SIMPLE
Signal and Image Processing Letters
Vol. 7., No. 1, 2025, pp. 86-92
ISSN 2714-6677
https://simple.ascee.org/

Topic Modelling of Disaster Based on Indonesia Tweet Using Latent Dirichlet Allocation

Aninditya Anggari Nuryono [a,1,*], Iswanto [b,2], Alfian Ma'arif [c,3], Rizal Kusuma Putra [d,4], Yabes Dwi Nugroho H [e,5], Muhammad Iman Nur Hakim [f,6]

[a,d] Informatics, Institut Teknologi Kalimantan, Balikpapan, Indonesia
[b] Department of Electrical Engineering, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia
[c] Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[e] Information System, Insitut Teknologi dan Bisnis Kalla, Makassar, Indonesia
[f] Automotive Engineering, Politeknik Keselamatan Transportasi Jalan, Tegal, Indonesia
[1] aninditya.nuryono@lecturer.itk.ac.id; [2] iswanto_te@umy.ac.id; [3] alfian.maarif@te.uad.ac.id; [4] rizal.putra@lecturer.itk.ac.id; [5] dwi@kallainstitute.ac.id; [6] m.iman@pktj.ac.id
* corresponding author

ARTICLE INFO

Keywords
Latent Dirichlet Allocation
Twitter
Disaster
Topic Modelling
Word Cloud

ABSTRACT

Twitter (now X) is a critical social media platform for disseminating information during crises. This study models disaster-related topics from Indonesian-language tweets using Latent Dirichlet Allocation (LDA). From a dataset of 8,718 tweets collected from official sources like BMKG and BNPB, we performed several preprocessing steps, including case folding, stop word removal, stemming, and normalization of slang and abbreviations. The optimal number of topics was determined using coherence scores, with the model achieving a peak coherence value of approximately 0.57. Keywords such as “banjir”, “kecelakaan”, “tanah longsor,” and others were used to collect data from Twitter accounts like "BMKG" (Meteorology, Climatology, and Geophysical Agency) and "BNPB" (National Disaster Management Agency). The results revealed that the most frequently discussed topics with high coherence values were “angin topan” “topan”, “virus corona”, “kecelakaan”, “tenggelam”, “badai”, “angin puting.” A word cloud was used to visualize these disaster-related topics.

1. Introduction

Social media is an online platform that enables users to share information. Among the most widely used social media platforms in Indonesia is Twitter, which is frequently used for sharing data and information [1], [2]. One significant type of information shared on Twitter pertains to natural and man-made disasters affecting Indonesia and the world. Government institutions, such as the Meteorological, Climatological, and Geophysical Agency (BMKG) and the National Disaster Management Agency (BNPB), actively use Twitter to broadcast urgent disaster-related information.

A disaster is defined as an occurrence, stemming from natural, non-natural, or human-related causes, that seriously threatens and disrupts the life of a community. Such events often lead to significant consequences, including loss of life, harm to the environment, destruction of property, and adverse psychological effects [3]. Examples of natural disasters are events like earthquakes, tsunamis, floods, wildfires, landslides, cyclones, droughts, and volcanic activity. In contrast, non-natural disasters are triggered by factors such as technological or modernization failures, widespread disease, epidemics, or accidents.

Topic modeling is one of the techniques in Natural Language Processing (NLP) used to identify frequently occurring themes in disaster-related discussions. In this paper, we propose using the Latent Dirichlet Allocation (LDA) method to identify patterns in disaster-related topics from Indonesian-language tweets. LDA is an unsupervised algorithm used to discover topics or themes within a corpus of documents [4]. Previous studies have widely implemented topic modeling using LDA [5], [6], Support Vector Machine (SVM) [7], LS-BTM [8], DBSCAN [9], LDA combined with KNN [10], and for various applications including sentiment analysis [11], [12], text summarization [13], transportation [14], healthcare [15], [16], and bioinformatics [17]. Research on topic modeling has also been conducted by Zhong et al., who introduced text mining using LDA on Wikipedia articles and Twitter data [18].

