Jurnal Ilmu Komputer Agri-Informatika

Tersedia secara daring di:
https://jurnal.ipb.ac.id/index.php/jika

Volume 12 Nomor 1 halaman 119–131
eISSN: 2654-9735, pISSN: 2089-6026

Topic Modelling on Beauty Product Reviews Using Latent Dirichlet Allocation

ADE SARAH HUZAIFAH[1]*, ROSSY NURHASANAH[1], R. A. FATTAH ADRIANSYAH[2]

Abstract

In contemporary society, beauty products have become essential, particularly for women. With their growing popularity, online review platforms now provide extensive information on product trends, customer satisfaction, and performance. However, the sheer volume of available reviews presents challenges in drawing meaningful conclusions. To address this, topic modeling techniques such as Latent Dirichlet Allocation (LDA) have been widely employed in text mining and information retrieval. LDA is a probabilistic model capable of uncovering latent structures within textual data and identifying similarities across documents. Recent studies suggest that topic modeling of product reviews in the cosmetics industry can yield valuable insights into consumer perceptions and product attributes. This study aims to identify thematic patterns in customer reviews of ten facial cleanser brands sourced from the Female Daily website. The research methodology consists of five main stages: data collection, preprocessing, topic modeling using LDA, visualization, and topic interpretation. The results reveal that Topic 2, which highlights preferred product advantages, is the most frequently discussed, accounting for 48.5% of the total reviews. Topic 1, which focuses on the effects of products on acne-prone skin, constitutes 38%, while Topic 3, emphasizing products with natural ingredients, makes up 13.5% of the reviews. These findings can assist businesses in developing products that align more closely with consumer preferences. Moreover, they support prospective buyers in making informed purchasing decisions by enhancing their understanding of product attributes based on user experiences.

Keywords: topic modelling, Latent Dirichlet Allocation, beauty product, product review, Female Daily

INTRODUCTION

Beauty products have become a necessity in society, especially for women. They are usually manufactured to improve the appearance of women. Therefore, this great opportunity certainly makes beauty industry players compete to produce various beauty products to meet market demand. The rapid growth of beauty products has also led to independent review sites that provide product-related information, introduce the latest brand or brand of a product, or can be accessed by users themselves to provide reviews of the products purchased. These online reviews are an important resource for both customers and manufacturers. They offer perspectives on product performance, customer happiness, and developing trends. However, it is difficult to derive insights from so many reviews. To address this, topic modeling approaches, such as Latent Dirichlet Allocation (LDA) have been employed to assess and categorize enormous volumes of textual data.

LDA is a generative probabilistic model that describes a set of observations using unseen clusters, which explains why certain parts of the data are similar. In the field of text mining and information retrieval, LDA is a powerful tool for modeling topics (Jelodar *et al.* 2019). In topic tracking systems, it has shown better performance than alternative models, such as vector space and unigram language models, by representing topics as polynomial distributions of features (Xiao-bo L 2011). Despite the difficulties caused by noise and a variety of material forms, LDA is very helpful in discovering latent topic frameworks in web pages (Phanidhar *et al.* 2023).

[1]Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan 20155;
[2]Department of Informatics Engineering, Faculty of Informatics, Universitas Mikroskil, Medan 20212;
*Corresponding Author: Email: adesarah@usu.ac.id;

The technique builds a language modeling framework for information retrieval by combining term models and probability mixture models (Gagan and Chirag 2022). Despite certain drawbacks, such as the need for accurate representation of single words and ambiguity due to common words appearing in different topics, LDA is still a popular option for academics in a variety of fields, including software engineering, social networks, and linguistics (Jelodar *et al.* 2019; Gagan and Chirag 2022).

The use of LDA in topic modeling has been extensively applied in many fields, such as online product reviews. LDA provides a more accurate identification of current customer satisfaction dimensions, such as location and service quality, in South Korea's accommodation industries while validating these dimensions. Topics related to points of competitiveness (e.g. value, staff professionalism) and points of uniqueness (e.g. ambient temperature, neighborhood attractions) were more prominent in the online reviews. The extracted topics represent areas that customers find important and can guide industry practitioners, but the valence of the reviews and individual customer characteristics should be considered (Sutherland *et al.* 2020).

