Analyzing Social Media Marketing Strategies of Indonesian E-commerce using Text Mining Techniques

Raretha Maren, Aisyah Larasati*
Industrial Engineering Department
State University of Malang
Malang, Indonesia
*Corresponding author: aisyah.larasati.ft@um.ac.id

Retno Wulandari
Mechanical Engineering Department
State University of Malang
Malang, Indonesia

Abstract

The application of social media in the field of marketing has grown popular globally. Social media becomes a new preference to many businesses for promoting and advertising. The transformation of conventional marketing into social media is due to its cost-effectiveness. Furthermore, the rapid dispersion of information has led many business people to switch their marketing media into the appropriate instrument by using Twitter. This study implements text mining and k-means for clustering tweets from the Twitter of Indonesian e-commerce, Blibli (@bliblidotcom). This study aims to segment the tweet contents which Blibli can focus on certain contents preferred by Twitter users as their marketing strategies and discover the best formulation of applying k-means. The optimal cluster for k-means accomplished by silhouette method with two distance metrics. The finding of this study provides cosine as the optimal formulation for text clustering problem. The outcome of existing experiments with cosine shows that 15 clusters as the best number. The result of tweet clustering according to the best k-means formulation indicates that the Twitter users tend to like the content about quiz programs named "Fun with Blibli". Hence, Blibli Indonesia can prioritize that content as marketing strategy in Twitter platform.

Keywords

E-commerce, Social Media Marketing, Text Mining, K-means, Silhouette Coefficient.

1. Introduction

With the advent of the digital era, social media has undergone the major transformations in various fields. The application of social media has become a core element for many people in the world. One of which is business people such as an e-commerce business (Manohar Singh and Singh 2018). Social media becomes the preferred e-commerce platform for advertising and product marketing due to its cost-effectiveness (Dwivedi et al. 2020). The simplicity of distributing information rapidly is the reason for media transformation development in the marketing field by utilizing the sophisticated platforms, such as Twitter. Twitter is a well-known social media with 19,5 million users in Indonesia. With the existence of retweet and like features, Twitter facilitates business people to propagate the tweets for promoting and advertising their products. One of the e-commerce businesses that allows Twitter as its marketing device is Blibli Indonesia. Blibli utilizes Twitter to provide information about discounts, product promotions, other offerings that can attract consumer to make transactions with the Blibli e-commerce application. An intense business competition requires Blibli Indonesia to implement a precise marketing strategy. Communication through the tweets obliges to be executed properly by conceiving preferable contents to acquire positive responses from consumer. Practically, Blibli Indonesia tends to post random tweets that are yet to fit consumer interests. Consequently, Blibli has become less recognized by the public. Hence, there are just few people who are willing to make transactions through the Blibli e-commerce application.

The integration of text mining and clustering method is the appropriate alternative to handle this problem. This is based on the ability of text mining to extract textual data into new insight that corresponds to the business purposes (Gupta et al. 2020). The next stage of text mining employs clustering method for content segmentation, which can be accomplished by using the k-means algorithm due to its low time complexity and rapid calculation for handling textual data (Syakur et al. 2018). In practice, the k-means method requires the proper cluster number to achieve optimal condition precisely. Accordingly, the replenishment of the silhouette coefficient method is required to acquire the appropriate cluster for maximizing the clustering performance result.

1.1 Objectives

There are two objectives in this study. First, to discover the preferable tweet contents of Twitter user or the followers of Blibli Indonesia, which later on, this result can be used as a fundamental of decision making to employ social media marketing strategy in Twitter platform for Blibli Indonesia. Moreover, this result is expected to provide information for Blibli e-commerce to focus on certain tweet contents which are preferred by Twitter users. Therefore, the higher probability of consumer being interested in Blibli, the more consumer will be willing to make a transaction through Blibli e-commerce application. Second, to invent the best formulation for text clustering problem based on the experiment of k-means algorithm with two distance measures, namely cosine and euclidean distance. Once the best formulation between two distance measures is discovered, it leads to provide the optimal results for content clustering.

