned with the suggestion in [16] that rapid emission reduction should take place to restore the ecosystem.

3) TOPIC 3 (ENVIRONMENTAL MANAGEMENT)

In this category, words that appeared in the cluster were "lingkungan" (environment), "tanam" (planting), "pohon" (trees), "sampah" (garbage), "walhi" (i.e., Indonesian nongovernment organization focusing on environmental issues), "mangrove," and others. It can be interpreted that the theme in topic 3 is about environmental management. The distribution of words on topic 3 can be seen in Figure 6(c).

The word "environment" achieved the highest score in this category. The algorithm could classify environmental management activities involving people and communities into one topic category. Environmental management is closely related to climate change, where activities such as planting trees and mangroves and managing waste can reduce the adverse effects of climate change. This result corroborates previous studies, which state that many environmental management activities are carried out in developing countries by involving governments and companies in a variety of social activities targeting the decrease of climate change effect [17]

4) TOPIC 4 (DEVELOPMENT ECONOMICS)

In this category, the dominant words were "ekonomi" (economy), "kemenkeu" (the Ministry of Finance), "menko" (coordinating minister), "tumbuh" (growth), "bank," "imf" (International Monetary Fund), "proyek" (project), and others. Based on these words, it can be interpreted that the theme in topic 4 is the development economics. The distribution of words on topic 4 can be seen in Figure 6(d).

This category captures development activities carried out by the government amid climate change issues. The words

