
CUSTOMER EXPERIENCE ANALYSIS SKINCARE PRODUCTS THROUGH SOCIAL MEDIA DATA USING TOPIC MODELING AND SENTIMENT ANALYSIS

Muhammad Habibi^{1,}, Kartikadyota Kusumaningtyas²*

^{1,2}Department of Informatics, Universitas Jenderal Achmad Yani Yogyakarta, Yogyakarta, Indonesia

**Email: muhammadhabibi17@gmail.com*

Submitted : 13 November 2022; Revision : 26 May 2023; Accepted : 7 June 2023

ABSTRACT

Currently, skin care products (skincare) are popular among the public. Both men and women are interested in buying skin care products. Moreover, there are many brands of skin care products that are divided into several types of facial and body care, such as moisturizers, toners, cleansers, and masks. Therefore, many consumers take the time to find information, for example, in terms of price, quality, and brand for decision-making. A lot of useful information is in the form of Twitter messages known as tweets which are sent from people who use skin care products because Twitter is one of the online social media where users can share their opinions and experiences. However, consumers still have to spend a lot of time searching, reading, and understanding the comprehensive collection of tweets before buying skin care products.

The purpose of this study is to analyze customer experience, analyzing automated tweets about skin care products. Tweets about skin care products will be subjected to a topic modeling process to find out what topics are being discussed. In addition, the topics that have been obtained will be subject to sentiment analysis in the form of positive and negative messages for skin care products. Consumers who are app users don't waste time reading and analyzing large amounts of data manually and they can decide to buy skin care products more easily.

The results of this study obtained 14 topics of discussion related to skincare. Meanwhile, the sentiment analysis results of 14 topics resulted in more positive sentiment class tweets overall. It related the category topic that has the number of tweets to the importance of skincare. In addition, categories related to ingredients for skincare products from nature, namely fruits and spices, are the topics that have the second highest number of tweets. The results of the analysis of tweets related to user experience on Twitter, it was found that users prefer skincare products that use ingredients from nature.

Keywords Topic Modeling; Social Media; Skincare; Customer experience

Paper type Research paper

INTRODUCTION

The massive use of social media by the community today makes the dissemination of information very fast. Twitter is one of the most popular social media platforms among internet users today. Based on data from Oberlo, Twitter's monthly active users in 2020 were 192 million daily active users [1]. The data generated by social media such as Twitter provides valuable information to many stakeholders regarding the behavior, preferences, tastes and characteristics of users. We can use this information as a reference in policy making, for example, to develop a marketing strategy for the company.

Currently, skin care products (skincare) are popular among the public. Both men and women are interested in buying skin care products. Moreover, there are many brands of skin care products that are divided into several types of facial and body care, such as moisturizers, toners, cleansers and masks. Therefore, many consumers take the time to find information, for example, in terms of price, quality and brand for decision making. A lot of useful information is in the form of Twitter messages known as tweets which are sent from people who use skin care products because Twitter is one of the online social media where users can share their opinions and experiences. However, consumers still have to spend a lot of time searching, reading, and understanding the comprehensive collection of tweets before buying skin care products.

The purpose of this study is to analyze customer experience, analyzing automated tweets about skin care products. Tweets about skin care products will be subjected to a topic modeling process to find out what topics are being discussed. In addition, the topics that have been obtained will be subject to sentiment analysis in the form of positive and negative messages for skin care products. Consumers who are app users don't waste time reading and analyzing large amounts of data manually and they can decide to buy skin care products more easily. Opinion mining with sentiment analysis combines natural language processing techniques based on machine learning, and text data mining with computational linguistics to analyze opinions or attitudes on various article topics. The purpose of this analysis is to evaluate the emotions and feelings of communication as aspects of product satisfaction [2]–[5]. There are several related studies as follow.

Research [6] is an opinion mining application to identify comment messages about good or bad airline services on Twitter. The opinion analysis process consists of two parts: message filtering to identify only subjective messages and comment analysis to classify positive (good) or negative (bad). This application can help customers to easily differentiate the services of this airline, and airlines to find out easily how to improve their services. Research [7] is a framework for classifying aspects of laptop reviews from one of the leading laptop review websites. It identified each paragraph of an individual laptop review page as subjective or not. Then, the aspects of each subjective paragraph will be classified. Therefore, this research can be used to analyze laptop reviews to help customers make decisions before buying.

