Jurnal Teknik Industri, Vol. 23, No. 2, December 2021
ISSN 1411-2485 print / ISSN 2087-7439 online

DOI: 10.9744/jti.23.2.111-120

Utilizing Elbow Method for Text Clustering Optimization in Analyzing Social Media Marketing Content of Indonesian e-Commerce

Aisyah Larasati[1*], Raretha Maren[1], Retno Wulandari[2]

Abstract: The massive increases in textual data from Twitter and text analytics simultaneously have driven organizations to obtain hidden insights to implement the proper marketing strategies for businesses. The vast information generated by Twitter enables most e-commerce businesses to utilize Twitter to implement social media marketing. One of those e-commerce businesses is Blibli Indonesia. Intense business competition has led them to perform marketing strategies to understand consumer tendencies. Focusing the marketing strategies on consumer preferences enables the increase of consumer interest in Blibli, which is in line with enhancing the opportunity to reach new consumers. This research aims to discover Twitter content based on k-means results to cluster the tweets of @bliblidotcom. The best cluster is determined with the elbow method by selecting the deepest curvature, three clusters. The result suggests that Twitter users like Park Seo Jun's content. Hence, Blibli can focus on that content as its business marketing strategy on the Twitter platform.

Keywords: E-commerce, Twitter, marketing, text mining, K-means, elbow.

Introduction

Simultaneously, the evolution of textual data and text analytics has encouraged most organizations to utilize them as their business concern [1]). The significant increase in textual data is driven by the emergence of social media platforms such as Twitter and Facebook [2]. The available textual data can be used as a data source to obtain new insight into precise social media marketing for businesses [3]). Twitter provides most of its information in textual form [4]). Associated with the phenomenon of e-commerce business in Indonesia, one e-commerce that utilizes Twitter platforms for social media marketing is Blibli Indonesia.

An intense business competition requires Blibli Indonesia to implement the proper marketing strategies. Communication through tweet uploads need to be carried out properly, according to consumer preferences, in order to acquire positive responses [5]. In practice, Blibli tends to upload random tweets that have not met consumer trends and interests. Moreover, the tweets of Blibli infrequently have major retweets, which indicates that the existing Twitter contents are yet to be adequately effective. In consequence, the existence of Blibli has been less popular in public, which is in accordance to have fewer possibilities for consumers who are interested in making transactions through the Blibli application. The retweet and like features on Twitter are metrics for evaluating content effectiveness, which also can be accepted as metrics for measuring content popularity and consumer preference [6]. Thus, the tweet content that has not had major retweets and likes indicates less desirable content.

The integration of text mining with the clustering method is the appropriate alternative to discover the preferable Twitter content that fits public preference. This is based on the ability of text mining to process textual data into beneficial information to meet analytical needs aligned with the goals of company business through several approaches [7]. The approach can be a clustering method to group the tweet contents using the k-means algorithm due to its convenience and relatively low time complexity of the implementation in handling large amounts of data [8]. In addition, the k-means algorithm is adequately optimal and highly utilized in text mining. The k-means algorithm proceeds to classify a set of data corresponding to the number of clusters determined by calculating the closest distance to the randomized centroids [9]. Each separate cluster has disparity characteristics. Thus, data partitioning is performed based on the characteristics of each cluster. Data with identical characteristics will be grouped into one cluster and vice versa [10]. However, the k-means algorithm cannot work correctly if the cluster parameter is selected subjectively since the algorithm is included in partitional clustering. Hence, it is

[1]Faculty of Industial Technology, Industrial Engineering Department, Universitas Negeri Malang, Jl. Semarang 5 Malang 65145, Indonesia.
Email: aisyah.larasati.ft@um.ac.id; raretha.maren.1705166@students.um.ac.id
[2] Faculty of Industial Technology, Mechanical Engineering Department, Universitas Negeri Malang, Jl. Semarang 5 Malang 65145, Indonesia. Email: retno.wulandari.ft@um.ac.id

*Corresponding author

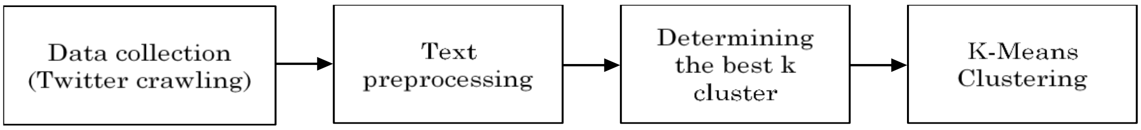
Data collection (Twitter crawling)
Text preprocessing
Determining the best k cluster
K-Means Clustering

Figure 1. Research flow diagram

obligatory to set the number of clusters at the beginning of the process before executing the algorithm [11]. Therefore, the elbow method can provide the optimal cluster as well as the performance of the clustering results [12].