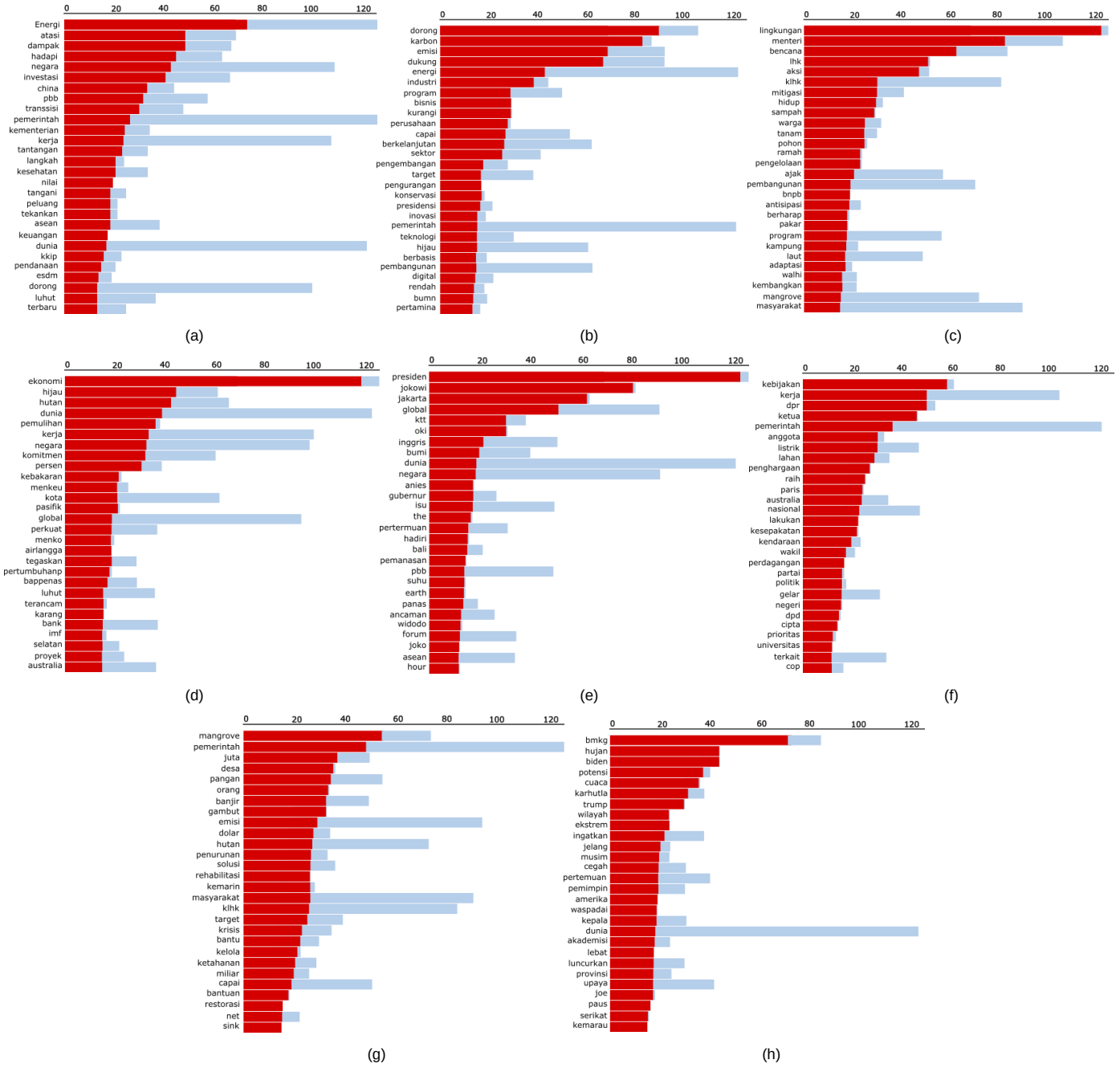

Figure 6. LDA visualization for each topic, (a) topic 1, (b) topic 2, (c) topic 3, (d) topic 4, (e) topic 5, (f) topic 6, (g) topic 7, and (h) topic 8. The Y axes contain top 30-most relevant terms for each topic. The X axes describe the number of each term.

"economy," "bank," and "ministry of finance" indicate the economic impact of the changing climate. Several words, such as "recovery" and "growth," indicate efforts to restore the community's economy, which is also in concordance with previous study regarding the relationship between climate change and its effect on the economy [18].

5) TOPIC 5 (INTERNATIONAL COOPERATION)

Some top frequent words that appeared in this cluster were "president," "jokowi" (i.e., the popular name of the current President of the Republic of Indonesia), "global," "konferensi" (summit), "isu" (issue), "inggris" (english), "pertemuan" (meeting), "forum," "asean" (Association of South East Asian Nation), "pemanasan" (warming), "temperatur" (temperature), and other words. Based on these words, international cooperation or conferences on global warming are suitable for explaining this category. The detailed distribution of words on this topic can be seen in Figure 6(e).

The model could capture the theme of international cooperation from the text, where this theme is very relevant to climate change. All countries are affected by climate change; therefore, cooperation should be undertaken to face the impact globally. This result has been pointed out in [19], which states that the objective of reducing emissions will not be plausible without the presence of international-level cooperation.

6) TOPIC 6 (POLICY/REGULATION)

Several top words that appeared in this category were "kebijakan" (policy), "kerja" (work), "dpr" (refers to Indonesia's House of Representatives), "pemerintah" (government), "perjanjian" (agreement), "nasional" (national), and others. The relevant theme for this group is policy or regulation. The distribution of words in this theme can be seen in Figure 6(f).

This category captured reporting on regulations at the local/national level carried out by the government and the DPR. This finding indicates that there have been efforts to address climate change issues by the government through policies and regulations to lessen the effect of climate change. This result is also in accordance with [20] about the relation between climate change and policymaking during the last 30 years, which indicate the importance of policy in addressing the climate change phenomenon.

7) TOPIC 7 (REHABILITATION)

In this category, some of the top frequent words that appeared were "mangrove," "solusi" (solution), "rehabilitasi" (rehabilitation), "restorasi" (restoration), and "makanan" (food). Based on these results, it can be interpreted that the theme representing this category is rehabilitation or restoration. The distribution of words in this theme can be seen in Figure 6(g).

This category describes efforts to prevent climate change's impact through rehabilitation and restoration activities. The algorithm classifies these restoration efforts as highly related to the preservation of mangrove forests. This result is considerably reasonable and is aligned with previous studies that suggest the conservation of mangrove forests in the Indonesian archipelago should be the highest priority to mitigate climate change [21].

8) TOPIC 8 (NATURAL DISASTER)

The words included in this topic were "bmkg," "hujan" (rain), "cuaca" (weather), "karhutla" (refers to forest and bushfires), "ekstrim" (extreme), "musim" (season), "waspada" (beware), and "kering" (dry). Based on this list, it can be interpreted that the topic representing this category is natural phenomena or disasters. The distribution of words on topic 8 can be seen in Figure 6(h).