2. Literature Review

The utilization of social media as marketing tools provides a million advantages and challenges for business people (Abed et al. 2016). Social media not only allows low-cost marketing, but also enhances proximity to the consumer since the proactive interaction between consumer and business people do exist. Related to the implementation of Twitter as a social media marketing tool, consumer engagement becomes an instrument for business people to explore the interest of consumer (McCorkle and Payan 2017). Once the business people find consumer tendency, they will conveniently assign and focus on those contents that fit with consumer interest as marketing strategies (Appel et al. 2020). The appropriate tweet contents play a significant role in the success of social media marketing. The analysis to discover the proper contents can be performed by observing the uploaded tweets from the Twitter account of business people (Yusril et al. 2020).

Putri et al. (2019) propose that text mining is a worthwhile method for analyzing tweet uploads since most of the available information on Twitter is saved in textual form. Text mining is the process of acquiring new insights from textual data by observing specific patterns (Hassani et al. 2020). The text mining process involves the integration of various disciplines such as machine learning, linguistics, and statistics to extract textual data into useful information to encounter analytical needs in achieving business objectives. The results of textual data extraction are term-weighting of words which can be used as the input data for advanced analysis process using machine learning (Weißer et al. 2020). As purposed by Kobayashi et al. (2018), that text mining involves the utilization of machine learning to attain new insights, clustering method can be applied to cluster content types corresponding to the characteristic of the tweets likewise. The proper term to define this research problem is text clustering.

Clustering is a technique to classify a set of data into several classes based on the level of similarity and dissimilarity, where data with similar characteristics will be grouped into one cluster and vice versa (Usino et al. 2019). The most prominent clustering algorithm due to its low time complexity and fast convergence is derived from partitional clustering, namely k-means (Orkphol and Yang 2019). In running its function, the k-means algorithm groups the data into one class by minimizing the distance between data item and random cluster center respectively (Kim and Gil 2019). In practice, k in k-means is specified subjectively by a human. Thus, this might provide suboptimal results. To overcome the shortcoming of k-means algorithm, one method can be applied to determine the best number of clusters is silhouette coefficient method (Yi et al. 2017). Silhouette coefficient method not only useful for discovering optimal cluster, but also serviceable for quantifying the quality of how precisely items are placed in the appropriate class based on cluster density (Hidayat and Yaqin 2019). The evaluation of similarity or dissimilarity in k-means is based on the distance measure. The techniques of selecting the proper distance measurement can be finished by experimenting several distance metrics to achieve optimal clustering results. The applicable distance measure includes cosine and euclidean (Al-Anazi et al. 2016, Harsemadi 2018).

To summarize, the application of text mining is a worthwhile method for extracting the unstructured data from Twitter, converting it into structured form by transforming and representing words respectively to numerical units as the input for the advanced analysis. The advance analysis utilized is text clustering to determine tweet contents of Indonesian e-commerce Twitter account, namely Blibli. The clustering process can be resolved by combining the k-means algorithm and silhouette method with the experiment of two distance measures, including cosine and euclidean, to find the best formulation of the cluster as well as acquiring maximize the content clustering results.

3. Methods

The research design is divided into four stages, including the preprocess of the textual data, the determination of the optimal clusters, the formulation of the k-means method, and analyzing tweets content. The data used in this case are 491 tweets which consist of two independent variables, namely text as an internal variable included in the model building, and retweetcount as an external variable for fundamental decision making. Raw data is processed with the utilization of anaconda environment using python programming 3.0, involving a number of modules such as "pandas", "emoji", "regular expression", "sastrawi", "nltk", "scikit-learn", "yellow brick", and "matplotlib".

The initial step is text preprocessing, a worthwhile step for transforming and cleaning textual data. Text preprocessing commences with eliminating emoticons, digits, usernames, and URLs attached to the tweets by using "emoji" and "regular expression" modules. The second step is case folding, to convert uppercase letter into lowercase since all the tweets do not apply consistent letter form. The third step is stemming, to homogenize words into basic form by removing the affixed, according to the confix stripping idea in "sastrawi" module. The fourth step is eliminating punctuation as the punctuation has no special meaning in textual data. The following step is filtering words, which is to remove unnecessary words based on a compiled dictionary of the authors in txt format by using "sastrawi" module. If a non-standard word is found, the authors will replace that word with a normative synonym. The next step is the tokenizing process by using "nltk" module, which functions is to parse the sentences into individual term in order to simplify word occurrence frequency as well as the term-weighting process. The final step of text preprocessing is term-weighting based on TF-IDF concept with the usage of a module named "nltk". TF-IDF mechanism is to calculate term-weighting from word occurrence and multiplied by document frequency with a logarithmic scale. The TF-IDF value will be used as an input for clustering (Roul et al. 2017).

