Advances in Health Sciences Research, volume 34
Proceedings of the International Conference on Health and Medical Sciences (AHMS 2020)

Topic Modelling of Germas Related Content on Instagram Using Latent Dirichlet Allocation (LDA)

*1st Muhammad Habibi
Departement of Informatics
Universitas Jenderal Achmad Yani
Yogyakarta, Indonesia
muhammadhabibi17@gmail.com

2nd Adri Priadana
Departement of Informatics
Universitas Jenderal Achmad Yani
Yogyakarta, Indonesia
adripriadana3202@gmail.com

3rd Andika Bayu Saputra
Pusat Studi dan Layanan Analitik Data
Universitas Jenderal Achmad Yani
Yogyakarta, Indonesia
dika.putra21@gmail.com

4th Puji Winar Cahyo
Pusat Studi dan Layanan Analitik Data
Universitas Jenderal Achmad Yani
Yogyakarta, Indonesia
pwcahyo@gmail.com

***Abstract*— Content generated by Instagram users related to the Healthy Living Community Movement (GERMAS) has provided new media information that is important for the community and, in particular, the health department. At present, Indonesia is facing a serious challenge in the form of a double burden of disease. Changes in people's lifestyles are suspected to be one of the causes of a shift in disease patterns (epidemiological transition) in the last 30 years. Discussions on what topics occur in the community related to health, as well as community complaints, have not been identified. The Data Mining technique makes it possible to analyze and extract any topics that are contained from the data captions from Instagram. This study uses Latent Dirichlet Allocation (LDA) as a method for modeling topics. The results of evaluating the number of topics using topic coherence yielded the eight most appropriate topic segments. Based on the results of content analysis on each topic segment, it was found that the most dominant topic related to GERMAS was a healthy lifestyle diet.**

Keywords*— *Topic Modelling, LDA, Text Mining, Data Mining, Latent Dirichlet Allocation

I. INTRODUCTION

The Healthy Life Society Movement (GERMAS) is a systematic and planned action carried out jointly by all components of the nation with awareness, willingness, and ability to behave healthily to improve the quality of life. [1]. At present, Indonesia is facing a severe challenge in the form of a double burden of disease. Changes in people's lifestyles are suspected to be one of the causes of a shift in disease patterns (epidemiological transition) in the last 30 years. In the 1990s, the biggest causes of death and illness were infectious diseases such as Upper Respiratory Infection (ARI), Tuberculosis (TB), and Diarrhea. But since 2010, non-communicable diseases (PTM) such as Stroke, Heart, and Diabetes have a more significant proportion in health care [2]. This shift in disease patterns has resulted in a burden on state health financing. The resources needed to treat PTM in addition to requiring high costs also require a long time, said Minister of Health. Therefore, GERMAS is a momentum for the community to cultivate a healthy lifestyle.

Communication technology is developing at this time, making communication as a significant need in everyday life. One of the most widely used communication media by the public is social media. Social media allows people to communicate quickly and easily. Instagram is one of the most popular social media in the community these days. Based on data from napoleoncat.com, 2019 Instagram users in the world are increasing rapidly. Especially in Indonesia, Instagram users quickly increased to 59.8 million as of October 2019. Indonesia ranks 4th most users in the world after the United States 110 million, Brazil 79 million, and India 69 million [3].

Instagram is a social media platform that facilitates us to share information. The information uploaded on Instagram consists of three essential parts, namely: The uploaded image, the caption that is the core of the message, and the hashtag. Hashtags, written with the # symbol, are used to index keywords or topics on Instagram [4]. To succeed in the implementation of GERMAS, one of the things done is by using the hashtag #germas in Instagram social media uploads. The use of the hashtag makes it easier for people to get information related to the healthy living community movement.

Content generated by Instagram users related to GERMAS has provided new media information that is important for the community and, in particular, the health department. Discussions on what topics occur in the community related to health, as well as community complaints, have not been identified. The Data Mining technique makes it possible to analyze and extract any topics that are contained from the data captions from Instagram.