The disaster category is an aspect that cannot be separated from the climate change phenomenon [22]. The topic modeling algorithm could capture this phenomenon into one category of natural phenomena and disasters. The most captured catastrophic phenomena were extreme rain and drought, forest fires, and the change of seasons. The word "bmkg" often appears in this category as it is government agency that always shares information with the public regarding disasters and their mitigation.

B. TOPIC ANALYSIS

There are some interesting findings based on the results of topic modeling in the previous subsection. In general, the model could describe important issues related to climate change such as renewable energy, disasters, international cooperation, policies, and development economics. It shows that the discussion on the issue in the media was quite comprehensive in discussing various aspects of climate change.

There are two main categories that can be observed from these results, namely climate change related to natural phenomena and climate change related to regulations or policies. In the category of climate change related to natural phenomena, the model succeeded in finding clusters of carbon emissions (topic 2); and catastrophic phenomena such as floods, droughts, and extreme weather (topic 8). In the climate category related to policy, the model succeeded in finding topic clusters such as renewable energy (topic 1), environmental management (topic 3), development economics (topic 4), international cooperation (topic 5), policy (topic 6), and rehabilitation (topic 7).

However, there are several topics that are missing from the clusters such as the themes of climate change with regard to agriculture [3], food security [23], health [24], and poverty [25]. The absence of the latter topics also becomes the limitation of this paper. There are at least two reasons which might cause this result. First, it is due to the limited timespan or dataset that was used in this research. Second, it might be due to the small frequency of headlines with the above themes indicating that those themes do not get serious attention.

V. CONCLUSION

In this study, modeling topics on climate change issues based on Indonesian media headlines have been conducted. Dataset from several online media was collected by scraping their websites. The data cleaning process was performed to prepare the data before implementing the clustering algorithm. Then, the LDA algorithm was applied to the dataset to model the topics.

Based on the results, the algorithm successfully categorized the main topics regarding climate change. Those topics were renewable energy, carbon emissions, environmental management, development economics, international cooperation, policy/regulation, rehabilitation, and disaster. However, the algorithm failed to categorize other important aspects such as agriculture, food security, health, and poverty.

For the future work, a bigger dataset and a longer timespan will be utilized to get a more comprehensive result. Furthermore, several clustering algorithms will be used to analyze the dataset to compare which algorithm yields better result than the other.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Conceptualization, Anang Kunaefi and Aris Fanani; methodology, Anang Kunaefi and Aris Fanani; validation, Anang Kunaefi and Aris Fanani; formal analysis, Aris Fanani; investigation, Anang Kunaefi; resources, Aris Fanani; data curation, Aris Fanani; writing—original draft preparation, Anang Kunaefi; writing—reviewing and editing, Anang Kunaefi and Aris Fanani; visualization, Anang Kunaefi; supervision, Aris Fanani; project administration, Aris Fanani; funding acquisition, Aris Fanani.

REFERENCES

[1] S. Fawzy, A.I. Osman, J. Doran, and D.W. Rooney, "Strategies for mitigation of climate change: A review," *Environ. Chem. Lett.*, vol. 18, no. 6, pp. 2069–2094, Nov. 2020, doi: 10.1007/s10311-020-01059-w.

[2] M. Measey, "Indonesia: A vulnerable country in the face of climate change," *Glob. Major. E-J.*, vol. 1, no. 1, pp. 31–45, Jun. 2010.

[3] N. Herlina and A. Prasetyorini, "Pengaruh perubahan iklim pada musim tanam dan produktivitas jagung (Zea mays L.) di Kabupaten Malang," *J. Ilmu Pertan. Indones.*, vol. 25, no. 1, pp. 118–128, Jan. 2020, doi: 10.18343/jipi.25.1.118.

[4] Y. Kinose *et al.*, "Impact assessment of climate change on the major rice cultivar Ciherang in Indonesia," *J. Agric. Meteorol.*, vol. 76, no. 1, pp. 19–28, Jan. 2020, doi: 10.2480/agrmet.D-19-00045.

[5] K.A. Harvian and R.J. Yuhan, "Kajian perubahan iklim terhadap ketahanan pangan," *Semin. Nas. Official Statist.*, 2020, pp. 1052–1061, doi: 10.34123/semnasoffstat.v2020i1.593.