Research [8] is the development of an application to classify comments about recipes from one of the popular food community sites into positive or negative sentiments. Another research is to analyze sentiment regarding the services of the Social Security Administrative Body (BPJS) for Health in Indonesia [9]. This study analyzes tweet sentiment about BPJS services into 4 sentiment classifications, namely satisfied, disappointed, happy and sad. This study uses two methods that are compared, namely the Naïve Bayes Classifier method and the Support Vector Machine. In different domain, a study [10] that discusses cosmetic product review comments to analyze positive and negative attitudes about various cosmetic products with sentiment analysis. The analytical method uses a machine learning technique called the Naïve Bayes Classifier to classify comments as positive or negative.

As the studies mentioned above, opinion mining with sentiment analysis is more useful for analyzing consumer opinions or attitudes towards products. This study focuses on comments about skin care products from Twitter by selecting text containing “#skincare”, then modeling the topic and categorizing these tweets according to sentiment polarity. In contrast to previous studies, this research uses Latent Dirichlet Allocation (LDA) to model the topic before it carried the sentiment analysis process out. Topic modeling is done to find out what topics are discussed in tweets related to skincare or facial care. Several studies that use LDA for topic modeling include research related to the discovery of the topic of Indonesia's infrastructure development in online news [10]. In addition, research that discusses modeling the topic of healthy life society movement (Germas) on Instagram data [11].

METHOD

The research stages used can be seen in Figure 1. We used the data in this study via the Twitter streaming API between January 1 and June 6, 2022. During that time, we managed to collect 4,750 tweets related to skincare using the keyword "face care. Twitter streaming API provides an interface to obtain a complete set of tweet attributes. However, in this experiment, only a few attributes are used. Table I describe the tweet attributes used in this study.

TABLE I. ATTRIBUTES OF TWEET DATA

Tweet Attributes	Description
Tweet ID	unique twitter user identity
Screen Name	username of the twitter account.
Tweet Text	a post on the social media site Twitter
Timestamp	a sequence of characters or encoded information identifying when a certain event occurred
Retweet	a re-posting of a Tweet
Likes	used to show appreciation for a Tweet

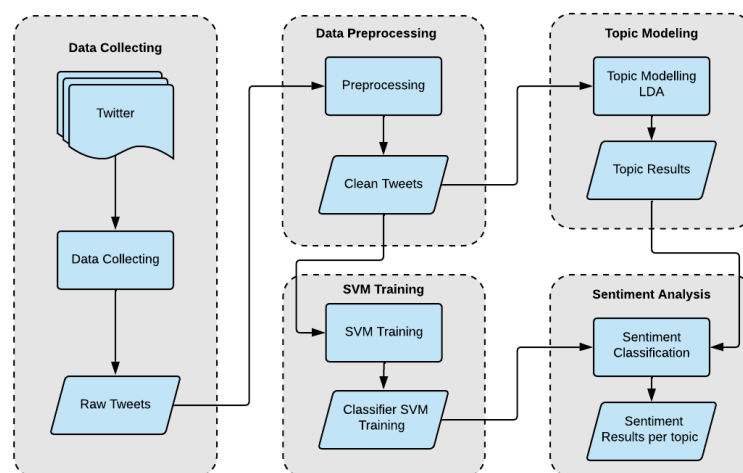


Figure 1. Research Stages

After data acquisition, the next step is data preprocessing. Generally, raw tweets contain lots of noise, misspelled words, and meaningless words, including various abbreviations and slang words. These words often interfere with the resulting tweet sentiment and degrade the performance of the classification model. Therefore, tweets must be pre-processed before actually extracting features from them. We perform preprocessing steps on the tweet data to be processed; these steps are:

- Tokenization: the process of dividing a text into certain parts.
- Normalize: Brings text to its standard form. The common normalization techniques used are as follows:
 - Case folding: Change capital letters to lowercase
 - Elimination of periods in terms.
 - Remove hyphens in terms.
- Cleaning: The steps for cleaning tweet data are as follows:
 - Delete the URL contained in the tweet
 - Remove @ sign-in username
 - Delete existing hashtags (#) in tweets
 - Delete the number contained in the tweet
 - Remove punctuation marks, such as question marks, exclamation marks, periods, and others.
 - Remove Unicode and symbols
- Stopword Removal: remove words that are considered meaningless using a Stop word list.