Anke Piepenbrink et al. demonstrated the use of LDA topic modeling as a novel approach for analyzing textual data [19]. Their analysis included 421 articles published in Organizational Research Methods (ORM), identifying 15 distinct topics. Jui-Feng Yeh et al. proposed a conceptual dynamic latent Dirichlet allocation (CDLDA) model for topic detection in conventional content [20]. Erwin B. Setiawan et al. applied LDA for detecting Indonesian-language spammers on Twitter, achieving an accuracy of 93.67% using logistic regression [21]. Besides topic modeling, LDA has also been employed for feature extraction [22] and for analyzing WhatsApp content [23].

While extensive research exists on topic modeling, a specific gap remains in the analysis of disaster-related discourse within the context of Indonesian-language tweets. This study aims to fill this gap by answering the research question: What are the dominant disaster-related topics discussed on Indonesian Twitter? To address this, we propose using Latent Dirichlet Allocation (LDA) to model and identify these topics from a corpus of Indonesian tweets. The resulting model provides insights that can support disaster response efforts by humanitarian and governmental organizations.

2. The Proposed Method

Topic modeling is a technique used to identify topics within a collection of documents. It consists of several entities: word, document, and corpus [13]. A word is the basic unit of discrete data in a document and represents the vocabulary, with each unique word indexed. A document is a sequence of words that form sentences and paragraphs. A corpus is a collection of documents, and corpora is the plural form of corpus. Topic modeling aims to discover hidden topics from unstructured documents. Initialization of the number of documents, topics, words, and iterations is performed in the topic modeling process. The coherence score is calculated to determine the optimal number of topics based on the highest coherence value.

2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is used to summarize, cluster, associate, and process large volumes of data, producing a list of weighted topics for each document. The Dirichlet distribution is used to determine the topic distribution per document. The output is used to allocate words to documents across different topics. Visualization of the LDA topic model is shown in Fig. 1.

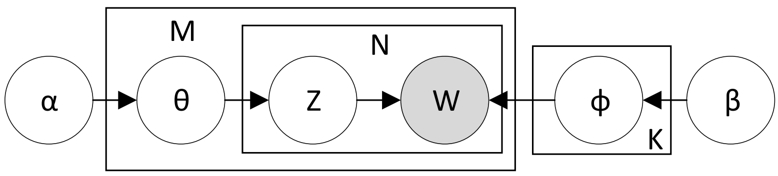

Fig. 1. LDA model

The model uses K for the number of topics, N for the word count in a document, and M for the total number of documents. The Dirichlet priors, α and β, are used for the per-document topic distribution (θ) and the per-topic word distribution (ϕ), respectively. Both ϕ and θ follow Dirichlet distributions, while z (topic assignment for a word) and w (word in a document) are multinomial.

The LDA workflow begins by initializing parameters for the number of documents, topics, iterations, and vocabulary size. The model then generates weighted outputs that can be normalized into probabilities. These probabilities represent the likelihood of a specific topic occurring in a document, and the likelihood of specific words within a topic.

The joint probability distribution is defined as:

$$P(z, w, \theta|\alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} [P(Zn|\theta) P(wn|Zn, \beta)] \quad (1)$$

The posterior distribution of latent variables is expressed as:

$$P(z, \theta|\alpha, \beta) = \frac{P(z, w, \theta|\alpha, \beta)}{P(w|\alpha, \beta)} \quad (2)$$

2.2. Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) normalizes the frequency of word occurrence across documents by assigning higher weights to terms that appear frequently in a specific document but less frequently in the entire corpus. Terms appearing frequently in many documents are penalized, as they are less informative.

$$idf(t) = 1 + log \frac{C}{1 + df(t)} \quad (3)$$

In this context, *idf(t)* is the inverse document frequency for term *t*, *C* is the total count of documents in the corpus, and *df(t)* is the number of documents that contain the term I

2.3. Data Preprocessing

A total of 8,718 tweets were collected in 2020 using the Tweepy library to access the Twitter Application Programming Interface (API). The data obtained from the Tweepy library can be retrieved after an account is registered and a token is obtained to access the API. After obtaining the data, the next stage is data preprocessing. Data preprocessing is the process of transforming unstructured data into structured data. Data preprocessing consists of tokenization, normalizing slang words, cleaning, removing Twitter symbols, stemming, case folding, and removing stop words [24], [25].