The clustering process applies k-means method with the utilization of "scikit-learn" module. Since k in k-means is a subjective case in determining the number of clusters, the best cluster can be observed by using the silhouette method that provides the highest score in the defined cluster range (Naeem and Wumaier 2018). The determination of optimal cluster is executed 15 times with a total cluster ranged from 2 to 16 and experimented with two distance measures. The best score obtained from the silhouette method is used as the real input k for k-means method to classify the tweet contents of the Twitter account of Blibli Indonesia. A number of the k-means parameters defined in this study are shown in Table 1.

Table 1. A Number of Parameters in K-means Algorithm

Parameter	Value	Description
n_cluster	2 – 16	The range of cluster number to form
init	k-means ++	Method for initializatizing of centroids which provide fast converging
max_iter	100	Maximum iterations of k-means clustering for a single run
n_init	10	Number of k-means algorithm run with different seeds
random_state	42	Specifies the random number generator for initializing centroid. Selecting 42 as an input parameter to avoid arbitrary results and to

		maintain reproducibility
metric	euclidean and cosine distance	Distance measurement between data item and cluster centers.

After discovering the best formulation of the cluster, an advanced analysis will be performed regarding the preferable tweet contents of Twitter users by inspecting the retweet calculation of formed clusters respectively (Indraloka and Santosa, 2017). The high-low value of retweet calculation reflects the interest level of Twitter users to tweet contents uploaded by Blibli Indonesia. Hence, the higher retweet calculations, indicate the more those tweet contents will be preferred by the public. Thus, tweet content that has the highest retweet calculation, can be created as a fundamental of decision making for Blibli Indonesia to perform the precise social media marketing strategy by focusing on some certain content tweets which Twitter users are interested in.

4. Data Collection

This research uses primary data, collected through the twitter crawling process based on keys and tokens from API Twitter. The process of twitter crawling utilizes R Studio 1.4.1 software, involving a number of modules, such as “xlsx”, “twitter”, “ROAuth”, and “RCurl”. The data is sourced from Indonesian e-commerce Twitter account (@bliblidotcom), with the date of uploaded tweets is commenced from 28 May 2020 to 4 March 2021.The total data obtained is 491 tweets that consist of two variables, namely text and retweetcount. Text is the tweets from the Twitter account of Blibli Indonesia that only use the Indonesian language, while retweetcount is the number of retweets from tweets respectively.

5. Results and Discussion

5.1 Discovering the Optimal Cluster with Silhouette Method

The silhouette value indicates the sensibility of data items is placed in the appropriate clusters. In this case, each word is portrayed as a data item to find silhouette value. The calculation of silhouette coefficient is executed 15 times by utilizing the “scikit-learn” module, with a total cluster is ranged from 2 to 16 and experimented with two distance measures, namely cosine and euclidean. In python programming, the rule of range function represented the cluster interval, is inclusive at the beginning and exclusive at the end. Thus, the number 16 is not included in the calculation. The higher silhouette value obtained, the better quality of clustering achieved, and vice versa. Table 2 shows the silhouette coefficient value of 15 clusters.

Table 2. The Silhouette Coefficient Result

The Number of Cluster (k)	Silhouette Coefficient with Cosine Distance	Silhouette Coefficient with Euclidean Distance
2	0,0634	0,0368
3	0,0794	0,0463
4	0,0941	0,0542
5	0,1104	0,0635
6	0,122	0,0711
7	0,1249	0,0783
8	0,1479	0,0862
9	0,1666	0,0973
10	0,1805	0,1054
11	0,1767	0,1037
12	0,1708	0,1001
13	0,1843	0,1081
14	0,1869	0,1102
15	0,1981	0,1181

The highest silhouette coefficient for both the experiment of cosine and euclidean distance provides the same number of clusters, which are 15 clusters, with the values obtained 0, 1981 and 0,1181 respectively. Keep in mind that in the silhouette coefficient concept, the measurement of cosine and euclidean distance has many distinctions in practice. Hence it will provide a different value in each specific distance measurement. According to Table 2, the silhouette value of cosine distance is slightly higher than the silhouette value of euclidean distance, which denotes that cosine distance shows better quality than euclidean distance. Thus, it concludes that the best parameter input of cluster number for running the k-means process is 15 clusters, obtained from the silhouette method based on cosine distance.