Topic Modelling

Topic modeling is a statistical-based process to find a set of topics in a particular set of text documents. Themes provide an intuitive syntactic representation of the ideas shared in text documents [12]. Therefore, we can divide the document into several categories according to the similarity of the topic model. In text form, a set of keywords that best describe a particular part of the text document represents the topic model. Topic modeling has received a lot of attention in the natural language and machine learning communities in recent years [13]. The main reason for using theme modeling in this experiment is its flexibility and reliability in revealing hidden themes in numerous text documents. Topic modeling is a powerful approach to finding statistical correlations between words in a text corpus to form syntactically relevant topics represented by word sets. The main purpose of topic modeling is to find word usage throughout the document and relate it to other words in various document segments [14].

The basic idea of topic modeling is to find a collection of words that often occur together in a given area of different segments of the document corpus according to statistical measures. From the perspective of the topic model, the document corpus consists of many topics, each consisting of a unique set of words or keywords [15]. In addition, for each word in each topic, there is a probability distribution of how certain keywords contribute to constructing the semantic meaning of a particular topic [14]. Figure 2 is a well-known example showing the topology of the topic model of a text

document produced by Blei [16]. From Figure 2, we can see that each topic contains a collection of words with a certain probability. This shows how important the words are in the topic. Each document then contains a set of terms that represent a set of topics.

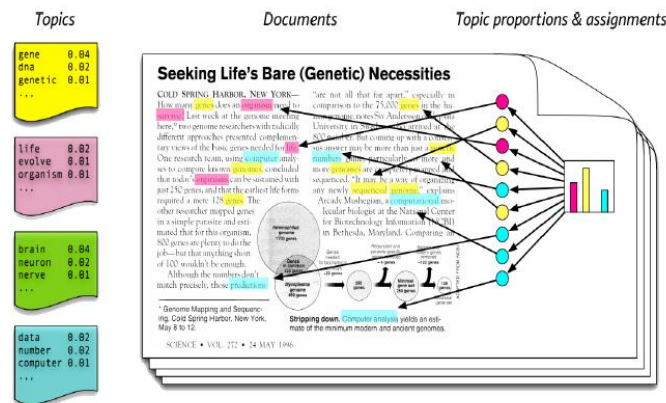


Figure 2. A general illustration of the topic model [12]

Several algorithms apply the topic modeling approach. One of the most widely used algorithms in topic modeling and used in this experiment is Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model for discrete data. Therefore, this algorithm is very suitable for text data [12]. The visualization of LDA can be seen in Figure 3.

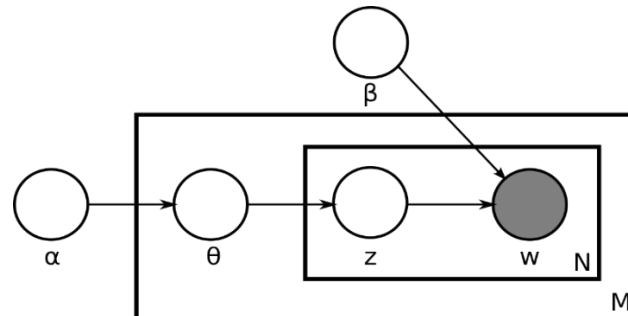


Figure 3. Visualization of the topic modeling using the LDA method [17]

LDA assumes the following generative process for each corpus D document [18]:

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of the N words w_n :
 - a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
 - b) Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n

From a machine learning perspective, we classified LDA as an unsupervised learning approach. This is because it automatically classifies unlabeled documents according to the similarity of the probability distribution of their terms [19]. Topic modeling is a powerful statistical-based approach for deriving different themes that occur in a text document set. However, when it comes to modeling, many criteria need to be considered. The first thing is the number of topics (k) should be determined at the beginning [20]. That could be a serious problem if there is no prior knowledge of the related documents. Hence, in this experiment, we iteratively generate various k -number of topics and evaluate each experiment to determine the best number of k . Then, for evaluating the best number of k -topic, we employed a metric named topic coherence. The coherence metric calculates how one term and another within a topic are statistically correlated with each other. The higher correlation score is indicating higher coherence and pointing to a better quality of the generated topic.

Sentiment Analysis

Sentiment analysis or opinion mining is a field of research that analyzes opinions, feelings, evaluations, judgments, attitudes, and people's sentiments towards product units, organizational services, individuals, issues, events, issues, and attributes [21]. Sentiment analysis often referred to as subjective analysis, opinion generation, judgment extraction, etc., has several relationships related to emotional computing, such as computer recognition and emotional expression [22].