Tokenization is the process of tokenizing tweets into several tokens. Each token represents a word, idiom, or interjection (e.g., hahah, heheh, wkwkw), number, emotion, or punctuation. Normalization involves changing non-standard words and abbreviations into standard forms (e.g., "jgn" to "jangan", "mgkn" to "mungkin", "abis" to "habis", "gue" to "saya"). Cleaning involves removing URLs, digits or numbers, punctuations, and Unicode characters. Removing Twitter symbols involves deleting "RT," "@username," and "#". Case folding is the conversion of the entire text in a document into the same form. This letter case conversion involves changing the text to lowercase. Stop word removal is the process of eliminating frequently occurring words that do not significantly influence text classification, such as "di" (in/at), "yang" (which/who), "dan" (and), "dengan" (with), "adalah" (is), and "dalam" (in/within). The results of preprocessing the obtained tweet data can be seen in Table 1.

Tweet data was collected from several Twitter accounts, namely @kompascom, @BNPB_Indonesia, @infoBMKG, @datagoid, @TMCPoldaMetro, @PTJASAMARGA, @CNNIndonesia, @TvOne_News', @tribunnews, @tempodotco, @tvOneNews, @Kemendagri_RI, @Kemenag_RI, @KemenkesRI, @NTMCLantas_Polri, @DivHumas_Polri, @Metro_TV, and @Beritasatu. The keywords used to collect disaster tweets included "gempa bumi" (earthquake), "erupsi gunung berapi" (volcanic eruption), "tsunami, tanah longsor" (tsunami, landslide), "banjir" (flood), "kekeringan" (drought), "angin puting beliung" (whirlwind), "abrasi" (abrasion), "kebakaran" (fire), "kebakaran hutan dan lahan" (forest and land fires), "karhutla" (forest and land fires), "kecelakaan transportasi" (transportation accident), "kecelakaan" (accident), "badai" (storm), "Corona" (Corona), "Virus" (Virus), "kecelakaan mobil" (car accident), "kecelakaan motor" (motorcycle accident), "kebakaran" (fire), "tabrakan mobil" (car collision), "tabrakan motor" (motorcycle collision), "tenggelam" (drowning), and "angin topan" (typhoon/cyclone). The information provided by these accounts varies, including weather warnings, disaster updates, and even retweets of information from other accounts. These accounts were selected because they provide upto-date disaster information. The Twitter API was used to download and retrieve tweets from these accounts.

Table 1. Sample of Tweet Data Preprocessing

Before
Menemukan kebakaran, kecelakaan, bencana alam, pohon tumbang, tawuran, Ormas anarkis, ODMK (Orang Dengan Masalah Kejiwaan), KDRT, balap liar, begal, jambret, todong, geng motor dan rampas di wilayah DKI Jakarta hub telp: 112. https://t.co/b6sGd8mUAG
Otoritas China menyebut korban meninggal virus corona sudah melebihi 900 orang. https://t.co/sR72CCHMag
Penembakan brutal yang terjadi di Thailand, Sabtu (8/2/2020) menewaskan 27 orang termasuk pelaku yang ditembak mati oleh tim khusus kepolisian. https://t.co/7MRcpHxn0K
WNI yang dikarantina berharap masyarakat Indonesia, khususnya masyarakat Natuna tidak terlalu cemas karena sampai saat ini mereka dalam kondisi sehat. https://t.co/zGApnDUy4a
Berikut daftar wilayah yang berpotensi terjadi banjir selama puncak musim hujan pada Februari 2020 ini. https://t.co/kfdLBfSEKQ
After
menemukan kebakaran kecelakaan bencana alam pohon tumbang tawuran ormas anarkis odmk orang kejiwaan kd balap liar begal jambret todong geng motor rampas wilayah dki jakarta hub telp
otoritas china menyebut korban meninggal virus corona melebihi orang
penembakan brutal thailand sabtu menewaskan orang pelaku ditembak mati tim khusus kepolisian
wni dikarantina berharap masyarakat indonesia masyarakat natuna cemas kondisi sehat
daftar wilayah berpotensi banjir puncak musim hujan februari