5.2 Analyzing Tweet Contents of Blibli Indonesia

In this case, the parameter of cluster number uses 15 to cluster tweet contents of Blibli Indonesia. After running the k-means model, the highest word occurrence will be counted based on the centroid of clusters respectively. The arranged word of each cluster is considered as a representation of tweet contents uploaded by Blibli Indonesia on the Twitter platform. The analyzing tweet contents utilizes the "scikit-learn" module to interpret the content of each cluster. Table 3 displays the discovered miscellaneous tweet contents on the Blibli Twitter account.

Table 3. Miscellaneous Tweet Contents of Blibli

Cluster	Words	Tweet Contents
0	bliblimart, properday, diskopop, buy	The promo offers for household needs in the events named "Blibli Mart", "Proper Day", and "Disko POP"
1	bliblihappyshopping, find, happy, guess, clue.	The riddle games in the event named "Blibli Happy Shopping".
2	bliblihysteria, blibliauction, cashback, shopping.	The auctions and cashback offer in the event named "Blibli Auction".
3	giveaway, happy, anniversary, dating.	The giveaway and dating event to celebrate the anniversary of Blibli Indonesia.
4	electronic, emoticonday, promo, discount.	The promo offers for electronic product in the event named "Emoticon Day".
5	funwithblibli, retweet, like, reply.	The prizes quiz program in the event named "Fun with Blibli".
6	price, inexpensive, primadona, women.	The promo offers for several products of women needs in the event named "Primadona".
7	extra, discount, bliblihappyshopping.	The extra discount offers in the event named "Blibli Happy Shopping".
8	bliblicam, camera, discount, promo.	The discount promo for several camera products in the event named "Blibli Cam".
9	blibliintips, tips, tricks, live, inspiration.	Miscellaneous tips and tricks requested by Twitter users in the event named "Blibli in tips".
10	bliblidgt, payment, quota, internet.	The payment service of electric bills and selling internet quota through Blibli application in the event named "Blibli DGT".
11	hysteriawithpsj, parkseojun, fansparty, brandambassador.	Introducing the new international brand ambassador Blibli through fans party with Park Seo Jun, South Korean Actor, in the event named "Hysteria With PSJ".
12	ilovemyself, selflove, giveaway, prize, happy.	The giveaway with self-love topic in the event named "I Love Myself".
13	claim, voucher, bliblihappyshopping, delighted, challenge.	The claiming of many vouchers provided by Blibli as many as possible in the event named delighted challenge in the series of "Blibli Happy

		Shopping" event.
14	army, bts, edition, Samsung, buds, giveaway.	The giveaway of Samsung products such as BTS Army Edition earbuds. BTS is a South Korean idol, while Army is the name of its fandom.

5.3 Analyzing the Clustering Density

The density of clustering results is evaluated by averaging the silhouette sample of each cluster, which is known as the silhouette index. The best cluster quality indicates the highest cluster density. Robani and Widodo (2016) propose that the closer the distance between data items in a cluster as well as the further the data items between other cluster groups indicate the better quality of clustering. The criteria to evaluate silhouette index ranged between -1 and 1. The index value that close to 1 provides good quality, and vice versa. Meanwhile the index value that closes to 0 provides an overlapping cluster. The overlapping cluster occurs as a result of finding the same data items in two different clusters. The cluster density visualization can be accomplished by utilizing the "yellow brick" module in python programming. The graph displays the silhouette index plotting for each cluster. The vertical thickness shows the data size in clusters and the dashed red line shows the global silhouette. Keep in mind that the global silhouette is a synonym of the silhouette coefficient. This visualization is not only serviceable to investigate cluster density, but also to examine cluster imbalance. Clusters that have vertically thick and short plotting interpreted as an imbalance of total clusters initialized, which is supported by a low score of silhouette index achieved. Thus, the solution to handle such problem is initialize more extensive cluster interval.

Most of the silhouette index obtained in this study has a positive value, plotted on the right side of 0 point. Besides that, the long plotting indicates the silhouette index of each cluster is high adequately and the initialized range cluster immaculately fits to the model. Thus, it can be concluded that the clustering of tweet content using the k-means and the combination of the silhouette method with the experiment of cosine distance provides the most optimal result. Figure 1 displays the cluster density of 491 data tweets.

