IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.13, No.4, October 2019, pp. 399~408
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.50574

Clustering User Characteristics Based on the influence of Hashtags on the Instagram Platform

Muhamad Habibi*[1], Puji Winar Cahyo[2]
[1,2]Department of Informatics, FTTI UNJANI, Yogyakarta, Indonesia
e-mail: ***[1]muhammadhabibi17@gmail.com**, [2]pwcahyo@gmail.com

Abstrak

Instagram merupakan media sosial yang sangat potensial digunakan untuk meningkatkan kesadaran terhadap suatu produk. Kurang lebih 70% pengguna menghabiskan waktu mereka untuk mencari suatu produk di Instagram. Selama ini, Banyak masyarakat yang melakukan penyebaran informasi atau promosi melalui Instagram dengan kurangnya memperhatikan sasaran. Sehingga tidak jarang informasi yang disebarkan merupakan informasi yang kurang tepat serta tidak sesuai dengan karakteristik pengguna. Penelitian ini bertujuan untuk melakukan klasterisasi karakteristik pengguna Instagram berdasarkan kecocokan hashtag. Metode yang digunakan dalam penelitian ini adalah metode K-Means Clustering. Berdasarkan hasil eksperimen, penelitian ini berhasil melakukan klasterisasi pengguna Instagram berdasarkan kecocokan hashtag pada caption teks. Selain itu, TF-IDF dapat digunakan sebagai salah satu feature yang cocok pada metode K-Means Klastering. Hasil analisis hashtag "#kopi" menghasilkan saran hashtag yang dapat digunakan untuk promosi suatu produk yang berkaitan dengan kopi, diantaranya adalah hashtag #coffeeshop dan #coffee dengan jumlah penggunaan sebanyak 14968 caption.

Kata kunci*— Klasterisasi, Instagram, K-Means, Media Social, Analisis Teks*

Abstract

Instagram is a social media that has the potential to be used to increase awareness of a product. Approximately 70% of users spend their time searching for a product on Instagram. Many people promote their products with a lack of attention to the target. So that not infrequently the information distributed is inaccurate information and not following user characteristics. This study aims to cluster the characteristics of Instagram users based on hashtag compatibility. The method used in this study is the K-Means Clustering method. Based on the results of the experiment, this research succeeded in clustering Instagram users based on the hashtag match on the text caption. Besides, TF-IDF can be used as a feature suitable for the K-Means Klastering method. The results of the hashtag "#kopi" analysis resulted in hashtag suggestions that can be used for the promotion of a product related to coffee, including the hashtag #coffeeshop and #coffee with total usage of 14968 captions.

Keywords*— Clustering, Instagram, K-Means, Social Media, Text Analysis*

Received October 14th,2019; Revised October 21th, 2019; Accepted October 23th, 2019

1. INTRODUCTION

Instagram is a popular social media that is experiencing very rapid development among internet users nowadays. Based on data from Techcrunch, Instagram has increased from 800 million users in September 2017 to 1 billion users in June 2018 [1]. In the Asia Pacific region, Indonesia is the country with the most user base Instagram. Namely, with the number of active users as much as 45 million [2]. Instagram is a social media platform that facilitates us to share information. Information uploaded on Instagram consists of three basic parts, namely: Uploaded image, caption, which is the core of the message, and hashtag. Hashtags written with the # symbol are used to index keywords or topics on Instagram. This function is made on Instagram and allows users to follow topics they are interested in easily.

Instagram is a social media that has the potential to be used to increase brand awareness and introduce a product. 70% of users spend their time searching for a brand of a product on Instagram. Instagram allows us to promote brands or products easily and authentically without selling directly to customers [3]. During this time, most people disseminating information or promotions on Instagram often pay less attention to the target. They do not pay attention to the characteristics of users who have an interest in the news (information) or follow a more specific hashtag. So that not infrequently, the information distributed is inaccurate information and not by user characteristics.

Based on these problems, we need a model that can cluster user characteristics based on hashtag compatibility. This study aims to classify Instagram users based on specific hashtag matches. This research implements Term frequency (TF) as a feature. Term frequency (TF) is a standard notion of frequency in corpus-based natural language processing [4]. The application of term frequency can be made to extract text from students' comments towards lecturers [5]. TF-IDF (Term Frequency-Inverse Document Frequency) is a metric that is commonly used in the process of text categorization [6]. TF-IDF consists of two-component values, namely term-frequency and inverse document frequency. The use of TF-IDF works well with text mining methods [7].