A support Vector Machine (SVM) is a non-probabilistic binary linear classifier. For a training set of points (x_i, y_i) , x is a feature vector, and y is the class. To determine the maximum margin hyperplane that divides the points with $x_i = -1$ and $x_i = 1$. The equation of the hyperplane is; $w \cdot x + b = 0$. For a data set consisting of features set and labels set, an SVM classifier builds a model to predict the classes for the new examples. It assigns a new case or data points to one of the categories [23].

Steps:

- Determine the optimal hyperplane
- Extend step 1 to problems that cannot be separated linearly
- Mapping data into easily classifiable high-dimensional spaces with linear decision surfaces.

DISCUSSION

Topic Analysis

The topic modeling in our experiment consists of two parts. The first part is building the LDA model, which comprises a statistically larger number of topics. Then by using lexical similarity, we then group topics that have the same words/terms. For the first stage of topic modeling, we determine the best number of k-topics for our LDA model. For this purpose, we iteratively compute the topic coherence of various k-topic counts. With Figure 4 depicts coherence scores in various topic starting numbers. In this experiment, we try to generate 1 to 30 topics from our corpus. From Figure 4 it can be seen that the highest coherence score is found in 14 topics with a value of 0.403. Thus, we can determine that the best k are 14 topics. We then focus on this number for further analysis. What themes are covered cannot be determined from the 14 topics retrieved. Because the topic involves a collection of words and will be assessed in the following step based on the rationale of human analysis.

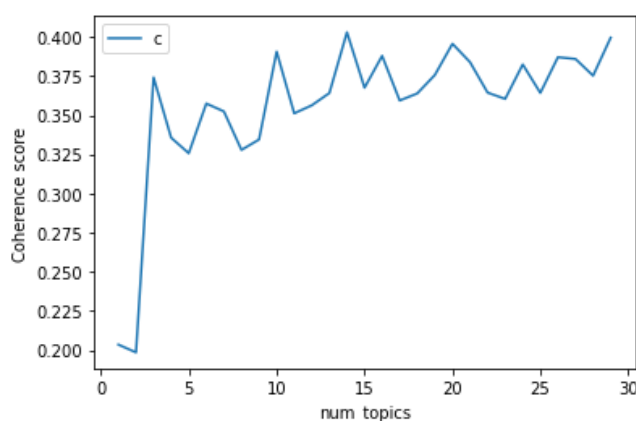


Figure 4. Coherence score of each number of topics

To better understand the structure of the words that make up each topic in the topic model, we then visualized a distance map of the 14 topics. This visualization provides insight into how topics relate to each other based on the terms/words they share. Figure 5 shows the distance map of 14 topics generated by the LDA algorithm on the document corpus used in this experiment. In Figure 5, we can see that it distanced some topics from each other, and some other topics overlap with other topics. Topic overlap indicates that there are similar words that build on different topics. The 14 k-topics with visualization shown in Figure 5 are the results of the first two stages of topic modeling used in this study.