This study aims to model topics on Instagram caption related to GERMAS. Latent Dirichlet Allocation (LDA) is used as a method for modeling topics. LDA is an unsupervised technique that automatically creates topics based on patterns of (co) occurrence of words in the documents that are analyzed [5]. Research related to modeling topics using LDA includes modeling topics related to hotel service online reviews [6], modeling topics on scientific articles [7]. Modeling topics on road traffic using twitter data [8], modeling topics related to "ethnic marketing" In 239 journal articles published by nine major publishers [9].

Research related to the use of Instagram data has been widely used, including detection of selfies on Instagram [4], clustering user characteristics based on hashtags on Instagram [10], ranking keywords based on image captions on Instagram using TF-IDF [11]. As for other studies, namely research exploring the habits and involvement of Instagram by the Indonesian Government Ministry [12].

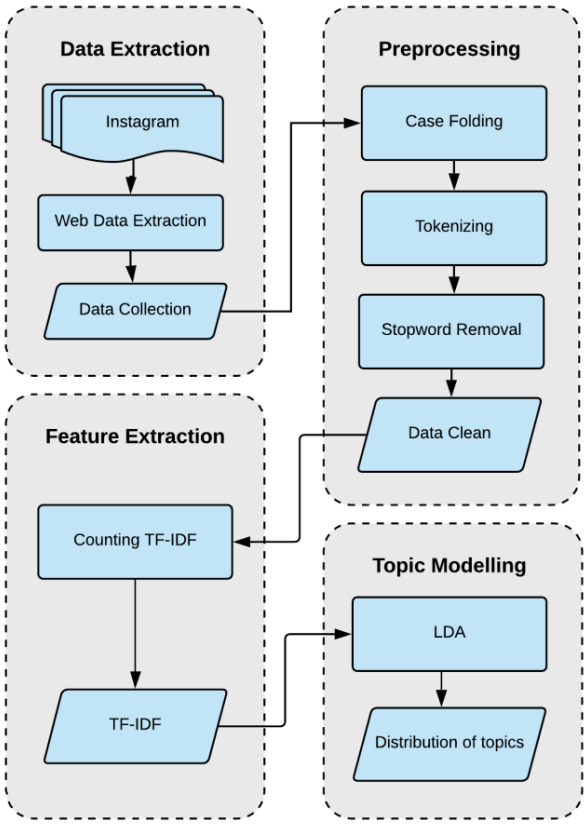

Fig. 1. System Architecture

A. Data Extraction

The first step in this research is data extraction. Instagram data caption is collected by using web data extraction. The contribution of web data extraction is to automate the process of information acquisition, especially information sourced from articles or freewriting on the internet [13]. The data used in this study uses data captions from Instagram with the #germas keyword. The data used are 80,745 thousand data captions, with data retrieval period 19 November 2018 to 21 November 2019. The following is an overview of the basic anatomy of the search using the #germas keyword on Instagram, which can be seen in Figure 2 and Figure 3.

B. Preprocessing

Preprocessing data is the process of preparing and cleaning text data before text analysis [14]. The steps in preprocessing used in this study include:

- • Case folding is the process of converting text into one upper or lower case letter.
- • Tokenizing is the process of dividing the text into tokens.
- Stop-word removal is the process of removing words from documents that do not have an essential role in providing patterns or information