This research aims to determine the appropriate social media marketing strategies for Blibli, according to tweet content that meets the general preferences (Twitter users). The results of this study are expected to provide information for Blibli Indonesia to focus on the preferred content of consumers as their marketing strategies on Twitter. Many more Twitter users are expected to attract Blibli applications by focusing on desirable content. Thus, the higher probability of receiving new consumers, the more consumers are willing to make a shopping transaction through Blibli Indonesia.

Methods

The research stages begin with the data collection stage using Twitter API to acquire keys and tokens, connecting Twitter to R Studio 1.4.1 environment. Next, a set of modules are involved in the data retrieval process, including *twitteR, RCurl, ROAuth,* and *xlsx*. The data is categorized as primary data, obtained via the Twitter crawling process on the Twitter of Blibli Indonesia (@bliblidotcom). This study contains 491 tweets uploaded from 28 May 2020 to 4 March 2021, along with two variables, namely text data variable, which is a set of tweets from @bliblidotcom, and retweet count variable, which is a total retweet of each tweet.

The next stage is cleaning the data via text preprocessing, which transforms textual data into a structured form to minimize noises and avoid errors during the clustering process [13]). After cleaning the data, the next step is to determine the best cluster number using the elbow method to discover the optimal cluster formulation. Finally, the value obtained from the elbow method is used as the parameter for the cluster number while applying the k-means algorithm for text clustering to find out the tweet contents of Blibli Indonesia. The text preprocessing until text clustering is performed by utilizing Anaconda environment with Python 3.0. Python modules used in this research are *pandas, emoji, regex, matplotlib, scikit-learn,* and *word cloud*. Figure 1 provides the flow diagram of the research methodology.

The differences between this study and other similar studies are the use of the Z method for stopwords removal in text preprocessing that infrequently explored in the field of text mining, and the elbow method, which is conducted by focusing on determining the k parameter based on the sum square error delta for clarity in selecting the best cluster point.

Text Preprocessing

Text preprocessing is a step for transforming words into the standard form of a language to reduce noises in the textual data, which can inflict quality reduction on further analysis [14]. Generally, text preprocessing consists of seven phases: character removal, case folding, stemming, removing punctuation, stopwords removal, tokenizing, and configuring a document matrix to represent word items in numeric form [15]. The phases and the sequence of text preprocessing techniques can be adjusted to the analysis requirements by enhancing other techniques but not reducing the general phases that need to be executed. The text preprocessing step involves many modules in Python.

The text preprocessing step is commenced by character removal, which eliminates emoticons, numbers, usernames, and URLs found in tweets by utilizing the *emoji* and *regex* modules. The second step is case folding for converting the upper-case into lower-case letters using *nltk* module since all text in tweets does not apply the letters form consistently. The third step is to transform the words into basic forms by removing affixes according to the confix stripping stemmer concept in the *sastrawi* module. The fourth step is removing a set of punctuations, as punctuation is only a symbol for reading intonation that indicates the text structure. Thus, it does not have a special meaning to be processed in text clustering. This step is carried out using the *regex* module and executed after stemming since it adjusts the systematics of repeated words in the Indonesian language rules inserting a hyphen (-) between the two words. Removing punctuation is performed at the end to avoid stemming errors in repeated words. The next step is stopwords removal which eliminates unnecessary and meaningless words in tweets according to the concept of Z-methods. The Z-methods removes most and least frequent words by ranking all existing terms in the dataset based on their occurrences and selecting the most and least words as the stoplist for the stopwords removal process, stored in txt format. Most frequent and singleton words in the Z-methods are evaluated as noises that indicate less informative terms [16]. Hence, those words are needed to be eliminated. The Z-method is a word-filtering procedure with a simple calculation and less calculation time to conduct the stopwords selection [17]. Moreover, the Z-method performance does not depend on the type of language and delivers promising results in terms of precision and recall [18].

In this study, the experiments to evaluate the clustering performance are conducted with and without the z-method in the filtering stopwords process. The best result is assessed by Davies Bouldin Index (DBI) value. The entire filter stopwords step is performed by applying various functions existing in the *nltk* and *sastrawi* module of Python.

After that, the next step is tokenizing, which parses the sentences into a word for facilitating the calculation of word occurrence and term weighting. The tokenizing step is performed through the *nltk* module. The following step of text preprocessing is to calculate the importance level of words according to the TF-IDF idea by utilizing the *nltk* module. The TF-IDF stands for Term Frequency-Invers Document Frequency, which calculates the term value derived from the word occurrence and then multiplied by document frequency with a logarithmic scale to prevent bias caused by high-frequency words compared to others [19]. The formulation of TF-IDF calculation is shown in equation (1).

$$W_{i,j} = tf_{i,j} \times idf_j = tf_{i,j} \times \log\left(\frac{N}{df_j}\right) \tag{1}$$

where $W_{i,j}$ is the weight of term j in document i, $tf_{i,j}$ is the term frequency of j in document i, N is a total document existed in the data, and df_j is the document frequency containing the term j.