LDA successfully analyzed 524 YouTube comments about the Lumpia Gang Lombok Semarang. This is one of the Micro, Small, and Medium Enterprises (MSMEs) in Indonesia. Comment analysis provides valuable insights into user preferences and perceptions of products, supporting MSMEs in understanding customer satisfaction and enhancing value for those enterprises. These insights, which provide a deeper knowledge of social media data, particularly on the YouTube platform, assist in the development and marketing of MSMEs (Alpiana *et al*. 2024).

LDA analyzed social media comments about the Infinix Note 30 smartphone, identifying 3 key topics: price and availability timeline, difficulty in obtaining the product due to its absence in the market, and post-usage ownership and aspects of the product that customers appreciate. These findings provide marketers with insight into consumer issues and how to improve the accessibility of their products (Alzami *et al.* 2023).

An LDA-based comparative study of consumer preferences for two competing items is presented. It is effective in discovering the main points of differentiation between the items, providing a more thorough overview of their advantages over competitors and potential areas for development. However, online product reviews have provided a good and reliable channel for not only understanding customer needs for one product or service but also analyzing product competition in the market. This provides valuable insights for product designers and e-commerce companies (Wang *et al.* 2018).

Product designers can use online reviews for data mining and decision-making to develop a cordless hairdryer design index system based on mapping the identified topics using LDA to user needs. The design index system was validated through a survey, which found that a product designed based on the index system had significantly higher consumer satisfaction, purchase intention, usage habits, and intention to continue using compared to a less compliant product (Miao *et al.* 2023).

Recent studies have also explored topic modelling for product reviews in the beauty industry. LDA has been used to identify key topics in online news about Somethinc brands (Puspita *et al.* 2024), beauty product reviews about Laneige Water Sleeping Mask (Salsabila and Trianasari 2021) and Lipstick products (Wang and Shao 2023). These studies have also shown that topic classification using LDA is a useful technique. Research on topic modeling has consistently demonstrated its ability to extract insightful information about consumer perceptions and product attributes. The efficacy of this approach in determining areas for product improvement, comprehending user satisfaction, and directing product development has been extensively established. However, the extant literature on this subject has focused primarily specific beauty products without broader comparisons across multiple brands.

The present study aims to analyze topics in consumer reviews of 10 brands of facial cleansing products from a Female Daily website, in order to understand what user considered when choosing a product. Contrary to the approach of preceding research, the present study

seeks to identify overarching thematic patterns in the discourse of consumers across diverse brands. By identifying the key attributes that consumers consider when selecting facial cleansing products, the findings can provide deeper insights into consumer decision-making factors and industry trends. The findings of this study can assist other users in making informed decisions regarding the purchase of beauty products, as each type of beauty product offers unique benefits.

METHOD

The workflow of this research is shown in Figure 1, where the stages are data collection, preprocessing, topic modelling with LDA, visualization and interpretation of topics.

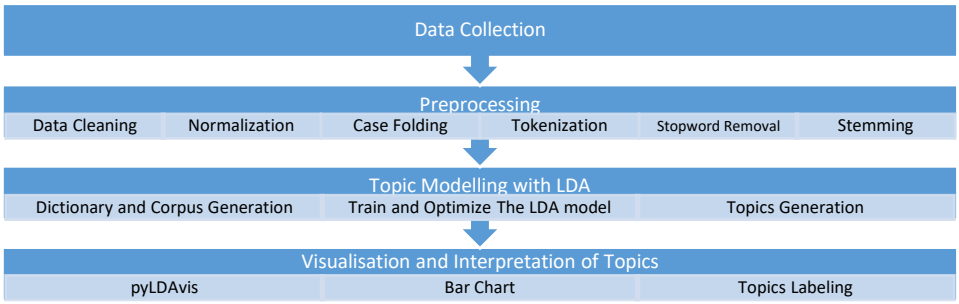

Figure 1 Proposed method

Data Collection

The data used in this study is review data for 10 brands of beauty products with the facial wash category where the product data is taken from the data scrapping process sourced from the Female Daily website using Web Scrapper, which is an application for scrapping data from the website and then the data that has passed the scrapping process will be saved into CSV form. The reviews collected are Indonesian-language reviews with a total of 3951 reviews. The dataset file contains the Product Name, Product Type, and Review.