K-Means method is used to classify user characteristics based on hashtags. K-Means is a clustering method that accommodates partitional algorithms. K-Means clustering aims to optimize a function to calculate the distance space between objects and the centroid (midpoint) cluster [8].

Research related to the use of Instagram data has been widely used, including detection of selfies on Instagram [9], ranking keywords based on image captions on Instagram using TF-IDF [10]. As for other studies, namely research exploring the habits and involvement of Instagram by the Indonesian Government Ministry [11].

2. METHODS

This study uses text caption data on Instagram that contains specific hashtags. The stages in this research are preprocessing, feature extraction, and then the clustering process is carried out. The steps of the study can be seen in Figure 1.

The first steps *carried out in this study are:*

2. 1 Preprocessing

Preprocessing data is a process where the text to be classified is cleaned and prepared before the document is analyzed [12] and done to avoid incomplete data, interference with data, and inconsistent data [13]. At this step, it begins by utilizing web data extraction. Text caption

data on Instagram is collected using certain words as search keywords. The next step is preprocessing data; the steps in preprocessing include removing the HTML elements contained in the data that has been collected. Next, look for the hashtags listed in the Instagram text caption.

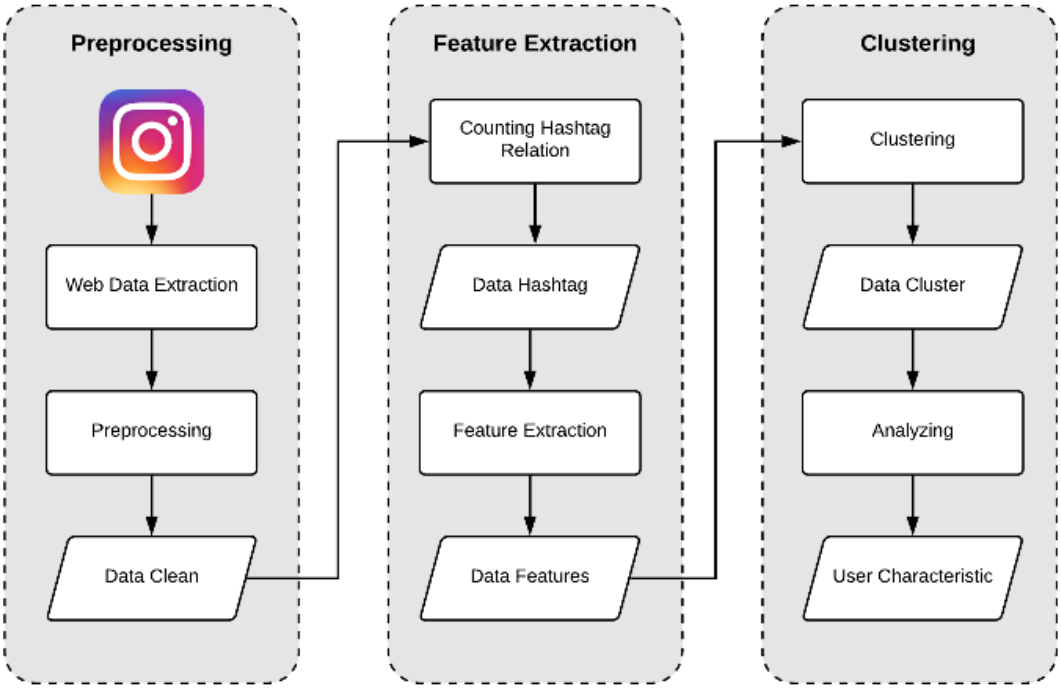

Figure 1 Research Stages

2. 2 Feature Extraction

Feature extraction is an extraction process to identify the entities in question [14]. This stage searches for the value of Term Frequency (TF) and IDF (Inverse Document Frequency) on the data text that has been collected. To see how important the appearance of tokens in a corpus. The calculation of the TF-IDF weight value is obtained from equation (1) below [15]:

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

Which is,

$tf_{t,d}$ = is the weight of a term t in document d
N = is the total number of documents
df_t = is the number of documents containing term t.