After undertaking the term-weighting with TF-IDF, the final step is to transform the value into an array format, following the rules in *scikit-learn* Python, which requires array format to operate the elbow method function. Moreover, the *scikit-learn* only accepts an array format to be processed. Hence, the matrix documents containing the value of TF-IDF of all terms should be converted into an array.

Elbow Method

The process of determining the best cluster utilizes the *scikit-learn* module in Python by using the elbow method. The elbow method is a procedure to optimize the k-means algorithm by evaluating the optimal clusters through the estimation of sum square error (SSE) in each defined range of the clusters [20]. The final figuration of the elbow method is a graph of the best cluster consistency by plotting the SSE value. The most extreme decrease and the most elbow-shaped is considered as the most optimal cluster number. In other words, the determination of optimal cluster is obtained by selecting the largest error decrease of the existing cluster [21]. In this study, the initialization of the cluster range applies the number of 2 to 20 for the k parameter and then selects the existing cluster with the most extreme error. The consideration regarding the cluster range is based on experiments. When reducing the cluster range, produces many misclustering cases. Therefore, the number of 2 until 20 is evaluated as the best fit for the case of this study. The optimal cluster selected from the elbow method is used as the input parameter of k-cluster for the k-means execution. For clarity, equation (2) is the SSE functionality to determine the sum square error value for each defined cluster.

$$SSE = \sum_{i=1}^{k} \sum_{j \in C_i} \left\| X_j - C_i \right\|^2 \tag{2}$$

where k is the defined cluster numbers, X_j is data j in cluster i, and C_i is randomly initialize the centroids.

The K-Means Algorithm for Tweet Clustering

The stage of the k-means algorithm is commenced by determining the optimal cluster using the elbow method as described in the aforementioned subsection. The procedure of the k-means is to divide a set of data into separate classes based on the calculation of the minimal distance between centroids and data points [22]. The k-means algorithm is included in partitional clustering that requires an appropriate intuition about the cluster interval. Thus, according to the data structure, the clustering results can be well-separated [23]. The determination of the optimal cluster is executed 20 times with the range of k is 2 to 20. The best value is obtained from the most extreme error through the elbow method, which will be used as the actual k-means input to cluster the tweets of Blibli Indonesia Twitter.

Table 1. Parameters of the k-means algorithm for text clustering of Blibli Indonesia's tweets

Parameters	Value	Description
k_cluster	2-20	The range of cluster numbers
max_iter	k-means ++	Determination of initial centroid to find the convergence point rapidly
n_init	100	Maximum iterations allowed for a single run k-means execution
random_state	10	Number of k-means executions using disparate centroids
metric	42	Random number generation process for the initial definition of the centroid

The k-means process is performed by using the *scikit-learn* module in Python. The entire k-means parameters set in this study as defined in Table 1.

This study conducted experiments for assessing the best performance of filter stopword with and without the Z-method. Thus, specifying which experiment gives the best performance is determined by selecting the smaller value of Davies Bouldin Index (DBI) as explained in the text preprocessing section. After investigating the optimal cluster number, further analysis is employed to discover trends in content tweets that meet the Twitter user preferences by tallying the average retweet in each content cluster formed by the k-means. The magnitude of the calculation is an embodiment of the preference level of Twitter users to the tweet content of Blibli. Clusters with high tally indicate the most effective content which can attract the public. Hence, the content with the highest retweet count can be used as a fundamental decision for Blibli Indonesia to apply business strategies through social media marketing on Twitter by focusing on the preferred contents of Twitter users. The tally of retweet count is performed using the *pandas* and *matplotlib* module in Python.

Davies Bouldin Index (DBI)

Davies Bouldin Index is one of the internal cluster validity indices to identify the clustering result by considering the ratio between minimum intra-cluster distance and maximum separation of cluster centers. The optimal compactness of the clustering result is evaluated by selecting the smaller value of DBI. The smaller value of DBI, represents the better clustering quality [24]. In this study, DBI idea is conducted by using scikit-learn module to identify the filter stopwords process and provide the performance of with and without the z-method in text clustering problem. The formulation of DBI is defined in equation (3).

$$DBI = \frac{1}{K}\sum_{i=1}^{k} max_{i\neq j}\left(R_{i,j}\right) \qquad (3)$$

where $R_{i,j}$ is the ratio between clusters i and j to the distance of related cluster centroids, while K is the optimal cluster number obtained from the elbow method in the defined experiments.