Preprocessing

Transforming unprocessed data into understandable data is the purpose of data preprocessing. Text preprocessing is an essential stage in Natural Language Processing (NLP). It has a major impact on how well various applications such as virtual assistants, web search, and text classification work (Tabassum and Patil 2020; Rajesh and Hiwarkar 2023). To produce reliable results, preprocessing procedures such as tokenization, stop word removal, and lemmatization must be properly chosen and arranged (Tabassum and Patil 2020; Al-Badrashiny *et al.* 2016). Good preprocessing reduces noise and increases efficiency by cleaning and filtering textual data (Rajesh and Hiwarkar 2023). In general, effective preprocessing methods are necessary to improve feature extraction and information retrieval in NLP applications (Tabassum and Patil 2020; Sharma and Parmar 2021). Preprocessing in this study includes a number of steps, including data cleaning, normalization, case folding, tokenization, stopword removal, and stemming.

1. Data Cleaning
   Data cleaning is a process of cleaning data by selecting data, eliminating inaccurate data, smoothing noisy data such as removing special characters and cleaning text from excessive spaces and punctuation marks. The purpose of this data cleaning is to eliminate errors in the data because clean and accurate data will produce higher-quality analyses and models.
2. Normalization
   Social media is often a place where non-standard language and slang words flourish. Social media users, especially the younger generation, often use slang to communicate more casually and expressively. For example, the word ‘ngegas’, which means angry or emotional, is often used on social media to describe someone who is angry. Although not in official dictionaries, social media users often use this word. The normalization of texts

that contain slang and non-standard language is crucial in the context of NLP in order to ensure a consistent and trustworthy analysis of the data. Strategies such as dictionary-based normalization can be used to replace slang words with more mainstream terms.

3. Case Folding
   Case folding is the process of converting all letters to lowercase. The main purpose of case folding in NLP is to reduce variation by removing words that are the same but written in both uppercase and lowercase. For example, ‘Kucing’ and ‘kucing’ would be considered the same. This can help simplify the word matching process so as to reduce data complexity while improving analysis accuracy.
4. Tokenization
   Tokenization is the process of breaking down text in the form of sentences into smaller parts, or tokens. An example of tokenization is the transformation of "saya pergi ke mall" into "saya", "pergi", "ke" and "mall". Each token can then be examined separately. For example, unnecessary phrases that do not provide important information for the model can be eliminated. After tokenization, the text can be transformed into a numerical representation. This is required by topic modeling methods. Tokenization also allows the model to recognize patterns and relationships between words in a document. This makes it easier to classify terms that often appear together into the same topic.
5. Stopword Removal
   Stopword removal is the process of removing common words that occur frequently in text but provide little information for analysis. These words are known as stopwords. They include words such as 'dan', 'atau', 'yang', 'di', and 'ke'. Stopword removal is an important preprocessing technique in text analysis that has a significant impact on various applications. It can decrease the size of the data set by 35 to 45%, thus improving the efficiency and accuracy of the text mining task (Ladani and Desai 2020). Stopword removal plays an important role in improving system performance in various text-processing applications.
6. Stemming
   Stemming is the process of converting words to their base form by removing suffixes or affixes. For example, words like ‘berlari’, ‘berlari-lari’, and ‘berlarian’ can be changed to ‘lari’. Stemming, the process of reducing words to their root form, is crucial in various Natural Language Processing (NLP) applications and Information Retrieval (IR) systems (Jivani 2011)(Patil et al. 2016). It plays a significant role in reducing inflectional forms and sometimes derivationally related forms of words to a common base form, thereby decreasing indexing dimensions (Jivani 2011)(Swain and Nayak 2018).

Topic Modelling with LDA

Topic modeling is a powerful technique for the analysis of a huge collection of a document. Topic modeling is used for discovering hidden structures from the collection of a document. The topic is viewed as a recurring pattern of co-occurring words. A topic includes a group of words that often occurs together. Topic modeling can link words with the same context and differentiate across the uses of words with different meanings (Barde and Bainwad 2017). Various methods have been developed, including Vector Space Model, Latent Semantic Indexing, and Probabilistic Latent Semantic Analysis, with LDA being the most popular (Sharma and Sharma 2017).

