Advances in Health Sciences Research, volume 34

Proceedings of the International Conference on Health and Medical Sciences (AHMS 2020)

Analysis of Popular Hashtags on Instagram Account The Ministry of Health

*1st Puji Winar Cahyo
Department of Informatics
Universitas Jenderal Achmad Yani
Yogyakarta, Indonesia
pwcahyo@gmail.com

2nd Muhammad Habibi
Department of Informatics
Universitas Jenderal Achmad Yani
Yogyakarta, Indonesia

3rd Adri Priadana
Pusat Studi dan Layanan Analitik Data
Universitas Jenderal Achmad Yani
Yogyakarta, Indonesia
adripriadana3202@gmail.com

4th Andika Bayu Saputra
Pusat Studi dan Layanan Analitik Data
Universitas Jenderal Achmad Yani
Yogyakarta, Indonesia
dika.putra21@gmail.com

***Abstract*— Using the Instagram platform to distribute information on this decade is considered quite useful. The Ministry of Health Republic Indonesia has an Instagram account with the username kemenkes_ri used as a medium for distributing health information. That information delivers through Instagram posts always includes hashtags that are used as search keywords for the content that has been distributed. Therefore, analysis is performed on all posts made by kemenkes_ri for one year in 2019. From hashtag monitoring can be seen the total frequency of posts every month for one year and top 10 hashtags that are often used. Furthermore, From the results of the most popular hashtag can proceed to search for topics that are often discussed using a combination of bigram features, weighting Term Frequency - Inverse Document Frequency (TF-IDF), followed by Topic Modeling using Latent Dirichlet Allocation (LDA).**

Keywords*— Instagram, hashtag, bigram, topic modeling, text mining

I. INTRODUCTION

Distribution of information through social media so quickly in this decade has triggered the government to take part in the use of social media. At this time, it was recorded that Indonesia was a country with a number of gadget users as many as 172 million people, with 60 million people in it were millennials [1]. The Indonesian Ministry of Health is a ministry that deals with the health sector, including the formulation and implementation of the health sector. In order to support the delivery of information with unlimited space and time, the Ministry of Health provides an Instagram social media account that is actively used. The name of Instagram account is kemenkes_ri with 513 thousand followers, and 1,215 posts were taken on January 3, 2020. The account contains health information as well as activities carried out by the Ministry of Health. From the large number of posts that have been made, of course, analysis can be done related to the discussion or text caption information in it.

Analysis of social media texts using natural language processing is currently common, and such analysis can be in the form of classification, clustering, or forecasting. Grouping data using cluster methods is done on Instagram hashtag caption analysis. The results of the report can find out the best hashtag suggestions to apply for the promotion of a product or service [2]. Other studies, using classification methods to determine the condition of disease severity in an area by using tweet text, the results of the survey can indicate the location with severity level so that it can be used as one of the parameters of rapid decision-making [3]. Analyze the prediction of individual welfare levels through Facebook social media conversations. Through this analysis, the level of welfare is not only determined by economic status but can be analyzed through the expressions they write [4].

From a number of previous studies on social media data analysis, in this study, we tried to analyze the trending hashtag and topic frequency on the Instagram account of the Ministry of the Republic of Indonesia. The results of the analysis are expected to provide information on how effective and popular the level of hashtag usage is at a certain time frame. In addition, the results of the topic frequency analysis are expected to provide information on what topics are often discussed on the Ministry of Health's Instagram account.

II. METHOD

The architecture used in trends analysis of health information on Instagram account of the Ministry of Health Republic Indonesia can be seen in Figure 1.

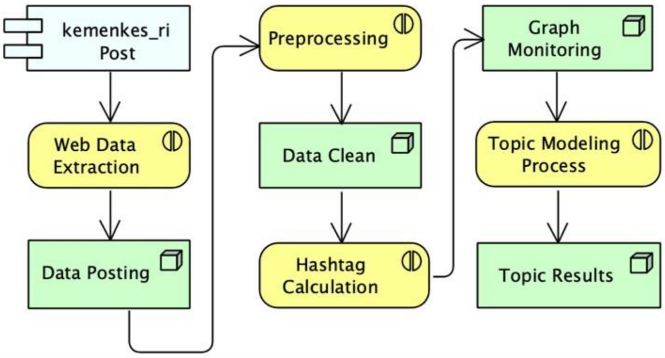
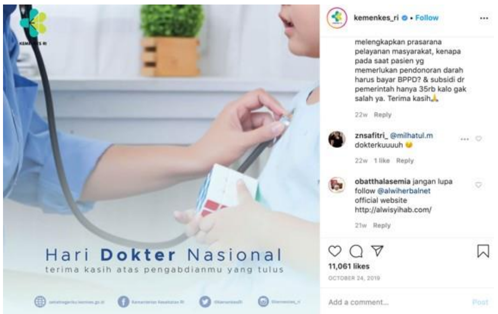