2. 3 Clustering

Clustering is the use of data mining techniques where groups of the same object are combined to form clusters, and this cluster is different from objects in other clusters [16]. At this stage, K-Means is used to classify user characteristics based on hashtags. The K-Means calculation is obtained from equation (2) below:

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n}(x_j - c_j)^2} \quad (2)$$

Which is,

d = distance
n = number of objects
j = (starts from 1 to n)
x_j = feature object to j with respect to x
c_j = centroid feature to j

In general, the steps for the K-Means algorithm [17] are:

a. Determine the number of clusters (k)
b. Determine the centroid
c. Does the centroid analysis change based on the means for each feature of the object (data)?
   - [Yes], change centroid to new centroid
   - [not], completed
d. Calculate distance space (proximity) of objects with centroids
e. Grouping objects based on the proximity of objects with centroid

The next step is to analyze and cluster the characteristics of users who follow the hashtag and create demographics of the data.

3. RESULTS AND DISCUSSION

The dataset used in this study was taken in 40 days, namely from 3 July to 13 August 2019, totaling 99,237 data captions. Text caption data from Instagram collected using the keyword "#kopi". The following is an example of text caption data that has been collected, as seen in Table 1.

Table 1 Example Caption Text data

No.	Caption Text
1	Bakar lemakmu sekarang kak! KOPI HITAM tanpa gula juga memiliki kalori 0% lohh!! Jangan salah memilih kopi hitam juga ya kak. Karena hanya yang berkualitas saja yang dapat memberikan manfaat yang baik buat tubuh kakak. Sumber: ottenmagazine #SavaUntukIndonesia #savacoffee #kopisava #ngopi #kopi #sava #kopiindonesia #kopinusantara
2	Cinta itu seperti kopi. Karena pahit kejujuran lebih berarti daripada manis kebohongan yang berujung melukai. Kopi Bubuk 250gr / 39k #kopi #kopihitam #kopiindonesia #coffeeshop #ngopi #kopinusantara #kedaikopi #barista #espresso #cafe #indonesia #kopisusu #kopimalam #manualbrew #pecintakopi #coffeetime #coffeelover #filosofikopi #coffeeaddict #kopilokal #kuliner #nongkrong #latteart #humbahas #medan #baristaindonesia #coffeeshop #vsco #vscofilter #danautoba
3	Mau es kopi kekinian yang lagi tren di kalangan anak-anak muda? Dan yang pasti punya banyak varian rasa yang bisa kamu nikmati! Hanya di @CetrooCoffee Ingin jadi mitra franchise #cetroocoffee? Hubungi kami di: - Telp/WA : 0813-2895-6766 - WEB : www.cetroocoffee.com #bisnis #waralaba #franchise #franchiseindonesia #kopi #coffee #kopiindonesia #bisnissukses #tipsbisnis #kedaikopi #kopisusu #bisnismurah #modalkecil #coffeemix #indonesiamerdeka #kopinikmat #bisniskopi #bisnisminuman #waralabaindonesia #bisniskopi #bisnisindonesia #bisnissukses #usahamodalminim #kopikekinian #usahakopi #waralabaminuman #waralabakopi #coffeefranchise"

The next step is to search term frequency for each caption of the text that has been collected. Based on the keyword "#kopi" there are 10 hashtags that have the most occurrence, including #cooffee, #cafe, #coffeegram, #coffeelover, #coffeeshop, #barista, #coffeetime, #kopihitam, dan #kopiindonesia. 10 hashtags with the most appearances, as seen in Table 2.

Table 2 10 hashtags with the most appearances

No.	*Hashtag*	*Term Frequency*
1	#coffee	33567
2	#coffeeshop	19318
3	#kopiindonesia	17442
4	#kopihitam	14945
5	#coffeelover	13489
6	#coffeetime	12889
7	#cafe	11923
8	#barista	11664
9	#coffeeaddict	11333
10	#coffeegram	10603

After getting 10 hashtags, the next step is to sort according to the level appearance of the hashtag. In the next step, hashtags that have a high rate of occurrence are grouped based on the relationship of appearance in a text caption. The hashtags that appear together in a text caption are grouped. 9 hashtag groups are obtained the most and are interrelated. The hashtag group consists of 3 different hashtag combinations, as shown in Table 3.

Table 3 The combination of the appearance of the hashtag

No.	Hashtag	Number of Documents
1	#kopi, #coffeeshop, #coffee	14968
2	#kopi, #coffeelover, #coffee	12143
3	#kopi, #coffeetime, #coffee	11355
4	#kopi, #coffee, #coffeeaddict	10227
5	#kopi, #cafe, #coffee	9991
6	#kopi, #coffeegram, #coffee	9905
7	#kopi, #barista, #coffee	9816
8	#kopi, #kopiindonesia, #coffee	8986
9	#kopi, #kopihitam, #kopiindonesia	7706

The following is a graph resulting from the appearance of a hashtag based on Table 3, as shown in Figure 2.