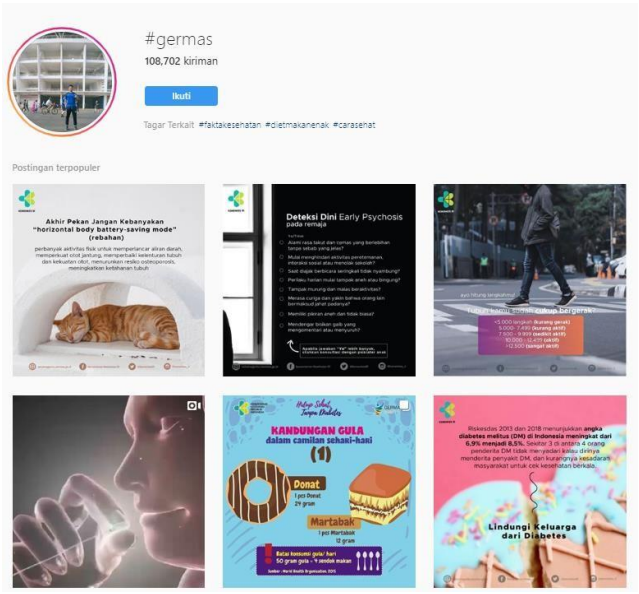

Fig. 2. Caption on Instagram data posts

C. Feature Extraction

Feature extraction is an extraction process to identify the entities in question [15]. A large corpus can usually contain more than 100,000 unique words, with most of these words appearing only in a few documents. Applying the LDA model to all terms in the corpus requires high computation. It is not very useful because most words have a distribution pattern that does not contribute to meaningful topics. For example, some words too often are informative such as "yang" and "dan" generally appear in every document, regardless of topic.

A useful technique for filtering out words that are too rare or too general is to use TF-IDF (Term Frequency-Inverse Document Frequency), which gives a low score for words that are very rare or very frequent. Another option is to have a minimum frequency cut-off for filtering out uncommon words and using a list of common words (and/or limiting Term Frequency - Inverse Document Frequency) to filter out words that are too general [5].

Term frequency (TF) is a standard notion of frequency in corpus-based natural language processing [16]. The application of term frequency can be made to extract text from students' comments towards lecturers [17]. TF-IDF (Term Frequency-Inverse Document Frequency) is a metric that is commonly used in the process of text categorization [18]. The use of TF-IDF works well with text mining methods [19].

D. Latent Dirichlet Allocation

Blei, et al [20] has introduced LDA (Latent Dirichlet Allocation) as a generative probabilistic model for groups of discrete data such as text corpora. LDA is an unsupervised machine learning technique. This method aims to model documents as emerging from multiple topics, where a topic is defined as a distribution over a fixed vocabulary terms

[21]. There are three generative process for each document in the collection [22]:

- Select a topic randomly from its distribution over topics for each document.
- Sample a word from the distribution over the words related to the chosen topic.
- Reiterate the process for all words in the document.

The visualization of LDA model representation can be seen in Figure 4 below.

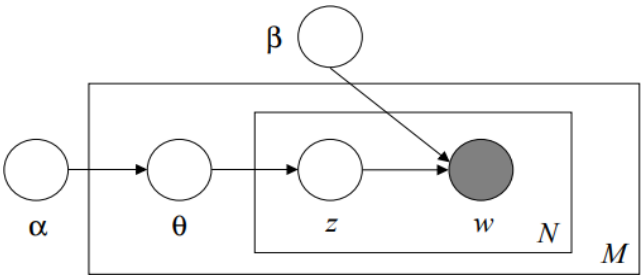

Fig. 3. LDA model representation [20]

Figure 4 shows the three levels of LDA representation. The first level is corpus-level parameters which are represented by symbol α and β. These corpus-level parameters are assumed to be sampled once in the process of generating corpus. Secondly, document-level variables (θ), sampled once of each document. Finally, word-level variables which symbolized by z and w. Word-level variables are sampled once for each word in each document.

II. Result and Discussion

This section will discuss the results of experiments that have been carried out. In our analysis, we made a visualization of a topic segment with a predetermined topic number of 15 topics. This visualization is to show the physical representation of the word frequency distribution for each topic. Figure 5 shows a viewing of the distance map between the topics produced.

After building a topic segment with a predetermined range, the next step is to evaluate the number of topic segments that are most appropriate. This study uses topic coherence to determine the most appropriate amount of topic segments.

