International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 4
ISSN: 2088-5334

Finding the Right Influencers on Instagram for Endorsement Product Using Text Mining

Hedy Pamungkas [a,*], Riswan Haryo Yudhianto [a], Bern Jonathan Sembiring [a], Indra Budi [a]

[a] *Faculty of Computer Science, University of Indonesia, Jakarta, Indonesia*
*Corresponding author: *hedy.pamungkas@ui.ac.id*

Abstract— **The growth of the Internet has significantly changed our way of life. Social media is being used not only for getting connected with people but also for media sharing, diffusing information, and even marketing. Social media marketing has a significant impact because it can target campaigns appropriately based on the segmentation of the products to be marketed. Furthermore, this approach can provide campaign costs that are less expensive than other traditional methods. One of the methods which are generally used is an endorsement. With the popularity of social media, a new type of endorsement called social media endorsement has been born. This type of endorsement incorporates social media influencers to promote products and goods. However, product and brand owners find it difficult to select suitable influencers to endorse their products. Therefore, we try to solve this problem by creating a recommender system using the clustering method. We collect data from Instagram, one of the most popular social media platforms. The result showed that three clusters produced out of the data had high quality. In addition, we applied Silhouette Coefficient to validate the result, which produced a positive result on 7.277×10^{-3}. With such a result, we conclude that this model could be used to categorize which brands or products an influencer normally endorse and recommend product or brand owners if an influencer is suitable based on clustering.**

Keywords— **Influencer; Instagram; endorsement; clustering; marketing.**

Manuscript received 4 Feb. 2022; revised 22 May 2022; accepted 6 Jun. 2022. Date of publication 31 Aug. 2022.

I. Introduction

The Internet has become inseparable from our daily life. It has shown significant growth in the past few years, especially in Indonesia. As the technology grows, so do its users. In 2018, it was recorded that there were 171,17 million Internet users in Indonesia; this number escalated to 196,71 million in 2020 [1]. This increase in number has supported other technologies that use the Internet to grow. One of them is social media. Initially used as a way to interact and connect with others virtually, social media has advanced so much this past few years. It has branched out to several different types, such as social network sites, forums, wikis, media sharing platforms, and rating and review communities [2].

Currently, the social media platforms with the most users are Facebook, with over 2 billion users, followed by YouTube with around 1,9 billion users, and Instagram in the third place with about one billion users [3]. Instagram is a social media platform that may be used to promote a brand or a product. On Instagram, 70% of people look for a product's brand. Instagram allows us to promote companies and products without selling directly to clients easily and authentically. During this period, most users using Instagram to disseminate information or promotions pay less attention to the intended audience. They are unconcerned about the qualities of users who are interested in current information. As a result, it is not uncommon for erroneous information to be sent, regardless of user attributes [4].

As a social media platform, Instagram is used not only for social interaction and sharing but also for promotion and product marketing [5]. For marketing and promotions, it is common to agree with public figures, such as celebrities, to advertise or promote brands or products to reach a wider audience. The practice is called endorsement [6]. This common practice has been around a long time and has become one of the most popular ways to promote products [7], [8]. In addition, there is a surge of a new type of public figures known as influencers, such as bloggers, vloggers, and celebgrams [9]. An influencer or social media influencer has successfully collected followers by building an online persona on social media [10] or influencing their followers through social media [11].

Product owners or companies usually offer these people to endorse their products by promoting them to their audiences through the post they send on their accounts. This process creates a new type of marketing called influencer marketing [12]. This type of marketing may bring fresh air to how companies or product owner promote their products. However, new problems have also arisen. Mainly, it is how to select the right influencer for the product. Considerations such as influencers' followers' demographics and influencers' persona should be included while selecting influencers for endorsement [6]. Therefore, a need for a recommendation arise to solve this problem. Some research on endorsement recommender systems has been done in the past, such as on [2], [6]. By using the recommender system, product owners are expected to choose which influencers fit for endorsing their products based on their profile and content so that the marketing can find its targeted segments.

To achieve said objective, we use social media analytics through the use of text mining from influencers' posts. One method that can be used is clustering by using K-Means. By doing so, we expect to be able to categorize influencers by their content so that we know which product is associated with respected influencers. In addition, the results of this study can be used as a model for endorsement recommender system for influencers and product owners.

II. Material and Method

A. Text Mining

Text mining is the process of extracting patterns from many text data sources in the form of usable information and knowledge [13]. Text mining is a subset of data mining that has additional steps to process unstructured textual data format for information acquisition [12]. As a subset of data mining, text mining can be used on cases that are otherwise unsolvable by usual means of data mining, specifically for textual data, such as emails, documents, or web pages. Therefore, it is usually used on written data sources [14], [15]. Text mining is done automatically using computers. Text mining aims to find words that represent the document, and their relations will then be analyzed using a statistical method such as classification or clustering. Generally, text mining processes are divided into pre-processing, feature selection, text representation, and analysis/evaluation.