Results and Discussions

The Best Cluster Number from The Elbow Method

The idea of selecting the best cluster number from the elbow method is commenced by the visualization, called the SSE (sum of square error) graph in each defined cluster. The point with the most elbow shape indicates the most extreme error, which evaluates as the best cluster number to perform the text clustering process. The SSE calculation process is executed using many Python modules functions, called the *scikit-learn* for calculating the SSE value and *matplotlib* for visualizing the elbow point. Figure 2 shows the SSE calculations using the elbow method in each defined k for the clustering process with and without the Z-method in the filter stopwords step.

Based on Figure 2, It is yet to be inferred which one is the best cluster for the k-means algorithm execution for both experiments since the resulting error value has only a few differences and decreases monotonically. Moreover, the elbow method is a semi-intuitive way to measure the best k point for clustering. Hence, for a more understandable and reasonable result of precise k point that has the most extreme error, the manual process should be done to select the proper cluster to perform the best formulation of the k-means algorithm for both experiments. Figure 3. displays the actual elbow graph with the highest error rate performed by manually processing the calculation of the SSE delta in each cluster.

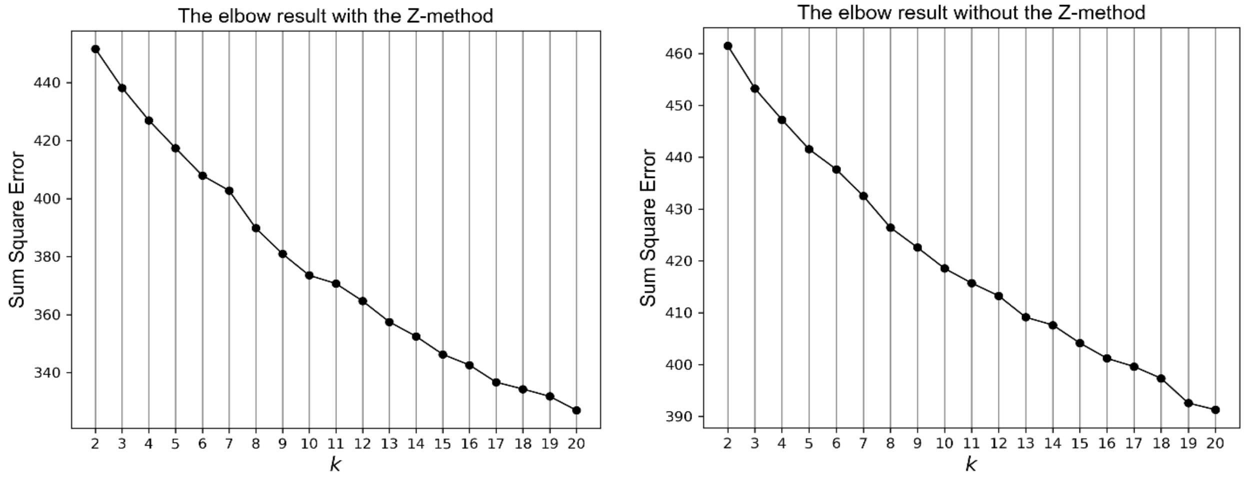

Figure 2. The elbow result with and without the Z-method

Table 2. The DBI value of with and without the Z-method

Experiments	DBI
Filter Stopwords with the Z-method	3,793
Filter Stopwords without the Z-method	4,237

The k point with the most error is at 3 clusters for both experiments, with and without the Z-method, which indicates the best-predicted cluster number for the k-means clustering process. For selecting the k point with the most error decrease precisely, the sum square error delta calculation is carried out through manual comparisons at each defined k point. In other words, the decrease in SSE is not observed based on the magnitude of the number obtained, the difference between the range of cluster points instead. As forecited in the first line, the most extreme error is 3 clusters since it produces the most decreasing error with 13,4596 for the SSE delta. Thus, 3 clusters are used as the input parameter of k-cluster for the Twitter content clustering of Blibli Indonesia Twitter using the k-means algorithm. Table 2 and 3 provide the SSE value based on Figure 2 and SSE delta based on Figure 3 in each experiment (with and without the Z-method applied in the stopwords filtering process) based on defined cluster thoroughly.

Davies Bouldin Index Analyses

The experiments in the stopwords filtering process conducted in this study are evaluated to recognize the performance differences between the experiments that apply the Z-method and do not apply the Z - method. The evaluation of the two experiments conveys more optimal results to undertake the text clustering. The Davies Bouldin Index (DBI) is used to determine the experiment that provides the optimal result by electing the smaller value. Table 4 displays the DBI value for both experiments.