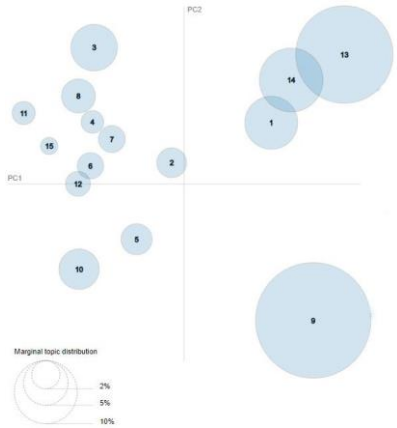

Fig. 4. Distance map between the topics produced

Topic Coherence prints a topic by measuring the level of semantic similarity between words with a high score on a topic. This measurement helps distinguish between topics that can be interpreted semantically and topics that are the result of human interpretation [23]. Topic Coherence is another way to evaluate topic models with a far higher guarantee of human analysis. Figure 6 shows the results of measuring the amount of topic suitability using topic coherence.

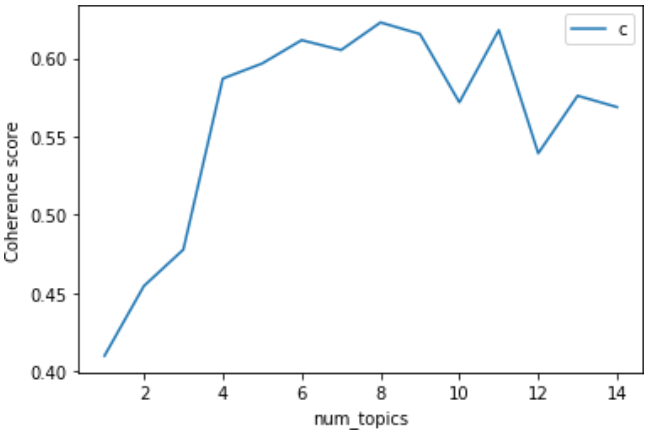

Fig. 5. Amount of topic suitability using topic coherence

Based on Figure 6, it can be seen that the most appropriate number of topic segments is eight topic segments. Therefore, in the subsequent analysis using eight topic segments. Figure 7 is a visualization of distance maps between topics produced for eight topic segments.

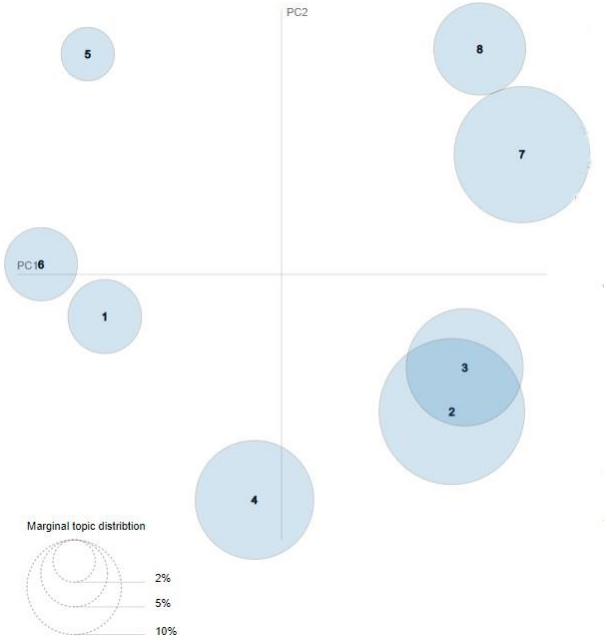

Fig. 6. Distance maps between topics produced for eight topic segments

For content analysis, we created two-word cloud visualization models, namely the word cloud for the entire segment and the word cloud for each segment in eight topic segments. Word cloud visualization will give an idea of what words are often used in specific conversations. This means we can see what words are the most significant that build a specific segment of the model being built.