3. Result and Discussion

The selection of the optimal number of topics in LDA is critical, as too few topics may cause topic overlap, while too many may lead to topic fragmentation. Based on the coherence score evaluation at 0.57 shown in Fig. 2, a coherent number of topics was identified around the higher threshold (close to 18), which indicates that the dataset is diverse and contains various disaster-related subthemes. This finding suggests that Twitter users discuss a wide range of disaster events, and a finer granularity of topics may improve the ability to capture specific patterns. The model with the highest coherence value was then processed into the LDA model along with the data dictionary. This model was subsequently represented using Word Cloud, a Python library. Word Cloud is a technique for representing text data based on the size of each word. Word Cloud illustrates the level of relatedness and importance based on the frequency of word occurrences. The results of data representation using LDA and Word Cloud for disasters are shown in Fig. 3 and Fig. 4.

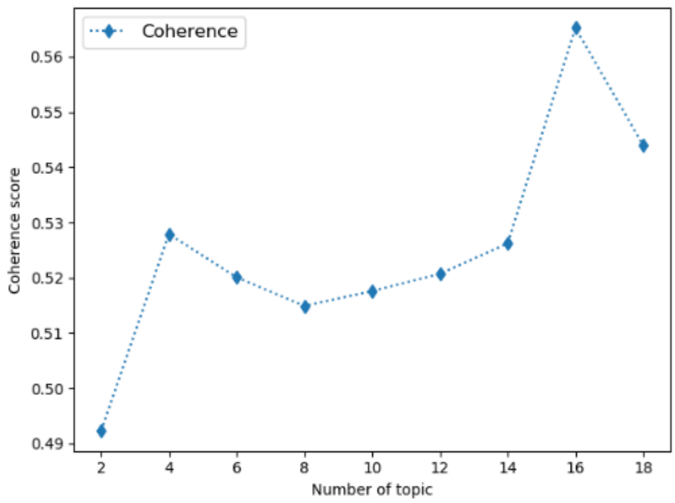

Fig. 2. Number of topics with coherence value

The LDA visualization in Fig. 3 reveals overlapping regions among several topics, specifically between topic 1 and 2, and between topic 4 and 6. These topic groupings indicate similarities between topics. This similarity can be observed from the distance between clusters, suggesting that the frequency and distribution of words within these topics are unique. The LDA visualization also displays the top 30 most frequently occurring words in the corpus. "Tenggelam" (drowning), "angin_topan" (typhoon/cyclone), and "kecelakaan" (accident) indicate that these words are frequently discussed and covered in news. Despite comprehensive preprocessing, some noisy terms such as informal expressions and colloquial particles (“ku,” “bgt,” “krn,” etc.) remain in the corpus. This highlights the challenge of processing informal and user-generated content in Bahasa Indonesia, where slang, abbreviations, and regional expressions are highly prevalent. Future work should consider enhancing the stop word list and incorporating more advanced natural language understanding tailored for Indonesian informal text.