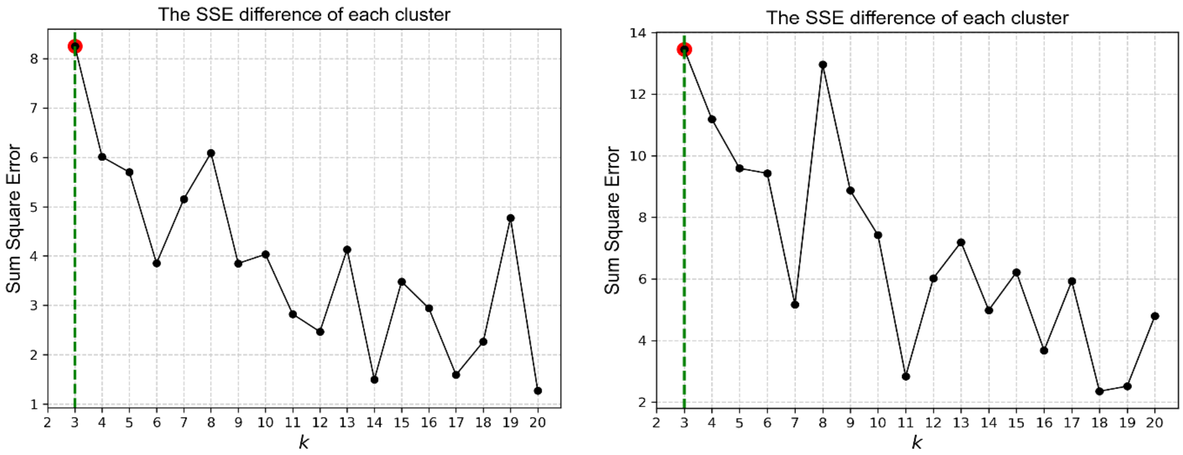

Figure 3. The SSE delta of elbow with and without the Z-method

Table 3. The SSE delta with the Z-method

Cluster number	SSE value	SSE delta
2	451,6376	-
3	438,178	13,4596
4	426,9913	11,1867
5	417,4011	9,5902
6	407,9702	9,4309
7	402, 8027	5,1675
8	389,8395	12,9632
9	380,9679	8,8716
10	373,5458	7,4221
11	370,7104	2,8345
12	364,6904	6,02
13	357,4961	7,1943
14	351,5107	4,9861
15	346,2964	6,2136
16	342,6194	3,677
17	336,6905	5,9289
18	334,3384	2,3521
19	331,8244	2,514
20	327,032	4,7924

Table 4. The SSE delta without the Z-method

Cluster number	SSE value	SSE delta
2	461,4826	-
3	453,2299	8,2527
4	447,2209	6,009
5	441,5222	5,6987
6	437,6697	3,8525
7	432,5189	5,1508
8	426,4298	6,0891
9	422,5793	3,8505
10	418,5445	4,0348
11	415,7211	2,8234
12	413,2544	2,4667
13	409,1204	4,134
14	407,6237	1,4967
15	404,142	3,4817
16	401,1995	2,9425
17	399,6078	1,5917
18	397,3432	2,2646
19	392,5719	4,7713
20	391,2999	1,272

Table 5. The interpretation of tweet content based on the result of the k-means clustering process

Cluster	Word items	Contents
0	bliblihisteria, belanja, parkseojun, histeriabarengpsj, voucher, cashback	The introduction of the newest brand ambassador of Blibli, named Park Seo Jun, a well-known actor from South Korean
1	Emoticonday, electronic, promotion, discount, printer	A series of many promotions for electronic products
2	bliblikam, discount, bliblibelanjaseru, bliblimart, promo	A series of many promotions for camera products

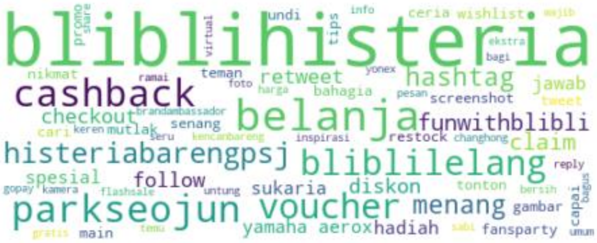

Figure 4a. The existing words in cluster 0

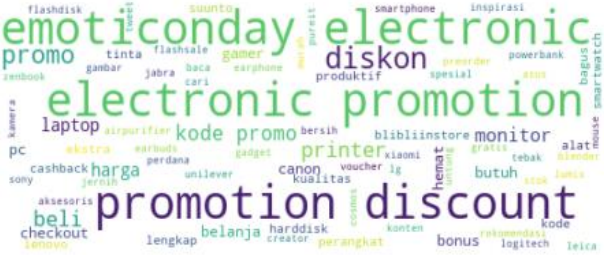

Figure 4b. The existing words in cluster 1

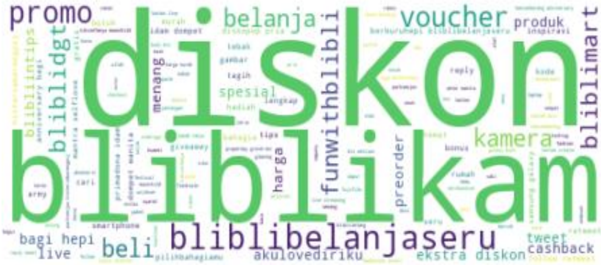

Figure 4c. The existing words in cluster 2

Based on Table 4, it concludes that filtering stopwords with the Z-method provide better clustering result. Furthermore, the Z-method results are used to determine the content clustering of Blibli Indonesia's Twitter with k equals 3 clusters as the content interpretation.