The algorithm for probabilistic topic modeling, known as LDA, is employed to identify topics within extensive sets of documents. It is assumed that documents have several topics, each of which is a probability distribution over a word count (Chundi and Go 2015). Without the need for pre-labeled data, the algorithm seeks to infer word distributions for topics and topic distributions for documents, exposing underlying structures (Chundi and Go 2015; Baranowski 2022). Figure 2 shows how LDA does topic modeling.

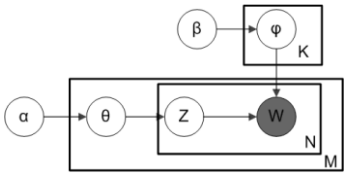

Figure 2 Graphical representation of the LDA model (Ameko *et al.* 2021)

The variable names are defined as follows:
$\boldsymbol{M}$ is the number of documents
$\boldsymbol{N}$ is the number of words in a given document
$\boldsymbol{K}$ is the number of topics
$\boldsymbol{\alpha}$ is the parameter of the Dirichlet prior to the per-document topic distributions
$\boldsymbol{\beta}$ is the parameter of the Dirichlet prior to the per-topic word distribution
$\boldsymbol{\theta_i}$ is the topic distribution for document *i*
$\boldsymbol{\varphi_k}$ is the word distribution for topic *k*
$\boldsymbol{Z_{ij}}$ is the topic for the *j*-th word in document *i*
$\boldsymbol{W_{ij}}$ is the specific word

We use the LDA algorithm calculation in (Jelodar *et al.* 2019) implemented with Python version 3.10.12, library Gensim version 4.3.3, and library pyLDAvis 3.4.1 where the parameters are shown in Table 1.

Table 1 Parameters

Parameter Name	Value
α	auto
β	auto
Passes	15
Start	2
Limit	40
Step	1
Coherence Score	C_v

Visualization and Interpretation of Topics

This research uses LDAvis, an interactive visualization tool designed for LDA topic modeling, to illustrate the relationships between topics. The visualization was created using the pyLDAvis Python module (Sievert and Shirley 2014), which makes it easier to explore topic distribution and keyword relevance.

The implementation process begins with data preprocessing, during which the reviews are tokenized and cleaned to ensure the extraction of meaningful topics. After applying LDA, the resulting topics are visualized as a set of circles, where the size of each circle represents its proportion of the total topics in the dataset. The inter-topic distance map represents subject relationships in a simple way, making it easier to identify closely connected themes.

Furthermore, LDAvis offers a ranked distribution of words, emphasizing those that are most representative of each topic based on their probability and relevance. The parameter λ in LDAvis enables users to adjust the balance between frequently occurring words and words that are unique to each topic, thus allowing for a more detailed examination of the topic structure. This HTML-based, interactive visualization takes a dynamic and flexible approach to exploring topics, supporting comprehensive interpretation at a later stage of analysis.

RESULT AND DISCUSSION

This study took review data from the Female Daily website from January 2022 to March 2023. Table 2 shows the number of reviews of each facial wash product brand used as a dataset.

Table 2 Total review data of ten products

Product Name	The Number of Reviews
Acnes Creamy Wash	394
Cetaphil Gentle Skin Cleanser	393
Corsx Low pH Good Morning Gel Cleanser	396
Garnier Bright Complete Brightening Face Wash Foam	397
Pond's Acne Solution AntiAcne Facial Foam	395
Senka Perfect Whip Facial Foam	391
Simple Refreshing Facial Wash Gel	395
Somethinc Low pH Gentle Jelly Cleanser	398
The Body Shop Tea Tree Skin Clearing Facial Wash	397
Wardah Lightening Micellar Gentle Wash	395

The dataset was then subjected to a preprocessing stage to ensure the data used in the analysis are in good quality and reliable. Table 3 shows the changes in the review data before and after the preprocessing stage.