Text mining is a set of tactics and procedures for recognizing and mining certain data, such as conclusions and mental states. It has generally been about content boundaries, such as offering an opinion on something, regardless of whether someone has a good, neutral, or negative review of anything. Text mining has traditionally been a topic or an administration whose survey has been made public on the Internet. This may explain why opinion mining and text analysis are frequently used as alternatives, despite our belief that seeing emotionally charged opinions is more precise.

B. Clustering

Clustering is a machine learning algorithm to group data into several separated groups based on their characteristics. By clustering, the data with similar characteristics will be grouped with another with similar characteristics [16]. Clustering can also be defined as the process of learning to put together scattered data into homogenous groups based on their distance value to the center of groups, usually called centroid [13]. Clustering can be done with a hierarchical or non-hierarchical approach or a combination of both. Most research use either hierarchical, which takes more time but produces a better result, or non-hierarchical, which is faster but harder to understand. However, the combined approach is more accurate [17], [18].

Clustering is a common machine learning technique used in social media analysis to reduce scatter in data [19]. It is mainly used for topic extraction or modeling [20], [21]. Discovered topics can be used for a specific purpose, such as recommendations or labeling. Clustering can be used to aid news analysis by automatically grouping items with similar characteristics. A cluster is a group of things that have been grouped due to their similarity or proximity [13].

Using clustering, we can identify dense areas, find overall distribution patterns, and discover interesting correlations between data characteristics. Data mining aims to identify ways to cluster enormous databases effectively and efficiently. Clustering in data mining involves many requirements, including scalability, the ability to handle many attribute types, the ability to manage high dimensionality, the ability to handle noisy data, and the ability to be simply translated [22]. This data clustering aims to reduce the clustering process's objective function, which normally aims to reduce variability within a cluster. And try to keep the differences between clusters to a minimum.

C. K-Means Clustering

The K-Means algorithm clusters data by finding the cluster center point (centroid) closest to the data. K-means clustering is one of the clustering methods available and one of the most popular ones [23], [24]. The benefit of utilizing the K-means approach for clustering is that it does not require a huge number of iterations to achieve effective clustering results, making it ideal for use with vast volumes of data [25]. K-means clustering tries to group a collection of data based on the tendency of each individual data to group with others and partition them into k clusters. The group's tendency is based on the distance of the data to the center of the cluster and the value of k is determined by the elbow method [26]. To measure the distance, Euclidean distance is commonly used, which is represented by the formula below:

$$\sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{1}$$

where x_i and y_i are points in Euclidean space with d dimensional. Sum of Squared Error is used for objective function which can be represented as below:

$$SSE = \sum_{i=1}^{d}\sum_{X_i \in C_k}(X_i - C_k)^2 \tag{2}$$

with cluster centroid:

$$C_k = \frac{\sum_{X_i \in C_k} X_i}{C_k} \tag{3}$$

The algorithm, based on k-means can be divided into the following steps [27]:

- Pick k random numbers and equivalent centroids Feature Selection
- For each data instance, assign it to the cluster which has closest centroid
- Update centroid and reassign instances

- Repeat until no changes

K-Means is used to group data by maximizing data similarity within a cluster and reducing data similarity between clusters. The cluster's measure of similarity is a function of distance. As a result, the smallest distance between the data and the centroid point is used to maximize data similarity.

D. Endorsement

Social media has become an indispensable marketing tool. Companies have begun to use social media to promote products, raise brand awareness, and improve customer interactions. Unlike traditional media, social media, on the other hand, empowers all users to create their material. Furthermore, because consumers trust friend's recommendations more than brand's advertisements, influencers can have a greater marketing impact on the audience than firms [28], [29]. As a result, businesses use influencer marketing to raise brand recognition and purchase intent among potential customers.

Endorsement is a practice to make brands noticeable to consumers to market the products. The practitioner is called an endorser, identified as anyone who enjoys public recognition for consumers' products by featuring in an advertisement [30], [31]. Endorsement is used to be done by celebrities in advertisements such as on TV, on the radio, or in a newspaper. However, the popularity of social media gives a new way to endorse a product which is through social media interaction [6] and the use of influencers [10], [32].

This change means a new measurement is needed to determine the efficiency of the endorsement. This comes in the form of engagement level. Engagement is the interaction between followers and their influencers through the post that the influencers upload [33]. Engagement can come in the forms of likes, views, or comments. However, engagement level cannot be measured solely by amounts of likes, views, or comments [34], [35].

E. Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) is a method used in statistics to evaluate how important a word is in a certain document or how crucial a category is in a set of files [36] and works well with text mining methods [37]. The term frequency (TF) and inverse document frequency (IDF) weighting models are combined in the TF-IDF weighting model occurs in a number of texts that are regarded as generic words deemed unimportant. This method works by checking the frequency of a word appearing in a document and how often it appears on other documents. It is divided into words' frequency and inverse texts' frequency. Words' frequency is the frequency of words' appearance in an article, while inverse texts' frequency notates how generally important a word is. The formula of TF-IDF can be seen below:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{4}$$

The goal of this research is to group influencers based on their contents and endorsements. To achieve it, we propose research methodology as can be seen on Figure 1. The approach taken in this research is mainly divided into these steps are Selecting platform for data collection, Data acquisition, data pre-processing, Feature Selection, Clustering and Evaluation.

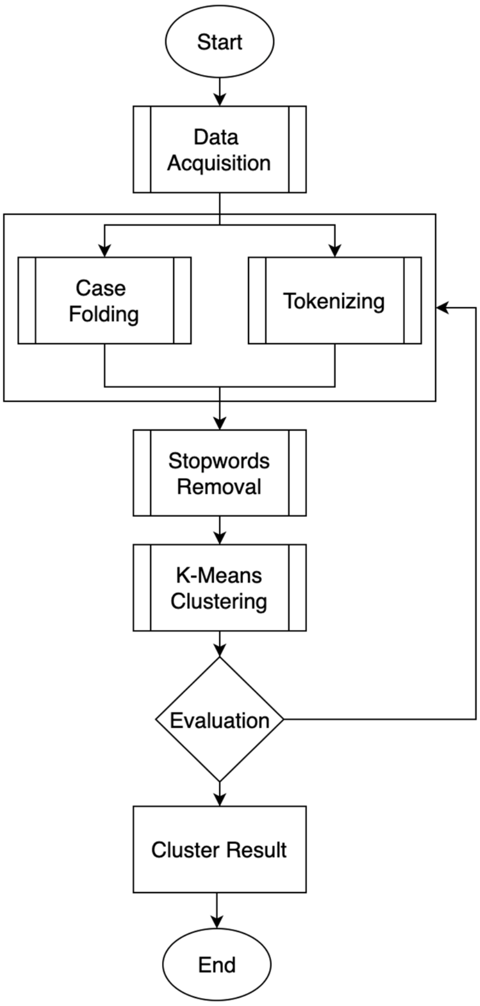

Fig. 1 Research Methodology

A. Selecting Platform for Data Collection

The platform which is chosen for the source of data collection is Instagram. We choose Instagram as this platform is one of the most widely used by people, and Instagram is also known for marketing and promotion through endorsement.

B. Data Acquisition

We employ a tool that can collect large amounts of data and turn it into information, and the data is collected using CSV (Comma Separated Values) format. We use hashtags and contents of influencers' posts as parameters for clustering.

C. Data Pre-Processing

After having finished data collection, the next step is data pre-processing. In this process, several steps need to be done. Common steps which are used are case folding, tokenizing, stemming, and tagging [10]. This process includes some activities, such as removing unnecessary characters or turning all characters into lower case [14]. Preprocessing improves the data, removes noise, and identifies which section of the image will be used in the following stage [36].

Data pre-processing is necessary to achieve accurate results and set data up to be ready for machine learning and data training [38]. For pre-processing the data, there are several things to do to prepare data. Based on previous research [14], generally, the methods are:

- Cleansing: to clean data from unnecessary words or characters to reduce noise in the data, such as punctuation marks or numbers
- Case folding: to change words in the reviews into a singular shape. For example, changing all words into a lower case or upper case
- Tokenization: splitting sentences into words, phrases, or symbols separated by spaces or special characters.
- Filtering: to put out stop words, which are usually general words with high appearance counts, irrelevant, or unimportant attributes to the text. These words are grouped into a list called a stop words list.

D. Feature Selection

After getting the data pre-processed, then we made the feature selection. This step is done to reduce data dimension to improve the accuracy and quality of the result [5]. A process that is commonly applied is stop word removal. Stop words are words that have high frequencies of appearance in text or document yet have no significant or relevant meaning throughout the document. By removing these words, we prevent these words from being included in analysis and evaluation, which can improve the results as these words can affect the result negatively.

E. Clustering

After making feature selection, we move into the main process of clustering the data. There are several research in the past using clustering in text mining, such as [13], [14], [16], [23]. We use K-means algorithm, which is popular, easy to use, and has good performance on the result. Clustering will be done to the data to group similar influencers to each different group based on their posts. The result will be a system to cluster the data based on their characteristics.

F. Evaluation

When clustering finishes, we evaluate the result by calculating the value of precision, recall, and purity of the cluster in the result. Next, we will check the validity of the cluster in the result. This is crucial to find the optimal divisions for the data.