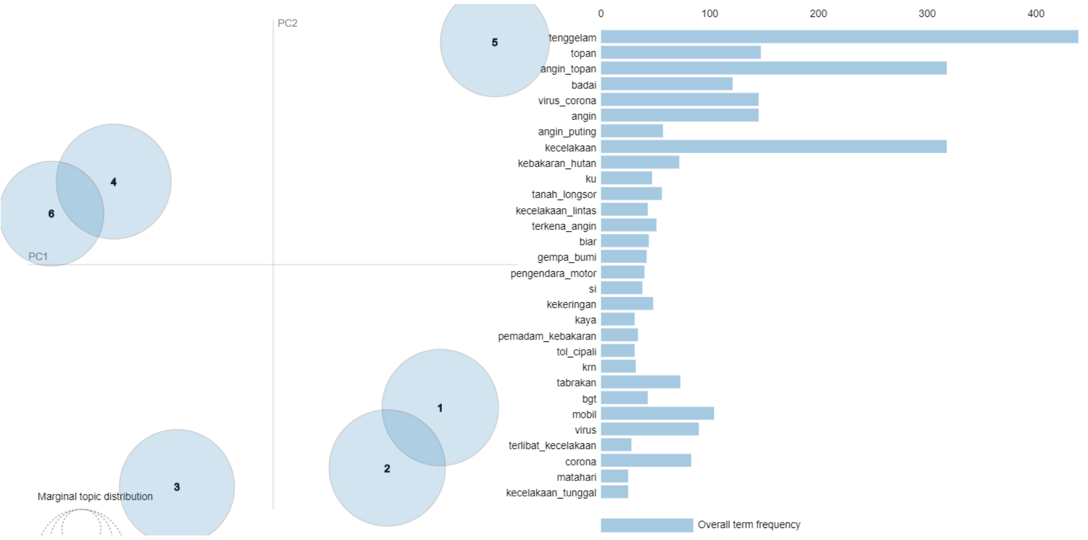

Fig. 3. LDA visualization disaster Indonesia tweet

In Fig. 4, disaster topics are illustrated using a word cloud. A word cloud is a visualization that conveys information by showing the most frequent and commonly appearing words from a document [26]. The size of the word is also crucial in indicating the frequency of important words within a text collection. In Topic #0, "angin_topan" (typhoon/cyclone), "topan" (typhoon), and "virus_corona" (coronavirus) have larger sizes. In Topic #1 to Topic #5, the words "kecelakaan" (accident), "tenggelam" (drowning), "badai" (storm), "angin_puting" (whirlwind), and "angin_topan" (typhoon/cyclone) dominate, leading to the conclusion that these words are frequently discussed and talked about by netizens regarding disasters on Twitter. Netizen is a term for social media users who communicate online.

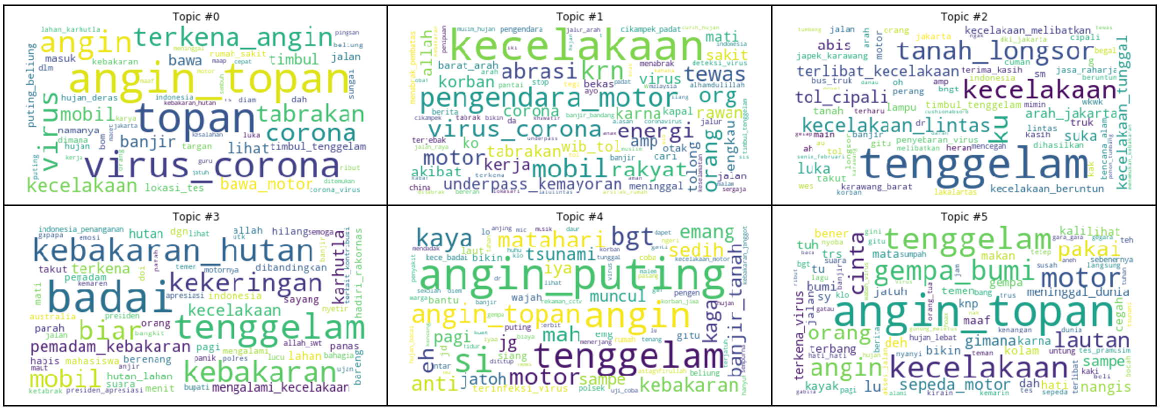

Fig. 4. LDA word cloud disaster Indonesia tweet

The results support the use of LDA not only for extracting disaster-related themes but also as a decision-support tool for governmental or humanitarian organizations. By monitoring dominant keywords and topic shifts over time, agencies like BNPB or BMKG can better understand public attention, misinformation trends, or regions of concern. Integrating real-time topic modeling with early warning systems may thus enhance disaster preparedness and response in Indonesia.