Table 3 Result of data preprocessing

Review Data	Cleaning Data	Case Folding	Stemming	Normalization	Stopword Removal	Tokenization
FW KECINTAAN AKU ❤ dia bikin muka seger, cerahan beneran tapi ngga bikin ketarik atau kering di mukaku. Apalagi kalo pulang kuliah kan kena polusi dan berminyak bgt begitu cuci muka pake ini lsg bener2 seger. Aku gatau sih krn skincare lain atau bukan tapi ini cukup mencerahkan mukaku dan kalo abis cucimuka pake ini rasanya bersiih aja gt. Cuma dia ga selalu available jd kadang make yg biru biasa tp ttp ga senampol yg ini pokonya luv bgtt ❤	FW KECINTAAN AKU dia bikin muka seger cerahan beneran tapi ngga bikin ketarik atau kering di mukaku Apalagi kalo pulang kuliah kan kena polusi dan berminyak bgt begitu cuci muka pake ini lsg bener seger Aku gatau sih krn skincare lain atau bukan tapi ini cukup mencerahkan mukaku dan kalo abis cucimuka pake ini rasanya bersiih aja gt Cuma dia ga selalu available jd kadang make yg biru biasa tp ttp ga senampol yg ini pokonya luv bgtt	fw kecintaan aku dia bikin muka seger cerahan beneran tapi ngga bikin ketarik atau kering di mukaku apalagi kalo pulang kuliah kan kena polusi dan berminyak bgt begitu cuci muka pake ini lsg bener seger aku gatau sih krn skincare lain atau bukan tapi ini cukup mencerahkan mukaku dan kalo abis cucimuka pake ini rasanya bersiih aja gt cuma dia ga selalu available jd kadang make yg biru biasa tp ttp ga senampol yg ini pokonya luv bgtt	fw cinta aku dia bikin muka seger cerahan bener tapi ngga bikin tarik atau kering di mukaku apalagi kalo pulang kuliah kan kena polusi dan minyak bgt begitu cuci muka pake ini lsg bener seger aku gatau sih krn skincare lain atau bukan tapi ini cukup cerah mukaku dan kalo abis cucimuka pake ini rasanya bersiih aja gt cuma dia ga selalu available jd kadang make yg biru biasa tp ttp ga senampol yg ini pokonya luv bgtt	facial wash cinta aku dia bikin muka seger cerah benar tapi tidak bikin Tarik atau kering di muka apalagi kalau pulang kuliah kan kena polusi dan minyak banget begitu cuci muka pakai ini langsung bener seger aku gatau sih karena skincare lain atau bukan tapi ini cukup cerah mukaku dan kalo habis cuci muka pakai ini rasanya bersih aja gitu cuma dia tidak selalu available jadi kadang make yg biru biasa tapi tetap tidak senampak yang ini pokok love banget	cinta muka segar cerah benar tidak tarik kering muka pulang kuliah kena polusi minyak cuci muka pakai langsung bener segar tidak skincare bukan cerah muka habis cuci muka pakai bersih tidak available kadang make biru tidak senampak pokok love	['cinta', 'muka', 'segar', 'cerah', 'benar', 'tidak', 'tarik', 'kering', 'muka', 'pulang', 'kuliah', 'kena', 'polusi', 'minyak', 'cuci', 'muka', 'pakai', 'langsung', 'bener', 'segar', 'tidak', 'skincare', 'bukan', 'cerah', 'muka', 'habis', 'cuci', 'muka', 'pakai', 'bersih', 'tidak', 'available', 'kadang', 'make', 'biru', 'tidak', 'senampak', 'pokok', 'love']

The next step is to create a text corpus from all the documents that have been preprocessed. This is the set of documents used for topic modelling. Then, the topic model is built using the LDA algorithm with parameters as in Table 1. This process is iterative to find the best model. The best model is used for topic modelling. This research uses topic coherence named C_v coherence (Röder *et al.* 2015) to determine the best model. C_v coherence measures how often words in a topic tend to appear together, that is, it measures the extent to which words in the topic are logically related. The formula for the calculation of the coherence measure is in Equation 1.

$$C_v = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \log \frac{D(w_i, w_j)+e}{D(w_i)+D(w_j)+e} \quad (1)$$

where n is the number of top words in the topic, $D(w_i, w_j)$ is the number of documents that contain both words w_i and w_j, $D(w_i)$ is the number of documents that contain word w_i and e is a smoothing parameter (Axelborn and Berggren 2023).

C_v values range from 0-1, with higher values indicating better coherence. A higher coherence value means the topics identified by the LDA model are more well-defined. This process also decides on the optimal number of topics. The number of topics affects how the LDA analysis results are interpreted. Table 4 shows the results of the topic coherence evaluation.

