Indonesia Infrastructure Development Topic Discovery on Online News with Latent Dirichlet Allocation

Ahmad Fathan Hidayatullah¹, Muhammad Rifqi Ma’ariff, Muhammad Habibie³, Siti Khomsah⁴

¹Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia
²,³Center of Data Analytics Studies and Services, Universitas Jenderal Achmad Yani, Yogyakarta, Indonesia
⁴Department of Data Science, Institut Teknologi Telkom, Purwokerto, Indonesia

E-mail: ¹fathan@uii.ac.id, ²muhammad.rifqi@gmail.com, ³muhammadhabibi17@gmail.com, ⁴siti@ittelkom-pwt.ac.id

Abstract. In the period of 2014-2019, the Indonesian government has defined infrastructure development as a priority. Related to this, various media, in both online and offline, routinely report the infrastructure development news to the public. However, it is quite difficult for people to obtain a summary of information from the internet about the infrastructure development that have been carried out by the government. This study aims to provide a brief summary about infrastructure development in Indonesia by performing topic modeling approach using Latent Dirichlet Allocation. We found that the use of bigram language model could help identify phrase from the corpus. Therefore, the keywords contained in the topics are more interpretable and acceptable. Moreover, a coherence score measurement was applied to find the best number of topics from our dataset. Based on the experiments, we obtained 40 topic models from our dataset. From those 40 topics, we inferred several topic labels, such as, oil and gas infrastructure; power plant infrastructure; information technology infrastructure and internet networks; road infrastructure; reservoir infrastructure, irrigation networks and water resources; railway infrastructure; and airport infrastructure.

1. Introduction
Infrastructure development plays an important role in a country. Good infrastructure development is one of an indicator which shows the growth, prosperity and economic stability. As a developing country, Indonesia is still increasing its infrastructure development in some fields. In addition, there are many regions in Indonesia, which need a lot of infrastructure development, such as road, bridges, airports, and ports. In the period of 2014-2019, the Indonesian government has defined infrastructure development as a priority. The purpose of this program is to push the economic growth and to realize the justice and equity for all people in all regions in Indonesia. Therefore, the Indonesian government has proposed a program which is called national strategic projects (in Indonesian: Proyek Strategis Nasional, or PSN)¹. The national strategic projects are managed by the Committee for Acceleration of Priority Infrastructure Delivery (in Indonesian: Komisi Penyediaan Percepatan Infrastruktur Prioritas, ²

¹ https://kppip.go.id/proyek-strategis-nasional/
or KPPIP). KPPIP works across ministries and institutions as bridges, assisting project owners in preparing and implementing project development.

Related to this, various media, in both online and offline, routinely report the infrastructure development news to the public. However, it is quite difficult for people to obtain a summary of information from the internet about the infrastructure development that have been carried out by the government. This is caused by the number of news articles on the internet which provide lots of information about infrastructure development. Therefore, it is necessary to provide a brief summary about infrastructure development in Indonesia. The summary will be obtained by analyzing the topics from the online news articles. The analysis aims to obtain an overview and an illustration of the progress and problems. The analysis would be performed by extracting related topics through online news media, which provide news about infrastructure projects and development in Indonesia.

Topic modeling research in domains related to governance has attracted many researchers to discover hidden topic from texts. DiMaggio, et al. [1] analyzed U.S. government arts funding for newspaper by applying Latent Dirichlet Allocation (LDA). Cogburn [2] applied topic modeling to illustrate the trends and topics in the internet governance and cybersecurity debates. Deng, et al. [3] proposed a multi-level topic model with LDA to analyze the concern of citizens on social media during man-made disasters. Topic modeling on social media platform was also investigated by Driss, et al. [4]. They applied both LDA and Latent Semantic Analysis (LSA) to bring messages and information from citizens to policy-makers. Sha, et al. [5] presented a Hawkes binomial topic model to analyze COVID-19 conversations on Twitter among U.S. governors and presidential cabinet members.

This study aims to provide a brief summary regarding the infrastructure development in Indonesia by extracting topic from online news. In this work, the topic extraction will be carried out using topic modeling approach method which is called LDA. LDA has a whole coverage of important processes in the development of topic models, so that it can be used as a reference framework for other researchers in developing specific domains on other topics [6]. In addition, LDA has shown remarkable performance in various natural language processing tasks to find hidden topics in text documents [7].