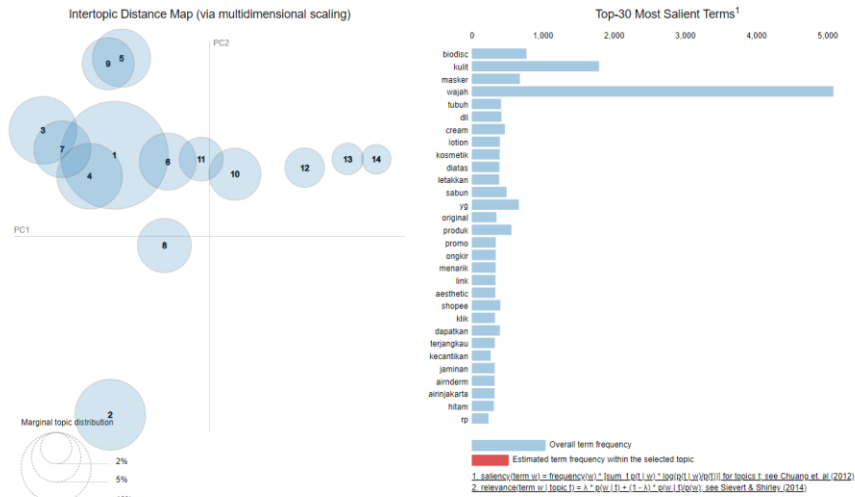


Figure 5. Term clusters visualization resulted by LDA

We have described the first stage of topic modeling in the previous paragraph, which is to determine the best number of k-topics for the LDA model. The second stage of topic modeling is to combine overlapping topics into one larger topic. The basic idea of topic representation in LDA is a topic consisting of a set of words. Therefore, overlapping topics will generally share a few words in their representation. Instead of manually selecting overlapping topics based on visualizations, we use binary cosine similarity to define aggregate topics. Thus, topics that share the most keywords in their topic representation will have a high similarity score and are considered similar topics [24]. Based on Figure 4, we can see that topics that have an enormous distance between topics mean that these topics have no relationship. However, topics that are close or even overlap between topics, means that the topic has a related relationship or may have a similar topic theme. Table II summarizes the topic categories of the aggregation process as a result of the second phase topic modeling in this study. The results of categorizing the topics into five category are based on Figure 5 of the intertopic distance map. Overlapping subject groups are groups of topics with nearly identical language that can be categorized into the same topic.

TABLE II. LIST OF TOPIC IDENTIFIED FROM LDA TERM CLUSTERS.

Topic Category	Associated LDA Term Cluster	Topic Description	Number of Tweet
Category 1	1, 3,4,6,7,10,11	The importance of skin care	2696
Category 2	2	Promo purchase skin care	71
Category 3	12,13,14	Skin care with fruit and spice extracts	809
Category 4	5,9	Skin care with exercise	774
Category 5	8	Skin care to the doctor's clinic	397

Table II shows that the topic aggregation process in the second stage of the topic modeling work in this study produced six topic categories. Each generated topic consists of several groups of words that have different occurrence weights. We consider topics that have the same occurrence of words to have the same topic category, we can see what words are most significant in building topic categories. This method helps distinguish topics that can be interpreted semantically and topics that are the result of human interpretation [25]. Based on the similarity of word occurrences on 14 topics, five topic categories were obtained, as shown in Table II column of topic description. The topic distance map illustrates proximity can group how many topics. Of the six topic categories, two of them have a dominant share with more than one hundred thousand related tweets. The other two contributed a fraction, with a few related tweets under thirty thousand. Figure 4 shows the overall proportions of all the topic categories.

Based on the proportions illustrated in Figure 6, category 1 topics are the highest-level issues discussed by the community. Table II shows that category 5 topics contain around 2,696 tweets, most of which talk about the importance of skincare. The second biggest issue discussed by

Indonesians on Twitter is related to skincare with fruit and spice extracts (topic category 3) with a total of 809 tweets or 17% of the total tweets received.

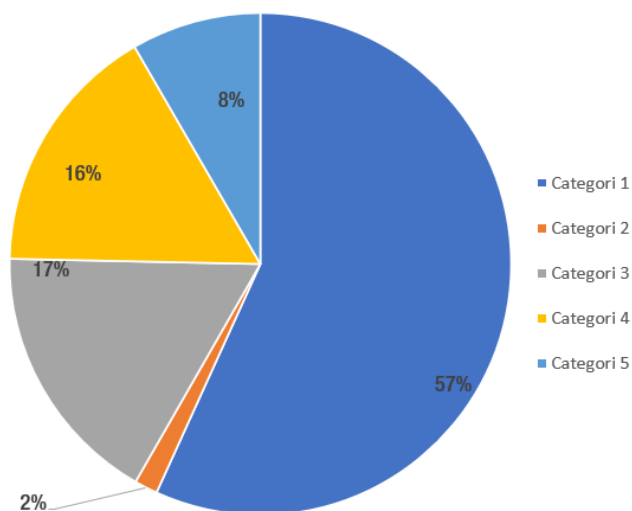


Figure 6. Proportion of tweets associated with each topic

Sentiment Analysis

After discovering the topics contained in the Covid-19 vaccination case, the next step is to conduct a sentiment analysis of each topic. Sentiment analysis is carried out by classifying sentiment into two classes, namely positive and negative classes. So that on each topic it will be known how many people have a positive or negative perspective. We carried sentiment analysis out using the Support Vector Machine (SVM) algorithm, which has been carried out by the training process with an accuracy of 81.73%. From the sentiment analysis process, the results of the sentiment class per topic were obtained, as shown in Figure 7.