Word cloud from all segments aims to summarize all keywords that appear in 8 topic segments. Based on the word cloud in Figure 8, it can be seen that the most dominant topic is a healthy lifestyle. As the word cloud shows, it is known that words such as life, health, weight, body, diet strongly dominate the most widely used keywords in all segments.

Fig. 7. Word cloud in all segments

Besides, we create a word cloud from each segment in eight topic segments to find the main topics for each topic. Figure 9 is a word cloud on eight topic segments.

Fig. 8. Word cloud on eight topic segments

Based on the word cloud results in Figure 9, we made a topic analysis for each topic segment. A topic analysis is made based on the relationship of words that often appear on each topic segment. The results of the topic analysis can be seen in Table 1.

TABLE I. TOPIC ANALYSIS PADA SETIAP SEGMEN TOPIC

Topic Segmentation	***keyword***	***Topic Analysis***
Topic 1	turun, via, #tipssehat, diet, coach, badan, berat, makan, uang, bervariasi	Healthy tips with diet coaching
Topic 2	#day, #tegalhits, #naikturunbb, #gobbideal, #slawihits, #brebeshits, #dietaman, #dietsehatalami, #herbalifebandarlampung, #dietjantung	Dietary lifestyle movements from regions in Indonesia
Topic 3	#lifestyle, #bahagia, #semangat, #bijaksana, #jalanpagi, #kaya, #gerakanindonesiasehat, #happy, #freedom, #morningwalk	Healthy lifestyle movement with morning walk activities
Topic 4	makan, sehat, kita, #dietsehat, kamu, bisa, tubuh, untuk, diet, badan	The importance of consuming healthy food for a diet
Topic 5	kesehatan, germa, masyarakat, sehat, senam, hidup, #hkn, kegiatan, puskesma, gerakan, #puskesmas	Healthy lifestyle movement health center with gymnastics activities
Topic 6	#dietalowcarb, #dieting, #diettips, #turunsize, #nutritionist, #gizipuskesmas, #dietgm, #dailyactivity, #makananbernutrisi, #makananrendahkalori	Tips on dieting by reducing carbohydrate and low-calorie foods

Topic Segmentation	***keyword***	***Topic Analysis***
Topic 7	propoli, #hwacademy, #griyasehathwa, #nutrisel, #natto, #gayahidupsehat, hwa, #anaksehat, #indonesiasehat, sehat	Health Wealth Academy expert and herbal medicine
Topic 8	senen, #gerakanmasyarakathidupsehat, #makanbuahsayur, #ppni, #nurse, ns, #garutupdate, kep, #ns, klinik	Healthy community movement by consuming fruits and vegetables

The most dominant topic of germas tweets is related to a healthy lifestyle diet. Topics that discuss diet can be seen in the topic segments 1, 2, 4, 6, and 8. While the topic segment 3 discusses the movement of healthy living on foot every morning, topic segment 5 discusses gymnastics activities at the puskesmas, and topic segment 7 discuss treatment herbs through Healthy Wealth Academy.

III. CONCLUSION

This research succeeded in modeling topics related to #germas on Instagram using Latent Dirichlet Allocation (LDA). This study, using the topic coherence to evaluate the number of most suitable topics, namely eight topic segments. Based on the results of content analysis on each topic segment, it was found that the most dominant topic was related to a healthy lifestyle diet. With the results of this analysis, it is hoped that it can help the community to start a healthy life by participating in the success of the germas program or the healthy life movement of the government.

ACKNOWLEDGMENT

The research team would like to thank the Department of Informatics and the Study Center and Data Analytic Services of Universitas Jenderal Achmad Yani Yogyakarta, who have supported the research team to add insight and knowledge through this research. Hopefully this research can bring benefits to the progress of the Indonesian nation.