2. Latent Dirichlet Allocation (LDA) Topic Modeling

Topic modeling provides a good way to analyze large amounts of text data that have not been classified [8]. Topic modeling is a text mining approach that is reliable in finding hidden text data and finding relationships between words from a corpus [9]. Topic modeling assumes that a document is a generative result of a set of hidden topics, where a topic is a probability distribution over words [8].

LDA was first introduced by [10] as a generative probabilistic model to automatically discover the hidden topics from a particular corpus data $D$. The probability calculation of LDA is represented as:

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

(1)

In equation (1), both $\alpha$ and $\beta$ represent the hyperparameter to obtain the likelihood of modeling. The $\alpha$ represents the Dirichlet prior parameter for the topic distribution in the document level and $\beta$ describes the word probability distribution for the particular topic. The topic distribution of document $d$ is represented in a form of vector $\theta$. The $z$ notation refers to the hidden topics from document $d$. The notation $M$ and $N$ represent the length of documents and the number of terms in the document respectively. Figure 1 illustrates the LDA model representation [10].

---

2 https://kppip.go.id/tentang-kppip/
3. Methodology

3.1. Data Collection and Preprocessing
We retrieved the dataset from some articles of online news media in Indonesia, which published during the period of 2014-2019. Before downloading the articles, we first determined some keywords which related to infrastructure development and national strategic projects, such as development, infrastructure, road, bridge, oil and gas. Furthermore, those keywords will be used to find the relevant articles through the internet. To obtain the content of the articles, we copied the URL of the articles, as the input parameter to the scraper. Table 1 shows the sample of the article URLs from several online news media. We utilized a Python scraper library called newspaper 3 to extract articles from online news media. The newspaper library works for more than ten languages including Bahasa Indonesia. Therefore, there is no special needed when we find the articles in Bahasa Indonesia. We collected 301 articles in Bahasa Indonesia with the total of 5,243 sentences, 92,880 words and 9,566 unique words.

| No | URL                                                                 |
|----|---------------------------------------------------------------------|
| 1  | http://www.beritasatu.com/satu/506680-oktober-tol-pemalangbatang-jalani-ului-kelayakan.html |
| 2  | http://jambi.tribunnews.com/2018/02/20/12-proyek-infrastruktur-yang-ambruk-di-indonesia |
| 3  | http://kupang.tribunnews.com/tag/pembangunan-infrastruktur-jembatan |
| 4  | https://news.okezone.com/read/2017/06/13/340/1714620/perbaikan-jalan-trans-kalimantan-dikebut-jelang-mudik-lebaran |
| 5  | https://tirto.id/luhut-sebut-target-infrastruktur-minimal-tercapai-90-persen-di-2019-cyBg |

The data preprocessing tasks in this work were slightly different from preprocessing data text in social media. In social media text, the text data are not clean and containing a lot of noisiness such as unstructured and non-standard words. On the contrary, the data of this study were more structured and did not contain non-standard words. In this preprocessing step, we performed some preprocessing tasks such as Unicode normalization, removing URLs, removing punctuations, removing numbers, case folding, removing additional white spaces and removing stop words.

3.2. Topic Modeling and Evaluation
The topic modeling process was conducted by utilizing a free Python library for topic modeling using Latent Dirichlet Allocation called Gensim 4. The topic modeling process was started by reading the preprocessed dataset. From this preprocessed data, we created a list of words and split the document based on tokens. We also build a bigram model to help identify phrases from our dataset. The next stage is to build a dictionary, then followed by creating an LDA model with TF-IDF (Term Frequency-Indexed Document Frequency). Finally, we compute coherence values from the each LDA model to evaluate the model.

3 https://newspaper.readthedocs.io/en/latest/
4 https://radimrehurek.com/gensim/index.html
3.3. Topic Analysis and Visualization

We analyzed the topic model based on the best number of the topics. From this analysis, we expected to obtain comprehensive and brief illustration of the infrastructure development which obtained from online news. Topic analysis was conducted by analyzing the relationships between words in each cluster group in the resulting LDA model. In addition, it will be seen also the relevance of topics between clusters to one another. Finally, we employed PyLDAvis\(^5\) as a visualization tools to analyze the extracted topics. PyLDAvis is built based on LDAvis \([11]\), which is an interactive visualization tool for Latent Dirichlet Allocation. The LDAvis result is presented in html format and it provides a flexibility in term of exploring topic-term relationships utilizing relevance \([11]\).