4. Conclusion

This study implements disaster topic modeling using Latent Dirichlet Allocation based on Indonesian tweets. Topic modeling is used to find topics that are related to each other. Frequently appearing topics have high coherence and frequency values. Overlapping topics indicate that they share similarities. The disaster-related topics most frequently discussed by netizens are "angin topan" (typhoon/cyclone), "topan" (typhoon), "virus corona" (coronavirus), "kecelakaan" (accident), "tenggelam" (drowning), "badai" (storm), "angin puting" (whirlwind), and "angin topan" (typhoon/cyclone). The preprocessing stage plays a vital role in removing unrelated words based on the existing corpus, including tokenization, stop word removal, stemming, case folding, normalizing slang words, and removing Twitter symbols. Word Cloud is used for visualizing interrelated topics.

References

[1] D. Gunawan, R. P. Siregar, R. F. Rahmat, and A. Amalia, “Building automatic customer complaints filtering application based on Twitter in Bahasa Indonesia,” *J. Phys. Conf. Ser.*, vol. 978, no. 1, 2018, doi: 10.1088/1742-6596/978/1/012119.

[2] N. Made, G. Dwi, M. A. Fauzi, and L. S. Dewi, “Cyberbullying identification in twitter using support vector machine and information gain based feature selection,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, no. 3, pp. 1494–1500, 2020, doi: 10.11591/ijeecs.v18.i3.pp1494-1500.

[3] S. Anson, H. Watson, K. Wadhwa, and K. Metz, “Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors,” *Int. J. Disaster Risk Reduct.*, vol. 21, Nov. 2016, doi: 10.1016/j.ijdrr.2016.11.014.

[4] S. Subramani, V. Sridhar, and K. Shetty, “A Novel Approach of Neural Topic Modelling for Document Clustering,” *Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018*, pp. 2169–2173, 2019, doi: 10.1109/SSCI.2018.8628912.

[5] A. F. Hidayatullah, E. C. Pembrani, W. Kurniawan, G. Akbar, and R. Pranata, “Twitter Topic Modeling on Football News,” *2018 3rd Int. Conf. Comput. Commun. Syst. ICCCS 2018*, pp. 94–98, 2018, doi: 10.1109/CCOMS.2018.8463231.

[6] K. S. Prabhudesai, B. O. Mainsah, L. M. Collins, and C. S. Throckmorton, “Augmented Latent Dirichlet Allocation (LDA) Topic Model with Gaussian Mixture Topics,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 2451–2455, 2018, doi: 10.1109/ICASSP.2018.8462003.

[7] K. B. Vamshi, A. K. Pandey, and K. A. P. Siva, “Topic Model Based Opinion Mining and Sentiment Analysis,” *2018 Int. Conf. Comput. Commun. Informatics, ICCCI 2018*, pp. 1–4, 2018, doi: 10.1109/ICCCI.2018.8441220.

[8] L. Li, Y. Sun, and C. Wang, “Semantic Augmented Topic Model over Short Text,” *Proc. 2018 5th IEEE Int. Conf. Cloud Comput. Intell. Syst. CCIS 2018*, pp. 652–656, 2019, doi: 10.1109/CCIS.2018.8691313.

[9] M. Mustakim *et al.*, “DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru,” *J. Phys. Conf. Ser.*, vol. 1363, p. 12001, 2019, doi: 10.1088/1742-6596/1363/1/012001.