III. Result and Discussion

A. Data Collection

We chose Instagram for data collection because Instagram is a popular social networking tool that is both informational and fun. Instagram is one of the most popular social media platforms, allowing anyone to create information with a large audience and share it on other platforms. It is connected to the state of the art of text mining, which is widely defined as the process of seeking and gathering insightful information from data by investigating and detecting a pattern.

Instagram offers a wealth of data to investigate because it allows people to express themselves instantly in the text, image, or video context. As a result, Instagram is one of the most popular marketing platforms because it allows individuals to interact based on their shared interests. This ensures that every Instagram campaign has a positive impact because it is simple to classify products so that everyone may see ads that are relevant to their interests.

We aim to use only text data from posts' captions and hashtags inside the posts. We collected 696 posts from one of the popular celebgram in Indonesia, which would be pre-processed. Table I displays some of the Instagram data bundled into CSV format to make it easier to read and process. There is a caption column that displays all of the text in each influencer's material and a date column that provides extra information about when the content was released.

TABLE I
Sample CSV data collection from Instagram

No	Captions	Date
1	*Kalo ootd, mood juga jadi naik, walau cuma di ujung komplek*	2021-06-06 11:12:36
2	*Humaira Instant Hijab @hanum_id*	2021-05-31 06:33:22
3	*InsyaAllah, sehat-sehat di dalam ya bayik, semuanya kita perjuangin sama-sama sampai kamu dan mama nyaman hingga waktunya lahir tiba... aamiin #31weekspregnant*	2021-05-27 08:02:38
4	*Haiii Moms! Mau ngeracunin produk anak lagi nih, aku mulai concern babycare yang HALAL karena lebih tenang aja di hati. kali ini pilihanku dari @momami.id karena aku pakai tisu pembersih giginya buat Freya. Rangkaian baby care dari momani ini udah lengkap banget loh. Bahan yang digunakan juga alami tanpa paraben dan SLS, dermatoligically tested dan hypoallergenic jadi gak perlu khawatir deh. Yuk cobain juga produk #momami, langsung cek ignya ya! #madeformi #momslittlehelper*	2021-05-23 11:40:45
5	*Berasa kayak princess karna dressnya @lnwfashion yang cakep bgt ... ini full brukat yang pasti ada furingnya, trus di list pitanya ada bukaan resleting untuk menyusuiii yang hampir gak keliatan, kece bgt kan*	2021-05-05 10:42:53

B. Pre-Processing Result

In this step, we aim to cleanse the data before doing the clustering to optimize the clustering result. In this optimization step, we execute a case folding and tokenization to remove emoticons, punctuation marks, mentions, hashtags, digits, and lowercase words. Table II displays the results of pre-processing using the sample data from Table I. As a result, the sample data will be easier to process in the following phase.

TABLE II
AFTER CASE FOLDING PROCESS

No	Captions	Date
1	*kalo ootd mood juga jadi naik walau cuma di ujung komplek*	2021-06-06 11:12:36
2	*humaira instant hijab hanumid*	2021-05-31 06:33:22
3	*insyaallah sehatsehat di dalam ya bayik semuanya kita perjuangin samasama sampai kamu dan mama nyaman hingga waktunya lahir tiba aamiin weekspregnant*	2021-05-27 08:02:38
4	*haiii moms mau ngeracunin produk anak lagi nih aku mulai concern babycare yang halal karena lebih tenang aja di hati kali ini pilihanku dari momamiid karena aku pakai tisu pembersih giginya buat freya rangkaian baby care dari momani ini udah lengkap banget loh bahan yang digunakan juga alami tanpa paraben dan sls dermatoligically tested dan hypoallergenic jadi gak perlu khawatir deh yuk cobain juga produk momami langsung cek ignya ya madeformi momslittlehelper*	2021-05-23 11:40:45
5	*berasa kayak princess karna dressnya lnwfashion yang cakep bgt ini full brukat yang pasti ada furingnya trus di list pitanya ada bukaan resleting untuk menyusuiii yang hampir gak keliatan kece bgt kan*	2021-05-05 10:42:53

Figure 2 shows the tokenization process results, including each word's frequency. However, it is clear from these results that many words do not have the correct meaning, which will impact the accuracy. As a result, a Stop words process is required to assist in making the correct word more meaningful. Second, we remove stop words from the data. Stop words have less significant meaning or importance [39]. However, stop words can also be crucial based on the context. Stop words are usually detected by their frequency of appearances across the data.