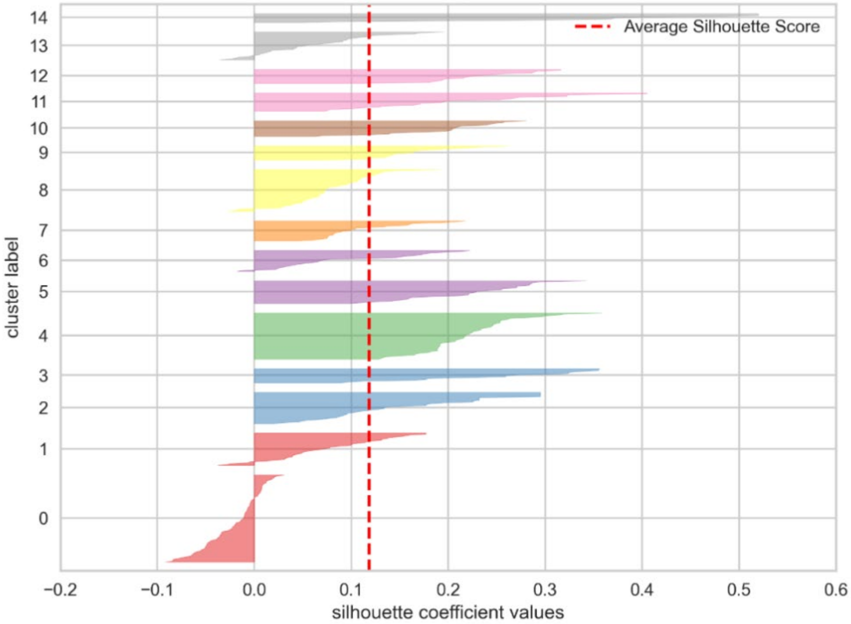

Figure 1. The Visualization of Silhouette Index

5.4 Determining the Tweet Contents for Business Strategy of Blibli Indonesia

The tweet contents preferred by Twitter users are determined through the average number of retweets in each cluster formed. The highest retweet indicates that Twitter users are interested in that tweet content. Hence, the average number of retweets is considered as a representation of the interest level of Twitter users in particular tweet content. The average number of retweets is achieved from retweetcount variable. The calculation concept is initiated by tallying the retweets in each cluster and then divided by the member of its cluster. The calculation is processed by utilizing groupby function in "pandas" module with sum and average aggregation in each existing cluster. Table 4 shows the retweet calculation of each cluster.

Table 4. The Retweet Calculation Results

Cluster Number	Number of Cluster Members	Sum of Retweets	Average Number of Retweets
Cluster 0	101	569	5,633
Cluster 1	38	129	3,394
Cluster 2	37	304	8,216
Cluster 3	17	451	26,529
Cluster 4	54	59	1,092
Cluster 5	27	3414	126,444
Cluster 6	25	65	2,6
Cluster 7	24	61	2,541
Cluster 8	49	74	1,510
Cluster 9	17	56	3,294
Cluster 10	19	18	0,947
Cluster 11	22	1109	50,409
Cluster 12	17	125	7,352
Cluster 13	33	384	11,636
Cluster 14	11	423	38,454

The priority of the tweet contents of Blibli Indonesia is obtained by observing the highest number of retweets of the clusters. The utilization of "matplotlib" module is to visualize the precise tweet content priorities by displaying the bar charts. Figure 2 shows the bar chart of the average number of retweets for each tweet contents.