Content Analysis of Blibli Indonesia Twitter

The clustering process is performed by applying k equals 3 clusters as the most optimal point for the input parameter of cluster number. Consequently, there are three content clusters discovered in the tweets of Blibli Indonesia. Miscellaneous content clusters are interpreted by reviewing arranged words and focusing on the ordered data with a minimum distance to its centroid in each cluster. The composition of terms in each cluster can be considered a representation of content types since those arranged words are also reflected as a characteristic of the cluster. This process is performed by using various functions in the two modules named *scikit-learn* to rank which data is close to its centroid and *nltk* to seize the words in each cluster based on ordered cluster centroids. Table 5 elaborates the interpretation of miscellaneous content clusters found in the tweets of Blibli Indonesia.

A set of words in each defined cluster will be convenient to interpret, if the words are visualized precisely. Therefore, the visualization process is carried out through the *word cloud* module. The resulting visualizations of three clusters are shown in Figure 4a,b and c.

Content Priorities of Blibli Indonesia Twitter

The preferable tweet contents are processed by investigating the highest retweets of each content cluster since most of the Twitter users tend to retweet the tweets they like. Therefore, the retweet tally can be examined as a representation of content effectiveness that fits with consumer preference. Specifying the contents that need to be prioritized by Blibli to attract public attention is carried out by selecting the highest retweet count of each content as the fundamental decision for performing social media marketing strategies. The idea behind computing the retweets is commenced by summing the total retweets of each content cluster and then dividing it by the members in each cluster. The entire calculation of this stage can be found out by using the *pandas* module with some existing functions. The family member of the cluster is derived from a function called *value counts*, while the total retweets and the average retweets are derived from a function called *pivot table* by assigning the parameter of aggregate function with two types, namely *sum* and *mean*. Table 6 represents the tally of retweets based on the number of retweets and members of the content clusters.

The interpretation of Twitter content priority that needs to be focused on by Blibli Indonesia is based on the highest retweet tally in each content cluster.

Table 6. The tally of average retweets of each cluster

Cluster	Number of members	Total retweets	Average retweets
0	66	1203	18,227
1	54	59	16,116
2	371	5979	1,093

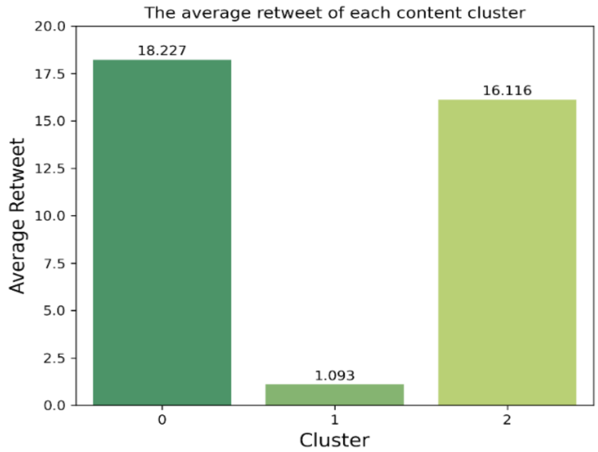

Figure 5. The graph of content priority for social media marketing strategies

Attain a crystal-clear result, the tally from Table 6 needs to be visualized by using a bar chart function in the *matplotlib* module of Python. Figure 5 denotes the graph of average retweet with the unit measure is a retweet count to make a consideration of appropriate social media marketing strategies for Blibli Indonesia in the Twitter platform.

From Figure 5, it can be inferred that the most preferred content tweet is cluster 0, which introduces the newest brand ambassador of Blibli Indonesia, named Park Seo Jun, a famous actor from South Korea. Meanwhile, the minor preferred content tweet is cluster 1, which holds events about special promos for some products related to cameras, and cluster 3, which holds the event named "Emoticon Day" about many promos for some electronic products. Therefore, the content that needs to be focused on as guidance to perform marketing strategy in Twitter is cluster 0, as the public tends to be fond of the contents about *Park Seo Jun* tweets than other cluster contents. Thus, to attract much attention from the public in performing its marketing strategy, Blibli Indonesia needs to consider the contents that meet the public preference by increasing the tweet of cluster 0 and reducing tweets of cluster 1 and 2, since those are yet to fit the Twitter user preference due to its lower retweet tally indication.