Fig. 1. Architectural model

A. Web Data Extraction

Web data extraction is the initial stage where data is collected from web page sources [5]. At this stage the information is taken from social media Instagram [6], the information is the posting data from the Ministry of Health account with the Ministry of Health account username with a total number of posts of 1,215 taken on January 03, 2020. An example post on the health ministry's Instagram account can be seen in Figure 2.



kemenkes_ri • Follow

melengkapkan prasarana pelayanan masyarakat, kenapa pada saat pasien yg memerlukan pendonoran darah harus bayar BPPD? & subsidi dr pemerintah hanya 35rb kalo gak salah ya. Terima kasih🙏

22w Reply

znsafitri_ @milhatul.m dokterkuuuuh 😉

22w 1 like Reply

obatthalasemia jangan lupa follow @alwiherbalnet official website http://alwisyihab.com/

21w Reply

11,061 likes

OCTOBER 24, 2019

Add a comment... Post

Fig. 2. Ministry of Health's Instagram post

B. Preprocessing

The preprocessing stage is the stage where the data will be normalized in the form and format of the data so that it produces the appropriate data to continue at the analysis stage [7]. In this preprocessing, caption text will be taken for each post, then proceed to the following steps:

- Cleaning symbols that are not recognized by the text: cleaning is focused on the presence of symbols from the text of the programming language that would be unnecessary. For example, in the use of backslash n (\ n), which is considered to enter.
- Tokenization in paragraphs to produce a sentence: at this stage, the paragraph will be broken down into several sentence forms [8].
- Tokenization in sentences to produce words: at this stage continues from the second stage, tokenization is carried deeper into the breaking of sentences into words.
- Stop word removal on unneeded words: the results of word tokenization are continued towards cleansing words that are not needed to enter the analysis phase. These words include conjunctions or additional words that are not necessary ("yes", "so", "wow", etc)[9].
- Fetching hashtags at each caption post: at this stage, the focus is on using hashtags that are implied in the caption, all hashtags will be taken and stored to proceed into the analysis phase [10].

The result of preprocessing stage is a collection of hashtags has been grouped according to using each month. The hashtag data will proceed to the hashtag calculation phase. In addition, the caption text data that has been cleaned and structured according to the format will be continued into the topic modeling process to find topics that are often discussed.

C. Hashtag Calculation

Hashtag Calculation Phase is the stage where hashtag data will be calculated based on the level of frequency of usage [11] each month. Then the amount will be calculated for one year to find out the top-ten hashtags most often used by the Ministry of Health in one year. The result of this hashtag monitoring is a graph of the level of development of the use of top-ten hashtags for one year.

D. Topic Modeling Process

The results of the top-ten hashtag are continued into a search for frequently discussed topic groups. The grouping of topics is taken from Instagram captions, and then feature extraction is done using the bigram technique [12]. Bigram technique is a way of decapitation based on two words that are interrelated to form the corpus module [13]. The results of the bigram corpus are counted into the conversion of the number of occurrences of interrelated words in the whole document by using the Term Frequency - Inverse Document Frequency (TF-IDF) feature [14]. TF-IDF is a numerical statistical value that represents the relevance of word categorization in a particular document [15]. The following formula from TF-IDF as shown in Equation (1) [2].

$$w_{t,d} = tf_{t,d} \times \log \frac{N}{df_t} \quad (1)$$

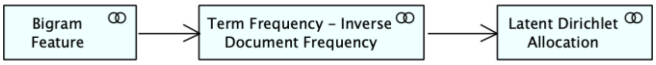
Which is

$tf_{",\$}$ = is the weight of a term t in document d

N = is the total number of documents

$df_{"}$ = is the number of documents containing term t.

The weighting using TF-IDF is continued to search for suitable topics based on hashtag usage. Topic search is done by utilizing the Python library [16], namely gensim [17] through a combination of TF-IDF and Latent Dirichlet Allocation (LDA) [18]. Overall the model used to determine the modeling topic is by Figure 3.



Bigram Feature → Term Frequency – Inverse Document Frequency → Latent Dirichlet Allocation

Fig. 3. The combination of bigram, TF-IDF and LDA

The results of the topic modeling in the form of topic groups that are often discussed by the top hashtags that have been obtained in the results of the hashtag calculation section.