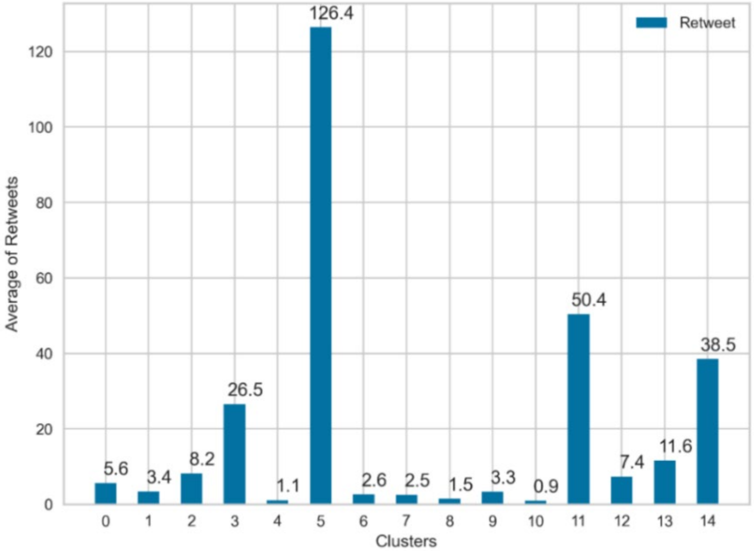

Figure 2. The Average Retweet of The Contents

According to Figure 2, the highest average number of retweets is cluster 5, which contains the prizes quiz program in the event named "Fun with Blibli". The second priority is cluster 11, which tweet content is about the introduction of Park Seo Jun as a new international brand ambassador of Blibli Indonesia, through a fans party named "Hysteria with PSJ". Park Seo Jun is a South Korean actor adored by the Indonesian people due to his capability in acting for many drama series. The third priority is cluster 14, which contains the tweet content about Samsung product giveaway through the collaboration of Blibli and BTS, a South Korean idol. Whereas the tweet contents with the lowest average number of retweets are cluster 8, 4, and 10, containing promos, cashback and discount for electronic products as well as the payment service of electricity bills through Blibli application.

6. Conclusion

The findings in this study provide information about the priority of tweet contents that needs to be focused by Blibli Indonesia as its marketing strategy recommendation. The results elaborate that most of Twitter users prefer the tweet contents about quiz program, giveaway, and collaborations of Blibli with South Korean artists to the tweet contents about promo, discount, and cashback offers. Therefore, Blibli e-commerce is expected to focus on the tweet contents with high-retweeted cluster, such as quiz program in the event named "Fun with Blibli" for its marketing strategy in the Twitter platform. Furthermore, this study also provides information for future works in formulating the parameter tuning of cluster numbers as well as the distance measures while running the k-means algorithm. Regarding the applied model, the combination of the k-means algorithm and silhouette method with the experiment of cosine distance shows the optimal result for the text clustering problem. Thus, this research can be used as a reference for the next researchers related to the appropriate algorithm combination to attain optimal condition rapidly.

Acknowledgements

The authors would like to appreciate the chancellor and the advisory lecturer of Universitas Negeri Malang (UM), who have funded this research publication through PNBP UM Research Grant 2021 under contract No. 5.3.780/UN32.14.1/LT/2021

References

Abed, S. S., Dwivedi, Y. K., and Williams, M. D., Social commerce as a business tool in Saudi Arabia's SMEs, *International Journal of Indian Culture and Business Management*, vol. 13, no. 4, pp. 1-19, 2016.

Al-Anazi, S., AlMahmoud, H., and Al-Turaiki, I., Finding Similar Documents Using Different Clustering Techniques, *Procedia Computer Science*, vol. 82, pp. 28–34, 2016.

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T., The future of social media in marketing, *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 79–95, 2020.

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., and Wang, Y., Setting the future of digital and social media marketing research: Perspectives and research propositions, *International Journal of Information Management*, vol. 59, no. 1, pp. 3-37, 2020.

Gupta, A., Dengre, V., Kheruwala, H. A., and Shah, M., Comprehensive review of text-mining applications in finance, *Financial Innovation*, vol. 6, no. 39, pp. 1-25, 2020.

Harsemadi, I. G., Perbandingan Distance Measure pada K-means Clustering untuk Pengelompokkan Musik terhadap Suasana Hati, *Seminar Nasional Teknologi Informasi dan Multimedia 2018*, vol. 1, no. 2, pp. 13-18, 2018.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., and Yeganegi, M. R., Text Mining in Big Data Analytics*, Big Data and Cognitive Computing*, vol. 4, no. 1, pp. 1-34, 2020.

Hidayat, W., and Yaqin, A., Business Trends Based on News Portal Websites for Analysis of Big Data Using K-Means Clustering, *2019 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, July 24–25, 2019.

Indraloka, D. S., and Santosa, B., Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia, *Jurnal Sains dan Seni ITS*, vol. 6, no. 2, pp. A51–A56, 2017.