Table 4 Topic coherence value

Number of Topics	Coherence Value
2	0.3290
3	0.4136
4	0.3811
5	0.3826
6	0.3919
7	0.3892
8	0.3684

Table 4 shows the highest coherence value is when LDA creates a model with 3 topics. This is the best model for topic modelling. pyLDAvis in Figure 3 is used to visualize this model. The visualization shows the distribution of words in each topic.

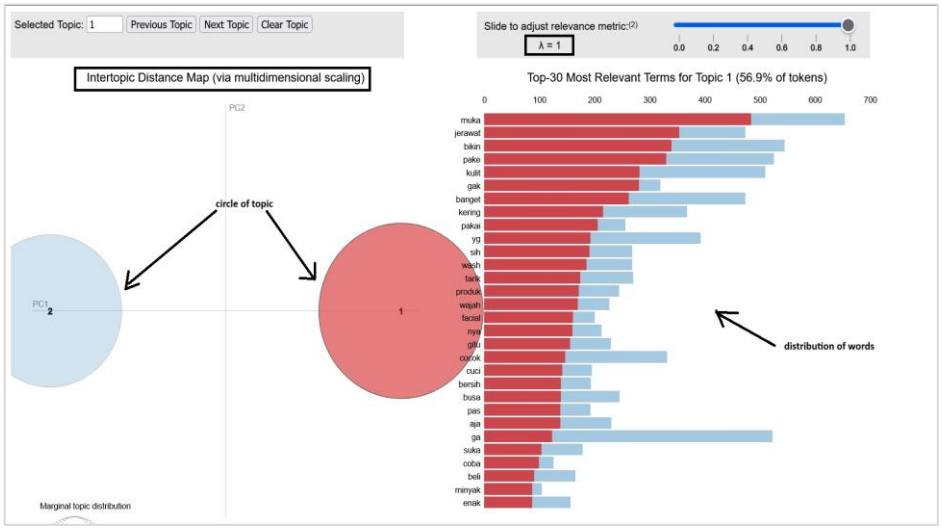

Figure 3 LDAvis display

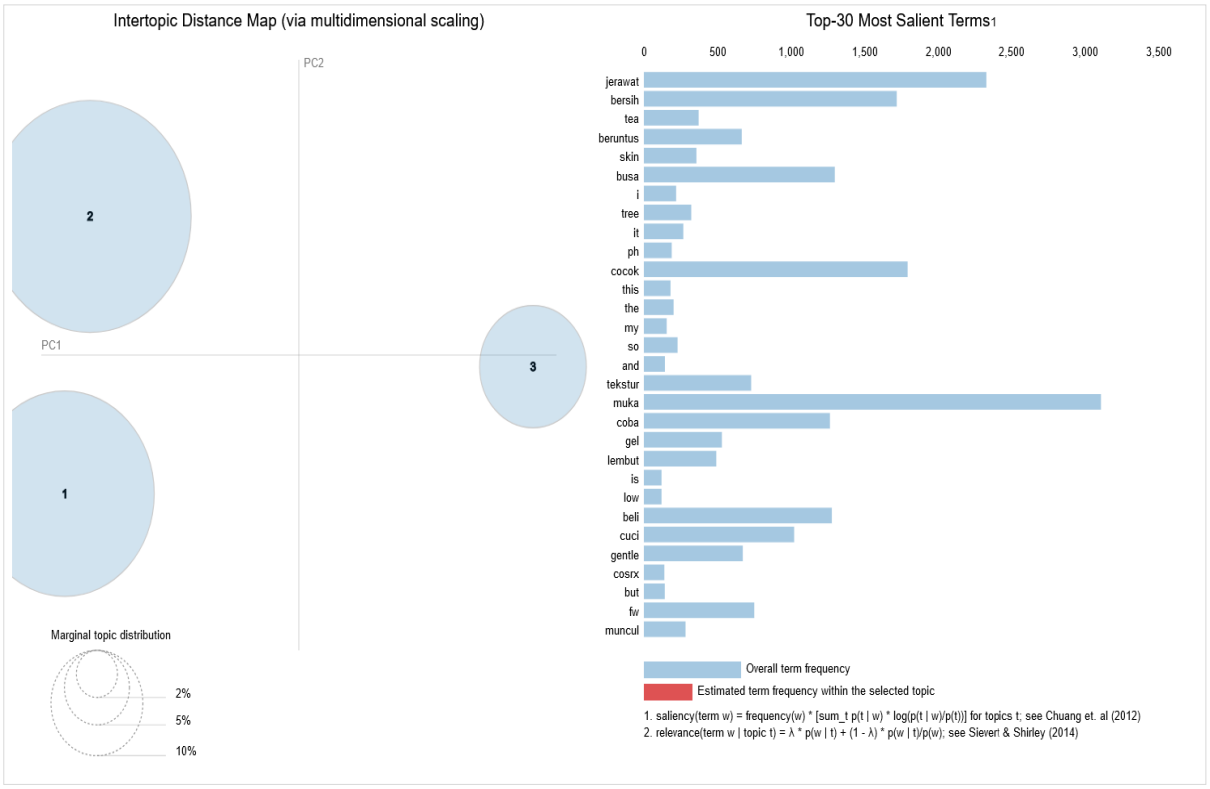

Figure 4 Topic modeling visualization

Figure 4 on the left shows a bubble diagram with distances between topics and bubble sizes representing topic prevalence in the corpus. Bubbles represent topics. The larger the bubble, the more reviews there are on that topic. Each bubble is separate, so the topics are distinct. It can be concluded that Topic 2 is the most dominant across the analyzed documents, where 48.5% of the words in the dataset are related to topic 2. The next order is Topic 1 with 38% and Topic 3 with 13.5%.