The content of *Park Seo Jun* is considered as a content cluster that attracts consumers' attention which can help the company to accelerate its sales and maximize consumer consistency to not change over to other e-commerce [25, 26]. In addition, the consumption pattern of the e-commerce community in Indonesia is influenced by the content tweets about the Korean wave phenomenon that has been rising in Indonesia nowadays [27]). Therefore, maximizing the preferable content can contribute to emerging consumer preference and curiosity towards Blibli Indonesia. Hence, a more incredible opportunity can be obtained for Blibli Indonesia to acquire new consumers willing to use the Blibli application for shopping transactions.

Conclusion

The implementation of k-means and text clustering concept to discover the tweet contents of Blibli Indonesia (@bliblidotcom) delivers three contents, which are considered as the best cluster point obtained through the elbow method. The invention in this study can provide information about the main content that needs to be focused on by Blibli as their marketing strategy in Twitter. Derived from the results, the content with the highest retweets tally is cluster 0, which holds the introduction of the newest brand ambassador of Blibli Indonesia from South Korea, named Park Seo Jun. Meanwhile, the content with the lowest retweets tally is cluster 1, which holds the tweets of special promos for some camera and electronic products. Assumed from the calculation of retweets, the content cluster that needs to be maximized for a marketing strategy is cluster 0, since that cluster is most demanded by Twitter users and provides various advantages for Blibli Indonesia to accelerate the purchasing and avoid consumer inconsistency behavior to change over to other e-commerce.

There are some limitations uncovered from this research. The k-means method is a partitional clustering that requires the best k clusters before executing the algorithm, which indicates the resulting cluster should be well-separated with no overlapping cases. In the meantime, the overlapping cluster seems to be found in this study, marked by the inclusion of two identical words in different clusters. Such overlapping cases are obligatory to be eluded. Hence, for future works, the authors would recommend using other metrics to measure the best k cluster. The more metrics are utilized, the more various results will be obtained. Therefore, the probability of attaining the best cluster number is more remarkable by selecting the one that is evaluated as the appropriate k, which fits with the characteristics of the data. In addition, a reason behind the overlapping cluster is data noise. *Noise* in text mining can be defined as less informative terms contained in the data. The textual data needs various treatments to reduce noises by eliminating non-essential terms. Such conditions can be handled by conducting stopwords removal process using Zipf's Law concept (the Z-method). This study also performs experiments that compare the performance of the clustering results with and without the Z-method on the filtering process. The results convey that appending the Z-method in the process provides a more optimal result because it results in a smaller DBI value. The smaller the DBI value obtained, the more compactness clustering achieved. Hence, the Z-method can be inferred to elevate the clustering quality. Thus, for future works, this study recommends appending the Z-method in filtering stopwords to acquire better clustering results. The key strength of DBI and the Z-method is in their computational time, which is not time-consuming than other methods. Once the process has a relatively short calculation time, it will quickly provide the most converging point and guarantee the more reliable clustering result performance.

Acknowledgement

The authors would like to thank the State University of Malang that have provided financial support through PNBP UM Research Grant 2021 under contract No. 5.3.780/UN32.14.1/LT/2021