III. Result and Discussion

From Instagram post data obtained from the Ministry of Health account of the Republic of Indonesia with a total of 1,215 posts in 2019. Then it can produce data on the results of using hashtags and search for topics that are often discussed by the following discussion.

A. Graph Monitoring

Monitoring charts are formed from the results of calculations on the top 10 hashtags that are often used for one year by the ministry of health through the social media platform Instagram. The calculation process uses a calculation based on querying the results of database computing using MongoDB [19]. The hashtag monitoring graph can be seen in Figure 4.

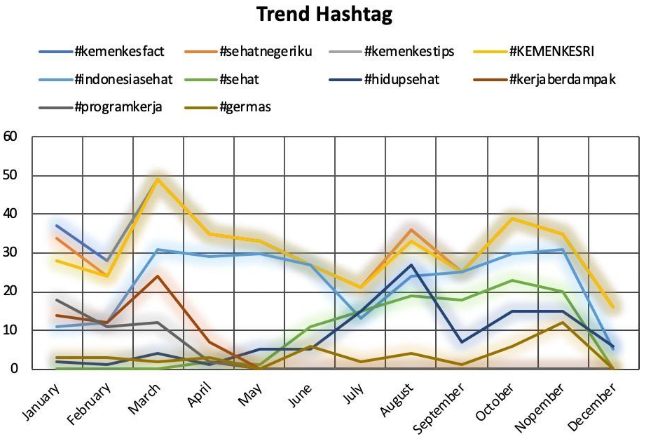


Fig. 4. Trend Hashtag (top ten)

It can be seen from Figure 4, the highest frequency of hashtag usage is in March, with an average usage of 27 times while the lowest hashtag usage is in December with an average of 8 times. Every Instagram post-activity needs to be analyzed in relation to the minimum frequency of posts every month, week, and even daily. This needs to be done to support the dissemination of information through a modern approach and ensure the existence of the Ministry of Health's own social media accounts.

As for the amount of each use of hashtag usage in one year, calculated from a total of 1,215 posts made by the ministry of health in 2019 can be shown in Figure 5.

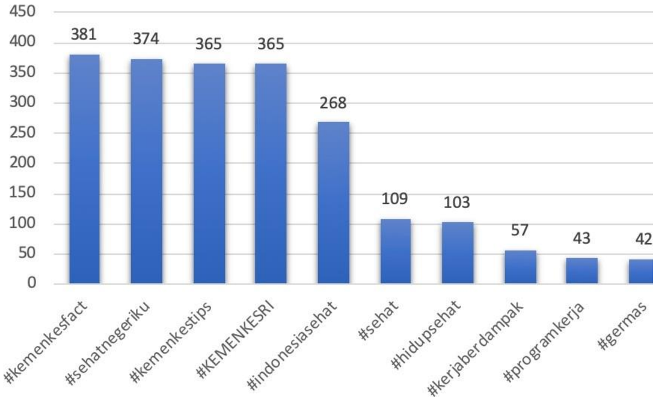


Fig. 5. Hashtag usage in one year

Figure 5 shows the most frequent use of hashtag #kemenkesfact with a total of 381 uses, followed by #sehatnegeriku for a total of 374 and #kemenkestips, #kemenkesri with a total of 365. A hashtag that is rarely used is #germas, with a total of only 42 users in a year.

These results contrast with the results of Rapat Kerja Kesehatan Nasional (Rakernas) on 12 February 2019, which one of them resulted in the launch of 8 innovations of the ministry of the health program in 2019 [20]. The information program included more emphasis on monitoring maternal and newborn deaths, public awareness-raising, and stunting prevention. With the least use of the #germas hashtag on the health ministry's Instagram post, germas socialization through social media is still considered relatively low, because #germas is ranked last of the top 10 hashtags.

Whereas the hashtag #kemenkesfact and #kemenkestips put more emphasis information for about health information and programs that have been carried out by the Indonesian Ministry of Health. It does not yet lead to the appropriate target keyword/hashtag target level; for example, the #stunting keyword or hashtag does not make it into the top ten hashtag levels.

B. *Topic Results*

The results of hashtag usage frequency show that #kemenkesfact is the hashtag with the most frequent total usage. For this reason, it is necessary to determine the topic that is often discussed in the hashtag. From the Topic Modeling Process that has been carried out, it is determined five groups of topics that are appropriate and have a level of closeness between topics with two-dimensional form, can be seen in Fig. 6.