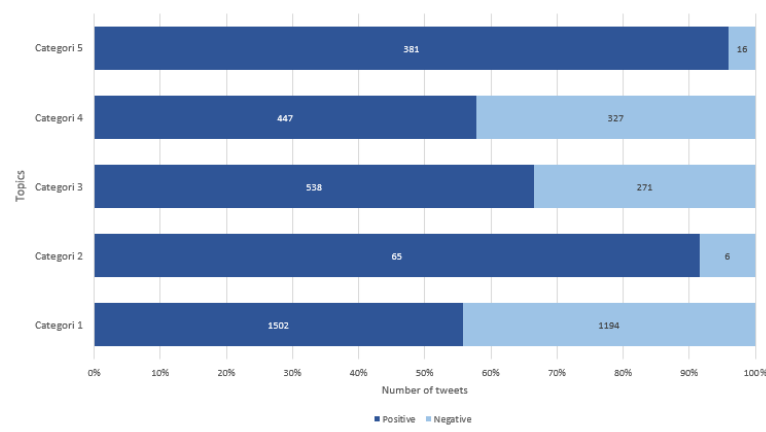


Figure 7. Sentiment proportion of each topic.

Figure 7 is a graph of the results of topical sentiment analysis. Based on the results of the per-topic sentiment analysis, it can be seen the number of positive or negative sentiment class tweets on each topic. In all topic categories, the distribution of the number of positive sentiment class tweets dominates over negative sentiment. This is because, at this time, public awareness about carrying out skin care has started early [26], [27]. Public awareness of the importance of taking care of facial skin is increasing along with the number of skincare products on the market [28]. In addition, skin care product ingredients derived from nature have become a topic of customer discussion because we consider them safer than the use of chemicals [29]. The use of abundant natural fruits and spices in Indonesia makes Indonesian facial care products no less competitive than skincare products from abroad [30].

CONCLUSION

In this study, we analyzed customer experience related to skincare products. To carry out this perception analysis, this study conducted two stages of analysis, namely topic modeling and sentiment analysis. The results of topic modeling obtained 14 topics of conversation related to skin care. Meanwhile, the sentiment analysis results of 14 topics resulted in more positive sentiment class tweets overall. We relate the category topic that has the number of tweets to the importance of skincare. In addition, categories related to skincare product ingredients from nature, namely fruits and spices, are the topics that have the second highest number of tweets. The results of the analysis of tweets related to user experience on Twitter, it was found that users prefer skin care products that use ingredients from nature.

REFERENCES

- [1] Y. Lin, "10 Twitter Statistics Every Marketer Should Know in 2021 [Infographic]," *Oberlo*, 25-Jan-2021. [Online]. Available: <https://www.oberlo.com/blog/twitter-statistics>. [Accessed: 17-Mar-2022].
- [2] X. Fang and J. Zhan, "Sentiment analysis using product review data," *J. Big Data*, vol. 2, no. 1, pp. 1–14, Dec. 2015.
- [3] H. Isah, P. Trundle, and D. Neagu, "Social media analysis for product safety using text mining and sentiment analysis," *2014 14th UK Work. Comput. Intell. UKCI 2014 - Proc.*, Oct. 2014.
- [4] B. Liu, "Sentiment Analysis and Opinion Mining," <http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016>, vol. 5, no. 1, pp. 1–184, May 2012.
- [5] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, *A Practical Guide to Sentiment Analysis*, 5th ed. Switzerland: Springer, 2017.
- [6] P. Pugsee, T. Chongvisuit, and K. N. Nakorn, "Opinion mining on Twitter data for airline services," *2015 5th Int. Work. Comput. Sci. Eng. Inf. Process. Control Eng. WCSE 2015-IPCE*, pp. 639–644, 2015.
- [7] T. Chatchaithanawat and P. Pugsee, "A framework for laptop review analysis," *ICAICTA 2015 - 2015 Int. Conf. Adv. Informatics Concepts, Theory Appl.*, Nov. 2015.
- [8] P. Pugsee and M. Niyomvanich, "Sentiment Analysis of Food Recipe Comments," *ECTI Trans. Comput. Inf. Technol.*, vol. 9, no. 2, pp. 182–193, Jan. 2015.
- [9] S. Rahmawati and M. Habibi, "Public Sentiments Analysis about Indonesian Social Insurance Administration Organization on Twitter," *IJID (International J. Informatics Dev.)*, vol. 9, no. 2, pp. 87–93, Dec. 2020.
- [10] S. Zuluaga *et al.*, "Indonesia Infrastructure Development Topic Discovery on Online News with Latent Dirichlet Allocation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1077, no. 1, p. 012012, Feb. 2021.
- [11] M. Habibi, A. Priadana, A. B. Saputra, and P. W. Cahyo, "Topic Modelling of Germas Related Content on Instagram Using Latent Dirichlet Allocation (LDA)," pp. 260–264, Jan. 2021.
- [12] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.
- [13] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, p. 101582, Dec. 2020.
- [14] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Mach. Learn.*, vol. 95, no. 3, pp. 423–469, Oct. 2014.
- [15] Y. Zuo *et al.*, "Topic modeling of short texts: A pseudo-document view," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, vol. 13-17-Aug, pp. 2105–2114.
- [16] D. M. Blei, A. Y. Ng, and M. T. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [17] M. Habibi, M. R. Ma'arif, and D. Subekti, "The Development of Social Media Intelligence System for Citizen Opinion and Perception Analysis over Government Policy," *Telemat. J. Inform. dan Teknol. Inf.*, vol. 19, no. 1, pp. 31–46, Jul. 2022.
- [18] P. B. Rahman and M. Habibi, "Topic Analysis for Coronavirus Disease (COVID-19) on Twitter Using Latent Dirichlet Allocation (LDA)," in *The 1st Universitas Muhammadiyah Yogyakarta Undergraduate Conference (UMYGRACE) 2020*, 2020, vol. 2020, pp. 1037–1043.
- [19] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019.