Based on Figure 2 it can be seen that the word with the hashtag "#kopi" is the root of the graph because searching data on Instagram uses the keyword "#kopi". Words with the hashtag "#kopi" are hashtags that always appear together with other hashtags so that in each group, there is a word with the hashtag "#kopi". After grouping the hashtag into 9 groups, the next step is cluster formation using K-Means Clustering. Hashtag data in each group were clustered into 5 cluster groups.

In this section, cluster discussion will be conducted on the hashtag 5 group, which is a hashtag group that contains a combination of hashtag #kopi, #cafe, and #coffee. The following is the result of hashtag clustering, as shown in Table 4.

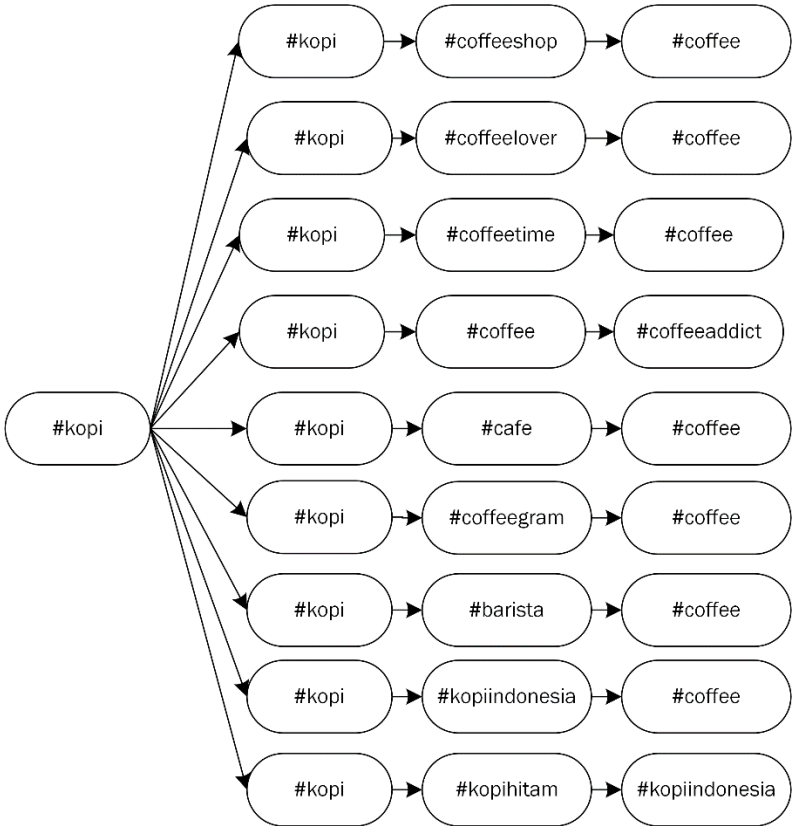

Figure 2 Graph of Hashtag Grouping

Table 4 Group 5 Hashtag Clustering Results

Cluster	Hashtag
1	#coffee #cafe #coffeetime #coffeeshop #coffeelover #espresso #barista #coffeelovers #love #food #coffeeaddict #latte #breakfast #coffeeholic #tea #kopi #instacoffee #cappuccino #latteart #caf #foodie #foodporn #coffeegram #coffeebreak #art #instagood #like #caffeine #photography
2	#coffeeshop #coffee #coffeetime #cafe #coffeelover #barista #espresso #coffeeaddict #latte #coffeelovers #coffeeholic #latteart #kopi #coffeegram #instacoffee #specialtycoffee #cappuccino #coffeebreak #coffeehouse #food #kedaikopi #foodie #coffeebean #breakfast #coffeelife #caf #caffeine #foodporn #baristalife
3	#milenials #kopi #kopimilenials #coffee #kedaikopi #kedaikopidenpasar #kedaikopibali #bali #kintamani #kopikintamanai #coffeeshop #cafe #cafebali
4	#coffee #cafe #instacoffee #toptags #coffeetime #cafelife #caffeine #coffeebreak #coffeefirst #coffeeshopvibes #butfirstcoffee #coffeeaddict #coffeegram #coffeeoftheday #ilovecoffee #coffeelover #coffeelovers #coffeecup #coffeeholic #coffiecup #coffeelove #coffeefliicks #coffeelife #coffeeplease #ig_coffee #thehappynow #kopi
5	#kopi #coffee #kopihitam #kopiindonesia #coffeeshop #ngopi #kopinusantara #kedaikopi #barista #espresso #cafe #indonesia #kopisusu #kopimalam #manualbrew #pecintakopi #coffeetime #coffeelover #filosofikopi #kopibali #latte #robusta #coffeeaddict #kopijakarta #kopilokal #kuliner #nongkrong #latteart