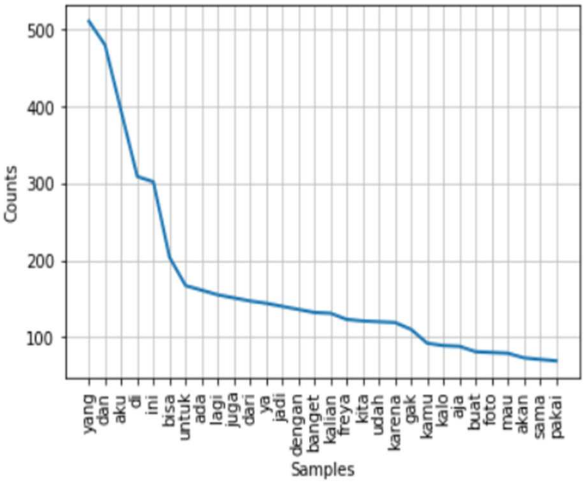

Fig. 2 Words in data and their frequencies

Then we removed the stop words. We specifically used the Indonesian stop words dictionary. Figure 3 depicts a better result since unnecessary words have been removed, making the graph appear more significant and displaying the frequency of each term, which has a greater impact on the results.

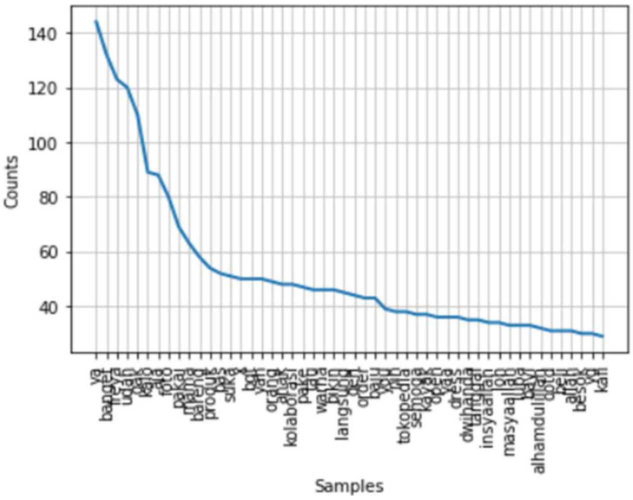

Fig. 3 Stop words and their frequencies

Then we did feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) to evaluate how important each word is. It works by checking the frequency of words that appeared in the data [36]. The result showed that there are 5123 features. Then, clustering was applied to the result. First, we tried to determine the optimal value of k by using the elbow method. The range for the value is between two and ten.

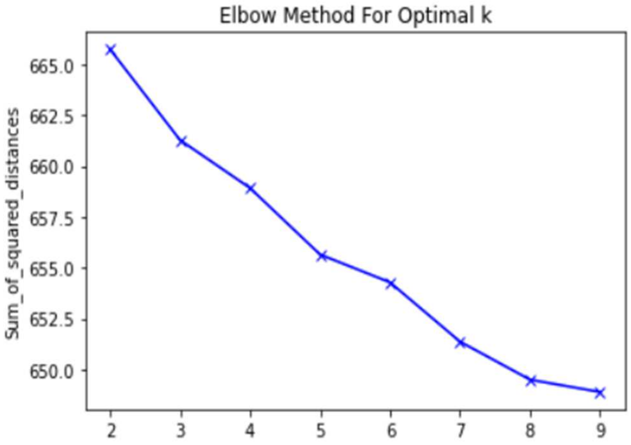

Fig. 4 Elbow method used to find optimal k value

C. Clustering Result

According to the graph in Figure 4, the optimal value for k was k = 5, which was obtained using the elbow method to determine the number of clusters used. Then, we divided the data into k cluster. We visualized it using Principal Component Analysis (PCA), and T-distributed Stochastic Neighbor Embedding (TSNE) scatter plot.

PCA was used to capture the global structure of the data better while TSNE was used to capture relations between neighboring clusters better. Three out of five clusters in total showed high quality based on Figure 5. It is also done to identify a category of terms that frequently appear in each of the clusters given in Table III. The product category and user persona are determined by looking back at the content uploaded to gain more accurate information about which product categories match the product category and user persona.

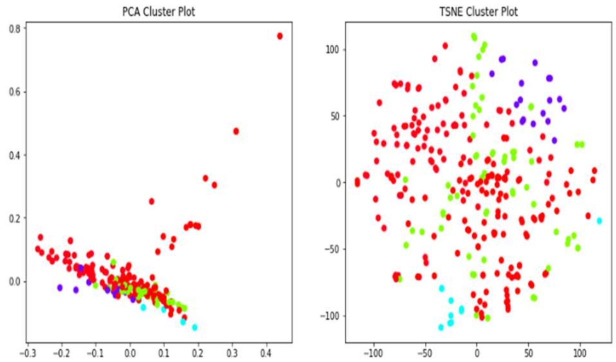

Fig. 5 PCA dan TSNE scatter plot of clustering

Because there is much content in Cluster 1 that references numerous brands of sleeping equipment such as pillows, bolsters, blankets, and other items, the category of bedding is found. Additionally, there are nightgowns for both mothers and children. This cluster occurred because this influencer is newly married and has children. Therefore, the time is appropriate for the owner of these bedding goods to employ his skills to promote their product.