References

1. Kinra, A., Beheshti-Kashi, S., Buch, R., Nielsen, T. A. S., and Pereira, F., Examining the Potential of Textual Big Data Analytics for Public Policy Decision-making: A Case Study with Driverless Cars in Denmark, *Transport Policy*, 98, 2020, pp. 68–78.
2. Casas, I., and Delmelle, E. C., Tweeting about Public Transit—Gleaning Public Perceptions from a Social Media Microblog, Case Studies, *Transport Policy*, 2017, 5(4), pp. 634–642.
3. Sivarajah, U., Irani, Z., Gupta, S., and Mahroof, K., Role of Big Data and Social Media Analytics for Business to Business Sustainability: A Participatory Web Context, *Industrial Marketing Management*, 86, 2020, pp. 163–179.
4. Rasool, A., Tao, R., Marjan, K., and Naveed, T., Twitter Sentiment Analysis: A Case Study for Apparel Brands, *Journal of Physics: Conference Series,* 1176, 2019, pp. 1-7.
5. Dastanwala, P. B., and Patel, V., A Review on Social Audience Identification on Twitter using Text Mining Methods, 2016 *International Conference on Wireless Communications, Signal Processing and Networking* (WiSPNET), 2016, pp. 1917–1920.
6. Yusril, A. N., Larasati, I., and Aini, Q., Implementasi Text Mining untuk Advertising dengan Menggunakan Metode K-Means Clustering pada Data Tweets Gojek Indonesia, *SISTEMASI: Jurnal Sistem Informasi*, 9(3), 2020, pp. 586-596.
7. Pratama, E. E., and Atmi, R. L., A Text Mining Implementation Based on Twitter Data to Analyse Information Regarding Corona Virus in Indonesia, *Journal of Computers for Society*, 1(1), 2020, pp. 91-100.
8. Ahmed, M., Seraj, R., and Islam, S. M. S., The k-means Algorithm: A Comprehensive Survey and Performance Evaluation, *Electronics*, 9(8), 2020, pp. 1-12.
9. Xiong, C., Hua, Z., Lv, K., and Li, X., An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers, 2016 7th *International Conference on Cloud Computing and Big Data (CCBD)*, 2016, pp. 265–268.
10. Sinaga, K. P., and Yang, M.-S, Unsupervised K-Means Clustering Algorithm, *IEEE Access*, 8, 2020, pp. 1-13.
11. Suresh Babu, S., and Jayasudha, K., A Survey of Nature-inspired Algorithm for Partitional Data Clustering, *Journal of Physics: Conference Series*, 1706(1), 2020, pp. 2-11.
12. Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., and Satoto, B. D., Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster, *IOP Conference Series: Materials Science and Engineering*, 336(1), 2020, pp. 1-6.
13. Reyhana, Z., *Analisis Sentimen Pendapat Masyarakat terhadap Pembangunan Infrastruktur Kota Surabaya melalui Twitter dengan Menggunakan Support Vetor Machine dan Neural Network*, Tesis, Jurusan Statistika, Institut Teknologi Sepuluh Nopember, 2018.
14. HaCohen-Kerner, Y., Miller, D., and Yigal, Y., The Influence of Preprocessing on Text Classification using a Bag-of-words Representation, *PLOS ONE*, 2020, 15(5), pp. 1-22.
15. Alam, S., and Yao, N., The Impact of Preprocessing Steps on the Accuracy of Machine Learning Algorithms in Sentiment Analysis, *Computational and Mathematical Organization Theory*, 25(3), 2019, pp. 319–335.
16. Saif, H., Fernandez, M., He, Y., and Alani, H., On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter, *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14),* 2014, pp. 810-817.
17. Sari, D., Sari, Y., Furqon, M., Pembentukan Daftar Stopword Menggunakan Zipf Law dan Pembobotan Augmented TF - Probability IDF pada Klasifikasi Dokumen Ulasan Produk, *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 4(1), 2020, pp. 406-412.
18. Mohammadi, M., Parallel Documenta Identification using Zipf's Law, *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*, 2016, pp. 21-25.
19. Bengfort, B., Bilbro, R., and Ojeda, T. *Applied Text Analysis with Python*, O'Reilly Media, Inc. 2016.

20. Liu, L., Peng, Z., Wu, H., Jiao, H., Yu, Y., and Zhao, J., Fast Identification of Urban Sprawl Based on K-Means Clustering with Population Density and Local Spatial Entropy, *Sustainability*, 10(8), 2018, pp. 1-16.
21. Umargono, E., Suseno, J. E., and S. K., V. G., K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median, *Proceedings of the International Conferences on Information System and Technology*, 2019, pp. 234–240.
22. Wu, C., Yan, B., Yu, R., Yu, B., Zhou, X., Yu, Y., and Chen, N., K-Means Clustering Algorithm and Its Simulation Based on Distributed Computing Platform, *Complexity*, 2021(1), 2021, pp. 1–10.
23. Enesi, I., Liço, L., Biberaj, A., and Shahu, D., Analysing Clustering Algorithms Performance in CRM Systems, *Proceedings of the 23rd International Conference on Enterprise Information Systems*, 2021, pp. 803–809.
24. Yang, Y., Han, D., Liang., S., Clustering Validity Index for Irregular Clustering Result, *Applied Soft Computing Journal*, 95(2020), 2020, pp 1-17.
25. Putri, R. K., Warsito, B., and Mustafid, M., Implementasi Algoritma Modified Gustafson-Kessel untuk Clustering Tweets pada Akun Twitter Lazada Indonesia, *Jurnal Gaussian*, 8(3), 2019, pp. 285–295.
26. Liya, I., Budiono, H., Sanjaya, V. F., and Suratmin, J. E, Pengaruh Hallyu Wave, Brand Ambassador, dan WOM terhadap Keputusan Pembelian pada Mie Sedap Selection Korean Spicy, *REVENUE: Jurnal Manajemen Bisnis Islam*, 2(1), 2021, pp. 11-26.
27. Sagia, A., and Situmorang, S. H., Pengaruh Brand Ambassador, Brand Personality dan Korean Wave terhadap Keputusan Pembelian Produk Nature Republic Aloe Vera, *Jurnal Manajemen dan Bisnis Indonesia*, 5(2), 2018, pp. 286–298.