[6] R. Cahyaningsih *et al.*, "Climate change impact on medicinal plants in Indonesia," *Glob. Ecol. Conserv.*, vol. 30, pp. 1–13, Oct. 2021, doi: 10.1016/j.gecco.2021.e01752.

[7] T.R. Keller *et al.*, "News media coverage of climate change in India 1997–2016: Using automated content analysis to assess themes and topics," *Environ. Commun.*, vol. 14, no. 2, pp. 219–235, Feb. 2020, doi: 10.1080/17524032.2019.1643383.

[8] P. Swarnakar and A. Modi, "NLP for climate policy: Creating a knowledge platform for holistic and effective climate action," 2021, *ArXiv: 2105.05621*.

[9] A. Minnich *et al.*, "ClearView: Data cleaning for online review mining," *2016 IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min. (ASONAM)*, 2016, pp. 555–558, doi: 10.1109/ASONAM.2016.7752290.

[10] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[11] A. Kunaefi and M. Aritsugi, "Characterizing user decision based on argumentative reviews," *2020 IEEE/ACM Int. Conf. Big Data Comput. Appl. Technol. (BDCAT)*, 2020, pp. 161–170, doi: 10.1109/BDCAT50828.2020.00002.

[12] H. Li, R. Lin, R. Hong, and Y. Ge, "Generative models for mining latent aspects and their ratings from short reviews," *2015 IEEE Int. Conf. Data Min.*, 2015, pp. 241–250, doi: 10.1109/ICDM.2015.28.

[13] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," *Proc. LREC 2010 Workshop New Chall. NLP Framew.*, 2010, pp. 45–50.

[14] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," *Proc. Workshop Interact. Lang. Learn., Vis. Interfaces*, 2014, pp. 63–70, doi: 10.3115/v1/W14-3110.

[15] A.G. Olabi and M.A. Abdelkareem, "Renewable energy and climate change," *Renew. Sustain. Energy Rev.*, vol. 158, pp. 1–7, Apr. 2022, doi: 10.1016/j.rser.2022.112111.

[16] J. Hansen *et al.*, "Assessing 'dangerous climate change': Required reduction of carbon emissions to protect young people, future generations and nature," *PLoS ONE*, vol. 8, no. 12, pp. 1–26, Dec. 2013, doi: 10.1371/journal.pone.0081648.

[17] J.A.P. de Oliveira and C.J.C. Jabbour, "Environmental management, climate change, CSR, and governance in clusters of small firms in developing countries: Toward an integrated analytical framework," *Bus. Soc.*, vol. 56, no. 1, pp. 130–151, Jan. 2017, doi: 10.1177/0007650315575470.

[18] D. Castells-Quintana, M.P. Lopez-Uribe, and T.K.J. McDermott, "Adaptation to climate change: A review through a development economics lens," *World Dev.*, vol. 104, pp. 183–196, Apr. 2018, doi: 10.1016/j.worlddev.2017.11.016.

[19] M.M. Ferrari and M.S. Pagliari, "No country is an island. International cooperation and climate change," Banque de France Working Paper, WP #815.

[20] J. Gupta, "A history of international climate change policy," *WIREs Clim. Change*, vol. 1, no. 5, pp. 636–653, Sep. 2010, doi: 10.1002/wcc.67.

[21] D. Murdiyarso *et al.*, "The potential of Indonesian mangrove forests for global climate change mitigation," *Nat. Clim. Change*, vol. 5, no. 12, pp. 1089–1092, Dec. 2015, doi: 10.1038/nclimate2734.

[22] J. Jetten *et al.*, "Responding to climate change disaster," *Eur. Psychol.*, vol. 26, no. 3, pp. 161–171, Jul. 2021, doi: 10.1027/1016-9040/a000432.

[23] T. Wheeler and J. von Braun, "Climate change impacts on global food security," *Sci.*, vol. 341, no. 6145, pp. 508–513, Aug. 2013, doi: 10.1126/science.1239402.

[24] World Health Organization (2021) "COP26 special report on climate change and health: The health argument for climate action," [Online], https://apps.who.int/iris/bitstream/handle/10665/346168/9789240036727-eng.pdf?sequence=1, access date: 16-Aug-2023.

[25] S. Hallegatte and J. Rozenberg, "Climate change through a poverty lens," *Nat. Clim. Change*, vol. 7, no. 4, pp. 250–256, Apr. 2017, doi: 10.1038/nclimate3253.