4. Result and Discussion

To build the topic model, we defined the number of topics as a parameter in LDA. We defined the number of topics from 5 until 50. Moreover, we also evaluated the topic coherence values in every 5 topics to obtain the best topic coherence value. Figure 2 shows the graph of the coherence score obtained from 5 to 50 topics.

\(^5\) https://pyldavis.readthedocs.io/en/latest/

![Figure 2. Coherence Score Graph](image)

It can be seen from the graph in Figure 2 that the coherence value increased significantly starting from 5 topics to 15 topics. After reaching 15 topics, the coherence value has slightly decreased for 20 topics. However, the coherence value obtained continues to fluctuate up to 50 topics and reach the highest score on the topic 40. This result is then used as a reference for further analysis, so that in-depth discussion will focus on 40 topics. The exact coherence values are shown in Table 2.

| Number of Topics | Coherence Value |
|------------------|-----------------|
| 5                | 0.5036          |
| 10               | 0.5083          |
| 15               | 0.5689          |
| 20               | 0.5594          |
| 25               | 0.5582          |
| 30               | 0.5662          |
| 35               | 0.5589          |
| 40               | 0.5709          |
| 45               | 0.5690          |
| 50               | 0.5705          |
Table 3. The most interpretable topics

| Topic# | Keywords | Label |
|--------|----------|-------|
| Topic#1 | gas, migas (oil and gas), gas_bumi (natural gas), pipa (pipe), pgn (Perusahaan Gas Negara/national gas company), bph_migas (Badan Pengatur Hilir Minyak dan Gas/the Agency’s Governing Body downstream Oil and Gas), holding_bumn, badan_usaha (business entity), bumi (earth). | Oil and gas infrastructure |
| Topic#2 | pembangkit_listrik (power plants), kantor_kementrian (ministry office), pembangkit (the ministry of power’s office), tenaga (power), listrik (electricity), produksi (production), berjalan_lancar (running smoothly), kapasitas (capacity), provinsi_jambi (Jambi Province). | Power plant infrastructure |
| Topic#3 | pembangkit_listrik (power plants), pembangkit (the ministry of power’s office), tenaga (power), listrik (electricity), produksi (production), berjalan_lancar (running smoothly), kapasitas (capacity), provinsi_jambi (Jambi Province). | Information technology and internet networks infrastructure |
| Topic#4 | pembangkit_listrik (power plants), pembangkit (the ministry of power’s office), tenaga (power), listrik (electricity), produksi (production), berjalan_lancar (running smoothly), kapasitas (capacity), provinsi_jambi (Jambi Province). | Information technology and internet networks infrastructure |
| Topic#5 | palapa ring (internet infrastructure project), bendungan (dam), palapa (internet infrastructure project), ring (internet infrastructure project), jepara (Jepara district), pelanggan (customer), penyelesaian (settlement), rudiantara (the name of Indonesian minister of communication and information), efisien (efficient), kartini (place name in Jepara). | Information technology and internet networks infrastructure |
| Topic#6 | palapa ring (internet infrastructure project), bendungan (dam), palapa (internet infrastructure project), ring (internet infrastructure project), jepara (Jepara district), pelanggan (customer), penyelesaian (settlement), rudiantara (the name of Indonesian minister of communication and information), efisien (efficient), kartini (place name in Jepara). | Information technology and internet networks infrastructure |
| Topic#7 | palapa ring (internet infrastructure project), bendungan (dam), palapa (internet infrastructure project), ring (internet infrastructure project), jepara (Jepara district), pelanggan (customer), penyelesaian (settlement), rudiantara (the name of Indonesian minister of communication and information), efisien (efficient), kartini (place name in Jepara). | Information technology and internet networks infrastructure |
| Topic#8 | palapa ring (internet infrastructure project), bendungan (dam), palapa (internet infrastructure project), ring (internet infrastructure project), jepara (Jepara district), pelanggan (customer), penyelesaian (settlement), rudiantara (the name of Indonesian minister of communication and information), efisien (efficient), kartini (place name in Jepara). | Information technology and internet networks infrastructure |
| Topic#9 | pertamina (State Oil and Natural Gas Mining Company), pembentukan_holding (holding formation), trivulan_pertama (first quarter), jangka_panjang (long-term), xl_axiata (telecommunications operator company), layanan_data (data service), miliar (billion), pgn_pertagas (national gas company), pmi (purchasing manager index), pipa_gas (gas pipe). | Oil and gas infrastructure & Information technology and internet networks infrastructure |
| Topic#10 | pertamina (State Oil and Natural Gas Mining Company), pembentukan_holding (holding formation), trivulan_pertama (first quarter), jangka_panjang (long-term), xl_axiata (telecommunications operator company), layanan_data (data service), miliar (billion), pgn_pertagas (national gas company), pmi (purchasing manager index), pipa_gas (gas pipe). | Oil and gas infrastructure & Information technology and internet networks infrastructure |
| Topic#11 | pesawat (aircraft), pita_lebar (broadband), program_pita (ribbon program), dpr ri (Indonesian people’s representative council), serat_optik (optical fiber), program_padat (solid program), pipa (pipe), produk (product), optik (optic), broadband (broadband). | Transportation & Information technology and internet networks infrastructure |
| Topic#12 | jaringan_irigasi (irrigation network), direktur_utama (president director), irigasi (irrigation), kementerian_pupr (minister of public work and human settlements), pembangunan (development), maret (March), operator (operator), daya_air (water power), jaringan (networks), tim (team). | Reservoir infrastructure, irrigation networks and water resources |
| Topic#13 | kereta_api (train), percepatan_pembangunan (accelerated development), pemalang_batang (Pemalang & Batang district), api (fire), kereta (train), kota_semarang (Semarang city), solo (Solo city), jalan_tol (toll road), dpr ri (Indonesian people’s representative council), anggota_komisi (commission members). | Railway infrastructure & Road infrastructure |