Based on the hashtag data in Table 4, an analysis of the characteristics of the resulting clusters can be carried out. Cluster 3 and Cluster 5 Have quite significant differences compared to the other Clusters (Cluster 1, Cluster 2, and Cluster 4). Cluster 3 is a hashtag cluster that has a lot of location hashtags, namely the city of Bali (#kedaikopidenpasar, #kedaikopibali, #bali, #kintamani #kopikintamanai).

Cluster 5 is a hashtag cluster with relation to the use of hashtag archipelago coffee or local coffee in Indonesia with hashtag relation in it as follows (#kopiindonesia, #kopinusantara, #indonesia, #kopilokal, etc.).

Whereas the hashtag group in this cluster (Cluster 1, Cluster 2, and Cluster 4) is a cluster that has the best hashtag relation. The difference in cluster 1 emphasizes the relationship to the use of social media hashtag in general, namely (#instacoffee, #coffeegram, #instagood, # photography). Cluster 2 emphasizes the relationship of the hashtag to various types of coffee, including (#espresso, #latte, #latteart, #caffeine). Whereas Cluster 4 is a hashtag relation cluster that emphasizes the use of English hashtags, including the following (#cafelife, #coffeebreak, #coffeefirst, #coffeeshopvibes, #butfirstcoffee, etc.).

From the five clusters, cluster graph representation can be made to make homogeneous data distribution easier. The cluster has the same characteristics of data between one data with other data. Cluster visualization is shown in Figure 3 as follows.

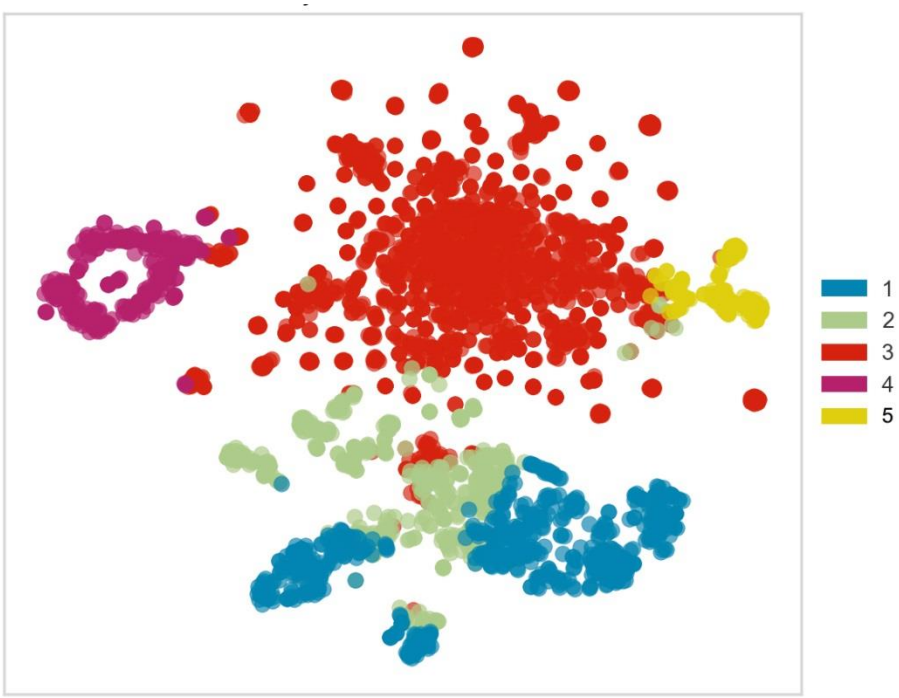

Figure 3 Cluster Visualization

In Figure 3 it can be seen that Cluster 5 has the least data distribution with a discussion of the hashtag in general. In this cluster, the emphasis on the relation of the characteristics of the Indonesian or Indonesian coffee hashtag. Examples of users in cluster 5 can be seen in Table 5. While Cluster 3 has the most data distribution compared to other clusters. The relationship of hashtag characteristics in this cluster emphasizes more on the discussion of Bali's local coffee. The following are examples of some of the users found in cluster 3, which can be seen in Table 6.