[10] M. A. Safi’ie, E. Utami, and H. A. Fatta, “Latent Dirichlet Allocation (LDA) Model and kNN Algorithm to Classify Research Project Selection,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 333, no. 1, 2018, doi: 10.1088/1757-899X/333/1/012110.

[11] S. J. Blair, Y. Bi, and M. D. Mulvenna, “Sentiment Classification of Social Media Content with Features Generated Using Topic Models,” Front. Artif. Intell. Appl., vol. 284, pp. 155–166, 2016, doi: 10.3233/978-1-61499-682-8-155.

[12] I. R. Putri and R. Kusumaningrum, “Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia,” *J. Phys. Conf. Ser.*, vol. 801, p. 12073, Jan. 2017, doi: 10.1088/1742-6596/801/1/012073.

[13] S. H. Banu and S. Chitrakala, “Trending Topic Analysis using novel sub topic detection model,” *Proceeding IEEE - 2nd Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics, IEEE - AEEICB 2016*, pp. 157–161, 2016, doi: 10.1109/AEEICB.2016.7538263.

[14] L. J. Sun and Y. Yin, “Discovering themes and trends in transportation research using topic modeling,” *Transp. Res. Part C Emerg. Technol.*, vol. 77, pp. 49–66, 2017, doi: 10.1016/j.trc.2017.01.013.

[15] S. Wang, M. J. Paul, and M. Dredze, “Exploring health topics in Chinese social media: An analysis of

Sina Weibo," in AAAI Workshop on the World Wide Web and Public Health Intelligence, pp. 20–23, 2014.

[16] M. Nuser and E. Al-Horani, "Medical documents classification using topic modeling," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 17, no. 3, pp. 1524–1530, 2019, doi: 10.11591/ijeecs.v17.i3.pp1524-1530.

[17] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-3252-8.

[18] Z. Tong and H. Zhang, "A Text Mining Research Based on LDA Topic Modelling," in Computer Science & Information Technology, pp. 201–210, 2016, doi: 10.5121/csit.2016.60616.

[19] A. Piepenbrink and A. S. Gaur, "Topic models as a novel approach to identify themes in content analysis," https://doi.org/10.5465/AMBPP.2017.141, Oct. 2017, doi: 10.5465/AMBPP.2017.141.

[20] J. F. Yeh, Y. S. Tan, and C. H. Lee, "Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation," *Neurocomputing*, vol. 216, pp. 310–318, 2016, doi: 10.1016/j.neucom.2016.08.017.

[21] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Detecting Indonesian Spammer on Twitter," in 2018 6th International Conference on Information and Communication Technology (ICoICT), pp. 259–263, 2018, doi: 10.1109/ICoICT.2018.8528773.

[22] P. M. Prihatini, I. K. Suryawan, and I. N. Mandia, "Feature extraction for document text using Latent Dirichlet Allocation," *J. Phys. Conf. Ser.*, vol. 953, no. 1, 2018, doi: 10.1088/1742-6596/953/1/012047.

[23] L. Zhang, "Data and Content Analysis for Social Network Using LDA Text Model," *J. Phys. Conf. Ser.*, vol. 1213, no. 2, 2019, doi: 10.1088/1742-6596/1213/2/022035.

[24] M. A. Fauzi, "Random forest approach fo sentiment analysis in Indonesian language," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 46–50, 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.

[25] F. Hidayatullah and M. Maarif, "Pre-processing Tasks in Indonesian Twitter Messages," *J. Phys. Conf. Ser.*, vol. 801, p. 12072, 2017, doi: 10.1088/1742-6596/801/1/012072.

[26] S. Lohmann, F. Heimerl, F. Bopp, M. Burch, and T. Ertl, "Concentri Cloud: Word Cloud Visualization for Multiple Text Documents," in *19th International Conference on Information Visualisation*, pp. 114–120, 2015, doi: 10.1109/iV.2015.30.