REFERENCES

[1] B. Komunikasi Pelayanan Masyarakat and T. Komunikasi Pemerintah Kemkominfo, "GERMAS Wujudkan Indonesia Sehat," 2019. [Online]. Available: https://www.depkes.go.id/article/view/16111500002/germas-wujudkan-indonesia-sehat.html. [Accessed: 23-Nov-2019].

[2] B. Komunikasi Pelayanan Masyarakat, "Pemerintah Canangkan Gerakan Masyarakat Hidup Sehat (GERMAS)," 2019. [Online]. Available: https://www.kemkes.go.id/article/view/16111600003/pemerintah-canangkan-gerakan-masyarakat-hidup-sehat-germas-.html. [Accessed: 23-Nov-2019].

[3] J. Pokrop, "Instagram Statistics That Matter for Marketers in 2019," 2019. [Online]. Available: https://napoleoncat.com/blog/instagram-statistics/. [Accessed: 24-Nov-2019].

[4] A. Priadana and M. Habibi, "Face Detection using Haar Cascades to Filter Selfie Face Image on Instagram," in 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 6–9.

[5] C. Jacobi, W. Van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," Digit. Journal., vol. 4, no. 1, pp. 89–106, 2016.

[6] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation," Tour. Manag., vol. 59, pp. 467–483, 2017.

[7] H. Jelodar et al., "Stabilization of an Inverted Robot Arm Using Neuro-Controller," Multimed. Tools Appl., vol. 78, pp. 183–198, 2018.

[8] A. F. Hidayatullah and M. R. Ma'Arif, "Road traffic topic modeling on Twitter using latent dirichlet allocation," in 2017 International Conference on Sustainable Information Engineering and Technology (SIET), 2017, pp. 47–52.

[9] S. Moro, G. Pires, P. Rita, and P. Cortez, "A text mining and topic modelling perspective of ethnic marketing research," J. Bus. Res., vol. 103, pp. 275–285, 2019.

[10] M. Habibi and P. W. Cahyo, "Clustering User Characteristics Based on the influence of Hashtags on the Instagram Platform," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 13, no. 4, pp. 399–408, 2019.

[11] B. A. Kuncoro and B. H. Iswanto, "TF-IDF method in ranking keywords of Instagram users' image captions," in 2015 International Conference on Information Technology Systems and Innovation (ICITSI), 2015, pp. 1–5.

[12] A. F. Azmi and I. Budi, "Exploring practices and engagement of Instagram by Indonesia Government Ministries," in 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), 2018, pp. 18–21.

[13] A. Priyanto and M. R. Ma'arif, "Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik)," Indones. J. Inf. Syst., vol. 1, no. 1, pp. 25–33, 2018.

[14] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," Procedia Comput. Sci., vol. 17, pp. 26–32, 2013.

[15] H. Siqueira and F. Barros, "A Feature Extraction Process for Sentiment Analysis of Opinions on Services," Proc. III Int. Work. Web Text Intell., 2010.

[16] M. Yamamoto and K. W. Church, "Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus," Assoc. Comput. Linguist., vol. 00, no. 0, pp. 1–45, 2000.

[17] M. Habibi and Sumarsono, "Implementation of Cosine Similarity in an automatic classifier for comments," JISKA (Jurnal Inform. Sunan Kalijaga), vol. 3, no. 2, pp. 38–46, 2018.

[18] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, 2002.

[19] M. Habibi, "Analisis Sentimen dan Klasifikasi Komentar Mahasiswa pada Sistem Evaluasi Pembelajaran Menggunakan Kombinasi KNN Berbasis Cosine Similarity dan Supervised Model," Universitas Gadjah Mada, 2017.

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

[21] D. M. Blei and J. D. Lafferty, "Topic Models," in Text Mining: Classification, Clustering, and Applications, 2009, p. 71.

[22] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proceedings of the First Workshop on Social Media Analytics - SOMA '10, 2010, pp. 80–88.

[23] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP), 2012, no. July, pp. 952–961.