Because the content in apparel, glassware, marketing from tourist sites such as villas, hotels, playgrounds, and testimonials from a public areas is so diverse, there are general categories in Cluster 2. This cluster emerged because these influencers are generally receptive to product promotion and conduct self-help promotions in public locations, giving the audience the impression that these influencers are honest in delivering promotions to the audience while also giving the brand owner trust.

The category of baby and child equipment is found in Clusters 3 and 4, marked by content showing babies and children using a baby and child equipment product. In addition to products, there are also promotions from service providers that serve baby care, such as *massage* and *spa*, while for children, there is more promotional content related to playgrounds and toys. Because influencers are caring for babies and children, this cluster shares the same features as Cluster 1, indicating that there is enough momentum to promote baby and child products.

Cluster 5 has the category of Muslim equipment, identified by material depicting this influencer wearing a headscarf and other Muslim clothing. Furthermore, influencers frequently promote their families, such as their children and husbands, who use Muslim products such as matching clothing and prayer gear. Because this influencer is a Muslim, his character is perfect for Muslim equipment products.

Based on the findings of the analysis of each of these clusters, it was discovered that these influencers have a strong family spirit, as evidenced by the content in each cluster that always depicts a family figure. So, from all of the clusters that have been formed, this can be used as a reference that these influencers are generally suitable to be used as alternative promotion using an endorsement approach for products and services related to families such as clothing, family-to-family travel, etc. However, based on the previous analysis, it is more appropriate for products and services related to babies and children because the momentum is right. These influencers are caring for their babies and children. However, this will only be temporary because it will inevitably grow, and promotional content will no longer be relevant.

Furthermore, these influencers provide honest testimonials that are marked using their promotional products daily, indicating that the influencer will provide honest testimonials, both good and bad. This will positively impact both the audience and the brand owners, ensuring that the audience is always interested in promotional products and that the brand owners maintain the quality of their products and services.

TABLE III
THE MOST FREQUENTLY OCCURRING WORDS AND ENDORSEMENT CONTENT CATEGORIES IN EACH CLUSTER

Cluster	The most frequent word	Category
1	*you, belum, yang, ya, kalo, love, aku, udah, he, tidur*	Bedding
2	*yang, dan, aku, banget, di, ini, kalian, untuk, jadi, bisa*	General
3	*foto, di, gais, paste, bareng, bayibayi, mahikakids, koleksi, ini, biasa*	Baby/Kids Equipment
4	*anak, dan, yang, ini, lagi, untuk, ada, dengan, bisa, kita*	Baby/Kids equipment
5	*ootd, tas, yok, ini, masyaallahtabarakallah, niiii, lah, hehe, dulu*	Muslim equipment

D. Evaluation

To ensure the result's validity, Silhouette Coefficient was applied to the model. Silhouette Coefficient will produce a number between -1 to 1 where -1 means incorrect clustering and 1 means highly dense clustering, which is high similarity between cluster's members and low similarity between neighboring cluster's members [40]. Calculation showed that the coefficient was 7.277×10^{-3} which meant a good cluster.

IV. CONCLUSION

This research aims to create a recommendation for selecting influencers suitable for products or brands based on their Instagram posts. We used clustering for categorized influencers' posts to discover what products they mainly endorse or what persona and activities are on their feeds. This hopefully will be useful for product or brand owners in selecting which influencers will suit their products or brands.

The result of the research showed that the model could cluster the data, albeit rather low. It was noticed from the Silhouette Coefficient, which was applied to the result. Some obstacles we faced while doing this research caused those problems. First, there were not enough data that we could provide to the model, and the lack of means or methods caused this to collect the data. Second, the data might not contain important keywords or hashtags that were distinct in other to distinguish the data into explicit clusters. Finally, there was a lack of data collected, which was implied in the clustering result.

For future work, we believe some improvements can be made for a better result and impact. First, collecting more data is required to get better clustering result out of the K-means algorithm. Second, different algorithms or approaches to clustering, such as the hybrid approach, can be adapted to produce better results. Last, more methods to validate clustering results should be applied for more robust validity.

ACKNOWLEDGMENT

The authors are grateful to the Faculty of Computer Science, the University of Indonesia, for all the knowledge, support in the process, and financing of the International Publication.

REFERENCES

[1] APJII, “Indonesian internet survey report 2019-2020,” *Asos. Penyelenggara Jasa Internet Indones.*, vol. 2020, pp. 1–146, 2020, [Online]. Available: https://apjii.or.id/survei.

[2] F. Mirzaalian and E. Halpenny, "Exploring destination loyalty: Application of social media analytics in a nature-based tourism setting," *J. Destin. Mark. Manag.*, vol. 20, p. 100598, 2021, doi: https://doi.org/10.1016/j.jdmm.2021.100598.