-
- [20] D. Newman, H. Jey, Lau, K. Grieser, and T. Baldwin, "Automatic Evaluation of Topic Coherence," 2010.
- [21] B. Liu, *Sentiment Analysis and Opinion Mining*, no. May. Morgan & Claypool Publishers, 2012.
- [22] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1, 2008.
- [23] S. G. Kanakaraddi, A. K. Chikaraddi, K. C. Gull, and P. S. Hiremath, "Comparison Study of Sentiment Analysis of Tweets using Various Machine Learning Algorithms," in *Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT 2020*, 2020, pp. 287–292.
- [24] M. Habibi, M. R. Ma'arif, and D. Subekti, "The Development of Social Media Intelligence System for Citizen Opinion and Perception Analysis over Government Policy," *Telemat. J. Inform. dan Teknol. Inf.*, vol. 19, no. 1, pp. 31–46, Feb. 2022.
- [25] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, 2012, no. July, pp. 952–961.
- [26] Y. Lisnawati, "Tingkatkan Kesadaran Pria Pentingnya Kesehatan Kulit Wajah - Citizen6 Liputan6.com," *Liputan6*, 14-Apr-2022. [Online]. Available: <https://www.liputan6.com/citizen6/read/4714931/tingkatkan-kesadaran-pria-pentingnya-kesehatan-kulit-wajah>. [Accessed: 01-Nov-2022].
- [27] G. Perkasa, "Kesadaran Merawat Kulit Dimulai sejak Usia Muda Halaman all - Kompas.com," *Kompas Lifestyle*, 11-May-2021. [Online]. Available: <https://lifestyle.kompas.com/read/2021/05/11/155659820/kesadaran-merawat-kulit-dimulai-sejak-usia-muda?page=all>. [Accessed: 01-Nov-2022].
- [28] S. D. Kusumaningrum, "KAJIAN PUSTAKA DALAM PENENTUAN TIPE DAN PERMASALAHAN KULIT WAJAH," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 1, no. 1, Aug. 2021.
- [29] U. Hidayatty Luthfia, "5 Buah yang Bisa Dijadikan Skincare Alami, Ampuh Bikin Wajah Flawless!," *idntimes*, 06-Sep-2019. [Online]. Available: <https://www.idntimes.com/life/women/ulfa-luthfia-hidayatty/buah-yang-bisa-dijadikan-skincare-alami>. [Accessed: 01-Nov-2022].
- [30] L. Prebreza, "5 Bahan Dapur yang Bisa Dijadikan Skincare Alami agar Wajah Sehat Halaman all - Kompas.com," *Kompas*, 17-May-2022. [Online]. Available: <https://www.kompas.com/tren/read/2022/05/17/100000265/5-bahan-dapur-yang-bisa-dijadikan-skincare-alami-agar-wajah-sehat?page=all>. [Accessed: 01-Nov-2022].