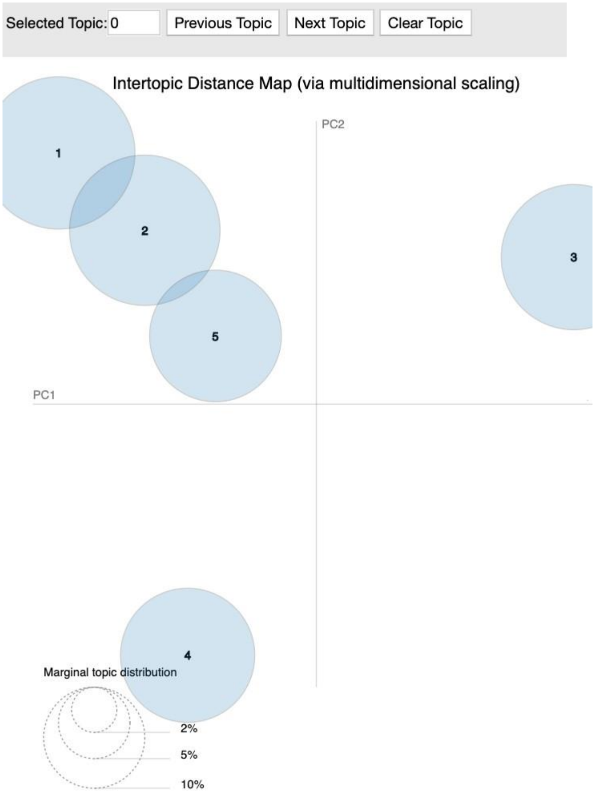

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribution
2%
5%
10%

Fig. 6. Intertopic Distance Map

They are viewed from Figure 6. topics 1 and 2, topics 2 and 5 have a high level of topic closeness, while topics 3 and 4 are at a distant point without contacting others. It can be said that topics 1, 2, and 5 have a higher level of inter-topic closeness compared to topics 3 and 4. By taking one topic, topic 3, it can be seen the relevant value of each term by Figure 7.

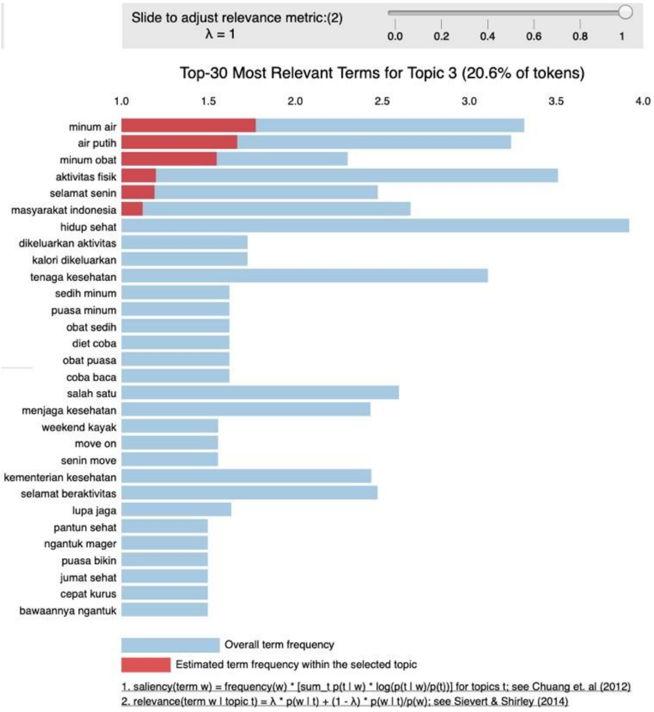


Fig. 7. Relevant Terms

As shown in Figure 7. there are 30 terms in the topic that corresponds to the hashtag #kemenkesfact. Within these terms, of course, several topics make sense and some that do not make sense in the logic of word reasoning. Unreasonable terms include dikeluarkan aktifitas, sedih minum, diet coba, salah satu, weekend kayak, senin move dan lupa jaga. Besides that, five terms have strong links with topic 3, including minum air, air putih, minum obat, aktifitas fisik, and masyarakat indonesia.

IV. Conclusion

This research was successful in representing data on the use of a hashtag in one year into a monitoring analysis graph every month. With the results of the analysis, it is necessary to study further the suitability of the use of hashtags on each post. In addition, there is a need for posting schedules or posting frequency every month because the total posts each month have a significant difference. Besides that, the use of bigram features in combination with TF-IDF still needs to be improved to produce better topic terms. The results of the topic terms obtained from the top hashtag namely #kemenkesfact more often discuss minum air, air putih, minum obat, aktifitas fisik, and masyarakat indonesia.

Acknowledgment

This research was funded by The Departement of Informatics, Faculty of Teknik dan Teknologi Informasi. The data supported by Center of Study and Data Analytic Services from Universitas Jenderal Achmad Yani Yogyakarta. With hope, this research can be beneficial for the State of Indonesia, especially in the field of health information analysis.