[3] A. Arora, S. Bansal, C. Kandpal, R. Aswani, and Y. Dwivedi, "Measuring social media influencer index- insights from facebook, Twitter and Instagram," *J. Retail. Consum. Serv.*, vol. 49, pp. 86–101, 2019, doi: https://doi.org/10.1016/j.jretconser.2019.03.012.

[4] 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, p. 399, 2019, doi: 10.22146/ijccs.50574.

[5] M. I. Akrianto, A. D. Hartanto, and A. Priadana, "The Best Parameters to Select Instagram Account for Endorsement using Web Scraping," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2019, pp. 40–45, doi: 10.1109/ICITISEE48480.2019.9004038.

[6] A. Arifianto *et al.*, "Endorsement Recommendation Using Instagram Follower Profiling," in *2018 6th International Conference on Information and Communication Technology (ICoICT)*, 2018, pp. 470–475, doi: 10.1109/ICoICT.2018.8528724.

[7] C.-W. (Chloe) Ki, L. M. Cuevas, S. M. Chong, and H. Lim, "Influencer marketing: Social media influencers as human brands attaching to followers and yielding positive marketing results by fulfilling needs," *J. Retail. Consum. Serv.*, vol. 55, p. 102133, 2020, doi: https://doi.org/10.1016/j.jretconser.2020.102133.

[8] D. Jiménez-Castillo and R. Sánchez-Fernández, "The role of digital influencers in brand recommendation: Examining their impact on engagement, expected value and purchase intention," *Int. J. Inf. Manage.*, vol. 49, pp. 366–376, 2019, doi: https://doi.org/10.1016/j.ijinfomgt.2019.07.009.

[9] C. Lou and S. Yuan, "Influencer Marketing: How Message Value and Credibility Affect Consumer Trust of Branded Content on Social Media," *J. Interact. Advert.*, vol. 19, no. 1, pp. 58–73, 2019, doi: 10.1080/15252019.2018.1533501.

[10] W. Tafesse and B. P. Wood, "Followers' engagement with instagram influencers: The role of influencers' content and engagement strategy," *J. Retail. Consum. Serv.*, vol. 58, p. 102303, 2021, doi: https://doi.org/10.1016/j.jretconser.2020.102303.

[11] R. Y. Kim, "The Value of Followers on Social Media," *IEEE Eng. Manag. Rev.*, vol. 48, no. 2, pp. 173–183, 2020, doi: 10.1109/EMR.2020.2979973.

[12] K. Sokolova and H. Kefi, "Instagram and YouTube bloggers promote it, why should I buy? How credibility and parasocial interaction influence purchase intentions," *J. Retail. Consum. Serv.*, vol. 53, p. 101742, 2020, doi: https://doi.org/10.1016/j.jretconser.2019.01.011.

[13] N. Yudiarta, M. Sudarma, and W. Ariastina, "Application of Clustering Text Mining Method for Grouping News on Unstructured Textual D ata," *Maj. Ilm. Teknol. Elektro*, vol. 17, p. 339, 2018, doi: 10.24843/MITE.2018.v17i03.P06.

[14] D. Indraloka and B. Santosa, "Application of text mining to cluster Shopee Indonesia's tweet data," *J. Sains dan Seni ITS*, vol. 6, 2017, doi: 10.12962/j23373520.v6i2.24419.

[15] S. P. Kristanto, J. A. Prasetyo, and E. Pramana, "Naive Bayes Classifier on Twitter Sentiment Analysis BPJS of HEALTH," in *2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)*, 2019, pp. 24–28, doi: 10.1109/IC2IE47452.2019.8940900.

[16] B. Ray, A. Garain, and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," *Appl. Soft Comput.*, vol. 98, p. 106935, 2021, doi: https://doi.org/10.1016/j.asoc.2020.106935.

[17] A. Ahani, M. Nilashi, O. Ibrahim, L. Sanzogni, and S. Weaven, "Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews," *Int. J. Hosp. Manag.*, vol. 80, pp. 52–77, 2019, doi: https://doi.org/10.1016/j.ijhm.2019.01.003.

[18] A. Fronzetti Colladon, B. Guardabascio, and R. Innarella, "Using social network and semantic analysis to analyze online travel forums and forecast tourism demand," *Decis. Support Syst.*, vol. 123, p. 113075, 2019, doi: https://doi.org/10.1016/j.dss.2019.113075.

[19] V. Arefieva, R. Egger, and J. Yu, "A machine learning approach to cluster destination image on Instagram," *Tour. Manag.*, vol. 85, p. 104318, 2021, doi: https://doi.org/10.1016/j.tourman.2021.104318.

[20] A. R. Pathak, M. Pandey, and S. Rautaray, "Topic-level sentiment analysis of social media data using deep learning," *Appl. Soft Comput.*, vol. 108, p. 107440, 2021, doi: https://doi.org/10.1016/j.asoc.2021.107440.