Among those 40 topic groups, we tried to discover some topics that are most related to the national strategic projects undertaken by the Indonesian government. The topics were inferred by observing the
correlation between the most 10 frequent words that appear in each topic. In this work, human judgement [12] was performed to help examine the keywords from the topic results and then give the topic label manually[13].

Table 3 and Table 4 show the most interpretable topics of this study. From the result, we found that the application of the bigram model has provided better results than without the bigram model. The use of the bigram model was able to capture some important phrase from our dataset. To determine phrase from the dataset, we set if two terms appear 10 times or more, then we will concatenate the two words and determine them as a phrase. As the result, we can see from the keywords for the resulting topic that the appearance of phrases could provide acceptable and more sensible topic label. For example, we can see from topic#1 that the appearance of the phrase gas_bumi (natural gas), bph_migas (Badan Pengatur Hilir Minyak dan Gas/the Agency’s Governing Body Downstream Oil and Gas), holding_bumn and badan_usaha (business entity) is more acceptable than if both words of each phrase are separated.

| Topic#25 | holding (holding company), holding_migas (oil and gas holding company), train (train), optimal (optimal), lrt_jakarta (light rail transit Jakarta), service (service),service_obligation (service obligation), asian_games (Asian Games), mantekh (place name in West Sumatera), presiden_jokowi (president of Indonesia) | Oil and gas infrastructure & Railway infrastructure |
| Topic#35 | knj (The Karimunjawa islands), penerbangan_perdana (first flight), menuju_karimunjawa (towards Karimunjawa), penerbangan (flight), arus_mudik (homecoming flow), media_sosial (social media), karimunjawa (Karimunjawa), semarang (Semarang), pertanian (agriculture), agus (person’s name) | Airport infrastructure |
| Topic#39 | pembangunan_waduk (reservoir construction), waduk (reservoir), mudik_lebaran (homecoming), selesai (finished), capaian (achievements), kepala_dinas (head of department), gas_irawan (person’s name/member of Indonesian people’s representative council), meningkatkan_kapasitas (increase capacity), tahun (year), keuntungan (profit). | Reservoir infrastructure, irrigation networks and water resources |

Based on Table 3 and Table 4, we concluded that there are two types of topic cluster. The first type of topic cluster is the topic which has one label only. This type of topic cluster only has one label because the majority of the most frequent words illustrate on a particular topic. On the other hand, the second one is the type of topic cluster, which has two topic labels. This type of topic clusters contains words that come from different topics. Among those topic clusters, based on the most 10 composed words, the topic number 9, 13, and 25 can be interpreted into two topics.

From the 40 topics obtained, we found 12 the most representative topic clusters that illustrate a brief summary about the national strategic project proclaimed by the Indonesian government in the period 2014-2019. Among those 12 topics, we inferred several topic labels, such as:

- Oil and gas infrastructure (Topic 1, 9 and 25).
- Power plant infrastructure (Topic 3).
- Information technology infrastructure and internet networks (Topic 5, 6, 9, and 11).
- Road infrastructure (Topic 13 and 8).
- Reservoir infrastructure, irrigation networks and water resources (Topic 12 and 39).
- Railway infrastructure (Topic 13 and 25).
- Airport infrastructure (Topic 35).