References

[1] R. Maulana, "screencapture-id-techinasia-pertumbuhan-pengguna-perangkat-mobile-indonesia-2020-01-03-13_07_03.pdf," *Techinasia*, 2019. [Online]. Available: https://id.techinasia.com/pertumbuhan-pengguna-perangkat-mobile-indonesia. [Accessed: 03-Jan-2020].

[2] 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.

[3] P. W. Cahyo and E. Winarko, "Model Monitoring Sebaran Penyakit Demam Berdarah di Indonesia Berdasarkan Analisis Pesan Twitter," Universitas Gadjah Mada Yogyakarta, 2017.

[4] H. A. Schwartz, "Predicting Individual Well-Being Through The Language of Social Media," in *Pacific Symposium on Biocomputing 2016*, 2016, pp. 516–527.

[5] T. Arabghalizi, B. Rahdari, and M. Brambilla, "Analysis and Knowledge Extraction from Event-related Visual Content on Instagram," *Res. Publ. Politec. di Milano*, pp. 1–12, 2017.

[6] S. Ravn, A. Barnwell, and B. B. Neves, "What Is ' Publicly Available Data '? Exploring Blurred Public – Private Boundaries and Ethical Practices Through a Case Study on Instagram," *J. Empir. Res. Hum. Res. Ethics*, vol. 15, no. 1–2, pp. 40–45, 2019.

[7] M. Habibi and P. W. Cahyo, "Journal Classification Based on Abstract Using Cosine Similarity and Support Vector Machine," vol. 4, no. 3, pp. 48–55, 2020.

[8] T. Jo, "Text Indexing," in *Text Mining. Studies in Big Data*, Vol 4., Springer, Cham, 2018, pp. 19–40.

[9] P. D. Nurfadila, Aji Prasetya Wibawa, Ilham Ari Elbaith Zaeni, and Andrew Nafalski, "Journal Classification Using Cosine Similarity Method on Title and Abstract with Frequency-Based Stopword Removal," *Int. J. Artif. Intell. Res.*, vol. 3, no. 3, 2019.

[10] J. B. Negrón, "#EULAR2018: The Annual European Congress of Rheumatology—a Twitter hashtag analysis," *Rheumatol. Int.*, vol. 39, no. 5, pp. 893–899, 2019.

[11] P. C. Karmaker and M. S. Hossen, "Performance Analysis of Frequency and Graph Theoretic Based Text Summarization," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–5.

[12] D. Alnahas and B. B. Alagoz, "Probabilistic Relational Connectivity Analysis of Bigram Models," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1–6.

[13] N. P. R and A. Ahsok, "Effect of Feature Reduction using Bigram Technique for detection of Forged Reviews," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 2481–2485.

[14] I. Alsmadi, "Term weighting scheme for short-text classification : Twitter corpuses," *Neural Comput. Appl.*, vol. 8, 2018.

[15] S. Qaiser and R. Ali, "Text Mining : Use of TF-IDF to Examine the Relevance of Words to Documents," vol. 181, no. 1, pp. 25–29, 2018.

[16] P. W. Cahyo and A. I. Wicaksono, "DJANGO FRAMEWORK AND PYTHON-GAMMU AS MIDDLEWARE SMS BROADCAST," *Compiler*, vol. 8, no. 1, pp. 27–34, 2019.

[17] M. D. Hoffman, D. M. Blei, and F. Bach, "Online Learning for Latent Dirichlet Allocation," in *Advances in Neural Information Processing Systems 23*, J. D. L. and C. K. I. W. and J. S.-T. and R. S. Z. and A. Culotta, Ed. Curran Associates, Inc., 2010, pp. 856–864.

[18] N. Lee, E. Kim, and O. Kwon, "Combining TF-IDF and LDA to generate flexible communication for recommendation services by a humanoid robot," *Multimed. Tools Appl.*, vol. 77, no. 4, pp. 5043–5058, 2018.

[19] M. R. Ma'arif, A. Priyanto, C. B. Setiawan, and P. W. Cahyo, "The Design of Cost Efficient Health Monitoring System based on Internet of Things and Big Data," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 2018, pp. 52–57.

[20] Kementerian Kesehatan Republik Indonesia, "Kemenkes luncurkan 8 inovasi program kesehatan," *kemkes.go.id*, 2020. [Online]. Available: https://www.kemkes.go.id/article/view/19021300037/kemenkes-luncurkan-8-inovasi-program-kesehatan.html. [Accessed: 31-Mar-2020]