Table 5 account list in cluster 5

Cluster	user
Cluster 5	@karibucafe @italianstylecaffeuno_germany @coffee.fannatic @ps.sunday @jess_andthebeanstalk @vermaak.paul @coffeewithfriend @dariennfd @superc_gjs @pyk_coffee @pykcoffee @sababa_cafe @mevillez_ @tareq_alshakaili @kingswearcoffee @coffee_lover_greece @petrapajek @oldpueblocoffee @aterliercamaieu @llea_official.co @beanoscafe_ad @saudisocialbutterfly @feroniamary @cafeteria_magnolia @jessdk1973 @cafeboladeoro @fore.moi @nescafesurat @bigbrewlittlebrew @dashcoffeeandbakery @elisabeth19.a @jude_sumu

Table 6 account list in cluster 3

Cluster	user
Cluster 3	@milenials.eskopi @dwikapuspitasari @_dhea4600 @27_wulan @nikendwulandari @ankyindtr @fridaard_ @steciasanjaya_88 @agung_nadiaa @dewi.sita @vivianclaudyaa @erykafbrnt @ayukharismafeb @devanibharadwipa @vidiagizelda @crownwina @anggundianap @arimascempaka @anggitapuspadewi @roindah_ @natasyamila_ @nadyaputri13 @igadewi03 @pandetia48 @santhiayumi @cathcat_ @galdinakharen

4. CONCLUSIONS

This research has succeeded in clustering Instagram users based on the suitability of specific hashtags on Instagram text captions. Based on experiments that have been carried out, TF-IDF can be used as a useful feature. TF-IDF can help the process of identifying the appearance of hashtags when searching. The results of the hashtag "#kopi" analysis produce hashtag suggestions that can be used for the promotion of a coffee-related product. These hashtags are #coffeeshop and #coffee with 14968 caption usages. K-Means method can be used as a suitable clustering method for the clustering process of Instagram user characteristics according to the hashtag effect.

ACKNOWLEDGEMENTS

This research was conducted in 2019 of Penelitian Dosen Pemula (PDP) scheme funded by the Directorate of Research and Community Service Directorate General of Research and Development Strengthening (DRPM) of the Ministry of Research and Higher Education (Kemristekdikti) of the Republic of Indonesia. The research team would like to thank the DRPM Kemristekdikti for allowing the research team to add insight and knowledge through research in this scheme. Hopefully, this research can bring benefits to the progress of the Indonesian nation

REFERENCES

[1] J. Constine, "Instagram hits 1 billion monthly users, up from 800M in September," 2018. [Online]. Available: https://techcrunch.com/2018/06/20/instagram-1-bill. [Accessed: 19-Aug-2019].

[2] A. ADI and A. HIDAYAT, "45 Juta Pengguna Instagram, Indonesia Pasar Terbesar di Asia," 2017. [Online]. Available: https://bisnis.tempo.co/read/894605/45-juta-pengguna-instagram-indonesia-pasar-terbesar-di-asia. [Accessed: 19-Aug-2019].

[3] A. Collins, "Instagram Marketing," 2018. [Online]. Available: https://www.hubspot.com/instagram-marketing. [Accessed: 19-Aug-2019].

[4] 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.

[5] M. Habibi and Sumarsono, "Implementation of Cosine Similarity in an automatic classifier for comments," vol. 3, no. 2, pp. 38–46, 2018.

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

[7] 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.

[8] D. J. Bora and A. K. Gupta, "Effect of Different Distance Measures on the Performance of K-Means Algorithm : An Experimental Study in Matlab," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 2501–2506, 2014.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] I. Hemalatha and A. Govardhan, "Preprocessing the Informal Text for efficient ALGORITHM FOR," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 1, no. 2, pp. 58–61, 2012.

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

[15] C. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2009.

[16] A. P. Clustering, "A Survey of Clustering Techniques and Algorithms," *2nd Int. Conf. Comput. Sustain. Glob. Dev.*, pp. 3014–3017, 2015.

[17] P. W. Cahyo, "Klasterisasi Tipe Pembelajar Sebagai Parameter Evaluasi Kualitas Pendidikan di Perguruan Tinggi," *Teknomatika*, vol. 11, no. 1, pp. 49–55, 2018.