The topic model results produced were then visualized using the PyLDAvis library. PyLDAvis has two panels in its visualization. As shown in Figure 3, the left panel illustrates the inter-topic distance
map via multidimensional scaling. We used the Jensen-Shannon divergence as the default computation metrics for the inter-topic differences [11]. The left panel depicts the overall view of the topic model result. In addition, we are also able to see how common each topic and how topic link with each other. The topic clusters generated in the left panel are indicated by circles with a specific number in each group/topic cluster.

The right panel is described by using bar charts. The bars show the most occurring terms of the topics. Here, we can see the 30 terms that are most relevant for a particular topic. From the right panel of our topic modeling result, we know from the bar chart that the terminology ‘pembangunan infrastruktur’ (infrastructure development) and ‘palapa ring’ are the most frequent appear in the corpora. This result indicates that the most discussed topics in our data are about infrastructure development and ‘palapa ring’. The term ‘infrastructure development’ is clearly explained that the Indonesian government places a high priority on infrastructure development during the period of 2014-2019. The second term, ‘palapa ring’, is a telecommunications infrastructure project in the form of fiber optic development throughout Indonesia. The appearance of the term ‘palapa ring’ as the second occurring terms indicates about the development of infrastructure in the telecommunication domain.

![Figure 3. Topic Modeling Visualization using PyLDAvis](image)

5. Conclusion
This research has succeeded in conducting topic modeling of the online news article data which contains information about infrastructure development in Indonesia. The application of topic modeling was performed utilizing the Latent Dirichlet Allocation method to discover topics related to infrastructure development. To build the topic model, this research has combined the LDA method with TF-IDF (Term Frequency-Indexed Document Frequency) and bigram language model. We found that the use of bigram language model could help identify phrase from the corpus. Therefore, the keywords contained in the topics are more interpretable and acceptable. From those 40 topics obtained, we inferred several topic labels, such as, oil and gas infrastructure; power plant infrastructure; information technology infrastructure and internet networks; road infrastructure;
reservoir infrastructure, irrigation networks and water resources; railway infrastructure; and airport infrastructure.

References

[1] DiMaggio P, Nag M and Blei D 2013 Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding Poetics 41 pp 570–606

[2] Cogburn D L 2019 Analyzing trends and topics in internet governance and cybersecurity debates found in twelve years of IGF transcripts Proc. of the 52nd Hawaii Int. Conf. on System Sciences pp 902–11

[3] Deng Q, Gao Y, Wang C and Zhang H 2020 Detecting information requirements for crisis communication from social media data: An interactive topic modeling approach Int. J. of Disaster Risk Reduction 50 101692

[4] Belkahla Driss O, Mellouli S and Trabelsi Z 2019 From citizens to government policy-makers: Social media data analysis Government Information Quarterly 36 pp 560–70

[5] Sha H and Hasan M A 2020 Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives preprint arXiv:2004.11692

[6] Verma A and Gahier A K 2015 Topic modeling of E-news in Punjabi Indian J. of Sci. and Technol. 8 pp 1–10

[7] Wang B, Liakata M, Zubiaga A and Procter R 2017 A hierarchical topic modelling approach for tweet clustering Int. Conf. on Social Informatics (New York: Springer) pp 378–90

[8] Alghamdi R and Alfalqi K 2015 A Survey of topic modeling in text mining Int. J. of Advanced Computer Sci. and Application 6 pp 147–53

[9] Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y and Zhao L 2019 Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey Multimedia Tools and Applications

[10] Blei D M, Ng A Y and Jordan M I 2003 Latent dirichlet allocation J. of Machine Learning Research 3 pp 993–1022

[11] Sievert C and Shirley K 2014 LDAvis: A method for visualizing and interpreting topics Proc. of the Workshop on Interactive Language Learning, Visualization, and Interfaces (Baltimore, Maryland, USA: Association for Computational Linguistics) pp 63–70

[12] Chang J, Gerrish S, Wang C, Boyd-graber J L and Blei D M 2009 Reading tea leaves: how humans interpret topic models Advances in Neural Information Processing Systems pp 288–96

[13] Hu N, Zhang T, Gao B and Bose I 2019 What do hotel customers complain about? Text analysis using structural topic model Tourism Management 72 pp 417–26