[21] A. Onan, "Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019, doi: 10.1109/ACCESS.2019.2945911.

[22] Y. D. Darmi and A. Setiawan, "Application of K-Means Clustering Method in Product Sales Grouping," *J. Media Infotama*, vol. 12, no. 2, pp. 148–157, 2017, doi: 10.37676/jmi.v12i2.418.

[23] P. W. Cahyo and M. Habibi, "Clustering followers of influencers accounts based on likes and comments on Instagram Platform," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 14, no. 2, p. 199, 2020, doi: 10.22146/ijccs.53028.

[24] B. Mulyawan, M. V. Christanti, and R. Wenas, "Recommendation Product Based on Customer Categorization with K-Means Clustering Method," *{IOP} Conf. Ser. Mater. Sci. Eng.*, vol. 508, p. 12123, May 2019, doi: 10.1088/1757-899x/508/1/012123.

[25] A. A. hussian Hassan, W. M. Shah, M. F. I. Othman, and H. A. H. Hassan, "Evaluate the performance of K-Means and the fuzzy C-Means algorithms to formation balanced clusters in wireless sensor networks," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 1515–1523, 2020, doi: 10.11591/ijece.v10i2.pp1515-1523.

[26] B. Bashari and E. Fazl-Ersi, "Influential post identification on Instagram through caption and hashtag analysis," *Meas. Control*, vol. 53, no. 3–4, pp. 409–415, 2020, doi: 10.1177/0020294019877489.

[27] S. Banerjee, A. Choudhary, and S. Pal, "Empirical evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means clustering algorithms," in *2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, 2015, pp. 168–172, doi: 10.1109/WIECON-ECE.2015.7443889.

[28] X. Yang, S. Kim, and Y. Sun, "How Do Influencers Mention Brands in Social Media? Sponsorship Prediction of Instagram Posts," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 101–104, doi: 10.1145/3341161.3342925.

[29] J. Han *et al.*, "FITNet: Identifying Fashion Influencers on Twitter," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, Apr. 2021, doi: 10.1145/3449227.

[30] A. P. Schouten, L. Janssen, and M. Verspaget, "Celebrity vs. Influencer endorsements in advertising: the role of identification, credibility, and Product-Endorser fit," *Int. J. Advert.*, vol. 39, no. 2, pp. 258–281, 2020, doi: 10.1080/02650487.2019.1634898.

[31] K. Osei-Frimpong, G. Donkor, and N. Owusu-Frimpong, "The Impact of Celebrity Endorsement on Consumer Purchase Intention: An Emerging Market Perspective," *J. Mark. Theory Pract.*, vol. 27, no. 1, pp. 103–121, 2019, doi: 10.1080/10696679.2018.1534070.

[32] S. Kim, J.-Y. Jiang, M. Nakada, J. Han, and W. Wang, "Multimodal Post Attentive Profiling for Influencer Marketing," in *Proceedings of The Web Conference 2020*, New York, NY, USA: Association for Computing Machinery, 2020, pp. 2878–2884.

[33] R. L. H. Yew, S. B. Suhaidi, P. Seewoochurn, and V. K. Sevamalai, "Social Network Influencers' Engagement Rate Algorithm Using Instagram Data," in *2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA)*, 2018, pp. 1–8, doi: 10.1109/ICACCAF.2018.8776755.

[34] C. Hughes, V. Swaminathan, and G. Brooks, "Driving Brand Engagement Through Online Social Influencers: An Empirical Investigation of Sponsored Blogging Campaigns," *J. Mark.*, vol. 83, no. 5, pp. 78–96, 2019, doi: 10.1177/0022242919854374.

[35] W. N. A. Rahman, D. S. Mutum, and E. M. Ghazali, "Consumer Engagement With Visual Content on Instagram," *Int. J. E-Services Mob. Appl.*, vol. 14, no. 1, pp. 1–21, 2022, doi: 10.4018/ijesma.295960.

[36] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 466–472, doi: 10.1109/ICACCS48705.2020.9074166.

[37] H. Aljuaid, R. Iftikhar, S. Ahmad, M. Asif, and M. Tanvir Afzal, "Important citation identification using sentiment analysis of in-text citations," *Telemat. Informatics*, vol. 56, p. 101492, 2021, doi: https://doi.org/10.1016/j.tele.2020.101492.

[38] K. Deb, S. Banerjee, R. P. Chatterjee, A. Das, and R. Bag, "Educational website ranking using fuzzy logic and k-means clustering based hybrid method," *Ing. des Syst. d'Information*, vol. 24, no. 5, pp. 497–506, 2019, doi: 10.18280/isi.240506.

[39] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature weighting method and its analysis using unstructured dataset," *CEUR Workshop Proc.*, vol. 2870, pp. 98–107, 2021.

[40] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987, doi: https://doi.org/10.1016/0377-0427(87)90125-7.