Kim, S.-W., and Gil, J.-M., Research paper classification systems based on TF-IDF and LDA schemes, *Human-Centric Computing and Information Sciences*, vol. 9, no. 30, pp. 1-21, 2019.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., and Den Hartog, D. N., Text Mining in Organizational Research, *Organizational Research Methods*, vol. 21, no. 3, pp. 733–765, 2019.

Manohar Singh, D., and Singh, G., Impact of social media on e-commerce, *International Journal of Engineering & Technology*, vol. 7, no. 2.30, pp. 21-26, 2018.

McCorkle, D., and Payan, J., Using Twitter in the Marketing and Advertising Classroom to Develop Skills for Social Media Marketing and Personal Branding, *Journal of Advertising Education*, vol. 21, no. 1, pp. 33–43, 2017.

Naeem, S., and Wumaier, A., Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K, *International Journal of Computer Applications*, vol. 182, no. 31, pp. 7–14, 2018.

Orkphol, K., and Yang, W., Sentiment Analysis on Microblogging with K-Means Clustering and Artificial Bee Colony, *International Journal of Computational Intelligence and Applications*, vol. 18, no. 03, pp. 1-22, 2019.

Putri, R. K., Warsito, B., and Mustafid, M., Implementasi Algoritma Modified Gustafson-Kessel untuk Clustering Tweers pada Akun Twitter Lazada Indonesia, *Jurnal Gaussian*, vol. 8, no. 3, pp. 285–295, 2019.

Robani, M., and Widodo, A., Algoritma K-Means Clustering Untuk Pengelompokan Ayat Al Quran Pada Terjemahan Bahasa Indonesia, *Jurnal Sistem Informasi Bisnis*, vol. 6, no. 2, pp. 164-176, 2016.

Roul, R. K., Sahoo, J. K., and Arora, K., Modified TF-IDF Term Weighting Strategies for Text Categorization, *2017 14th IEEE India Council International Conference (INDICON)*, Roorkee, India, December 15–17, 2017.

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*, Surabaya, Indonesia, November 9, 2017.

Usino, W., Satria, A., Hamed, K., Bramantoro, A., A, H., and Amaldi, W., Document Similarity Detection using K-Means and Cosine Distance, *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 165-170, 2019.

Weißer, T., Saßmannshausen, T., Ohrndorf, D., Burggräf, P., and Wagner, J., A clustering approach for topic filtering within systematic literature reviews, *MethodsX*, vol. 7, pp. 1-10, 2020.

Yi, J., Zhang, Y., Zhao, X., and Wan, J, A Novel Text Clustering Approach Using Deep-Learning Vocabulary Network, *Mathematical Problems in Engineering*, vol. 2017, pp. 1–13, 2017.

Yusril, A. N., Larasati, I., and Aini, Q., Implementasi Text Mining untuk Advertising dengan Menggunakan Metode K-means Clustering pada Data Tweets Gojek Indonesia, *Jurnal Sistem Informasi*, vol. 9, no. 3, pp. 586-596, 2020.

Biographies

Aisyah Larasati is a senior lecturer, associate professor, and a head of undergraduate program in Industrial Engineering at State University of Malang, Indonesia. She achieved bachelor degree in Industrial Engineering from Institut Teknologi Sepuluh Nopember (ITS) Surabaya, Indonesia, Masters in Industrial Engineering from Institut Teknologi Sepuluh Nopember (ITS) Surabaya, Indonesia, and Master in Industrial Management from Hogeschool van Arnhem en Nijmegen (HAN University), Netherlands. Afterwards, she earned her PhD degree in Industrial Engineering and Management from Oklahoma State University, USA. Her passion in data analytics, statistics, and service quality has led her to publish a number of journals and papers.

Retno Wulandari is a senior lecturer and a head of undergraduate program in Mechanical Engineering at State University of Malang, Indonesia. She achieved bachelor, masters, and PhD in Mechanical Engineering from Brawijaya University (UB) Malang, Indonesia. She has been teaching for over 21 years. She has published many journals and conference papers. Her research fields include energy conversion machine, fluid mechanics, conversion and management energy, thermo fluid, thermodynamic, and heat transfer.

Raretha Maren is a candidate for Bachelor Degree in Industrial Engineering from State University of Malang, Indonesia. She has become a lab assistant for web mining and text mining courses. She has participated in a number of lecturer projects and written some international paper publications. Her passion is in data mining, text mining, and web mining.