The horizontal bar chart on the right shows the most common words used in each topic. The length of the bar represents the frequency of topics arising from a single word. The left and right parts of the visualization are connected, where the left part selects a topic and the right part shows the most frequently occurring words to interpret it. pyLDAvis shows that there are 30 main words for each topic. Bar chart used to show the percentage of each word on each topic. The bar charts for each topic are shown in Figure 5, Figure 6, and Figure 7 respectively.

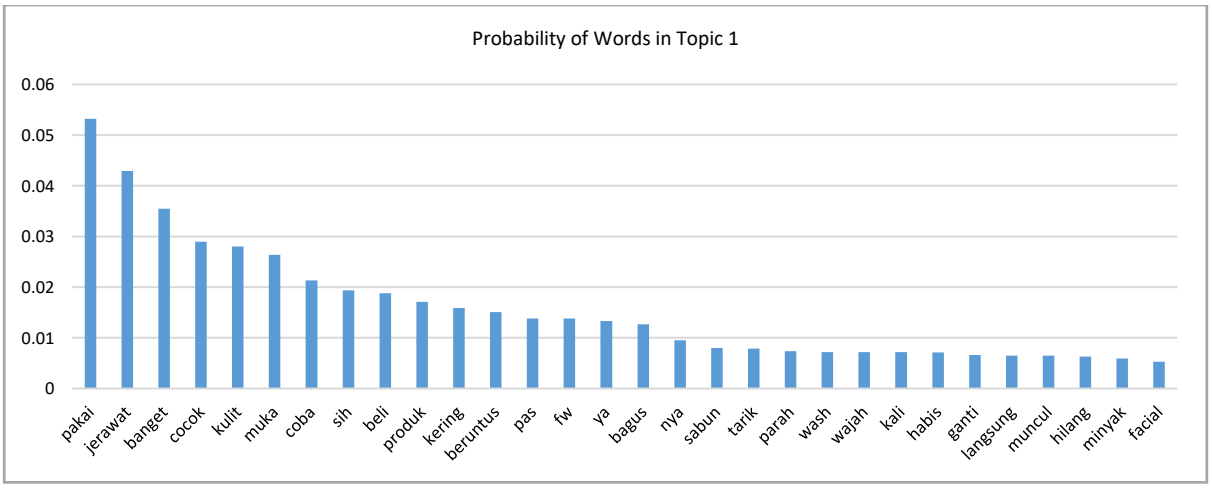

Figure 5 Word distribution for Topic 1

Figure 5 shows the word distribution for Topic 1, highlighting key terms such as "jerawat" (0.0429), "cocok" (0.029), "kulit" (0.028), "muka" (0.0264), "kering" (0.0159), and "minyak" (0.0059). These terms suggest that Topic 1 contains user conversations about the impact of products on acne-prone skin.

A careful examination of this word distribution reveals that consumer perception is heavily influenced by the product's impacts on their skin condition. The frequent usage of phrases such as "jerawat", "kering", and "minyak" implies that customers largely judge the efficacy of these products depending on whether they aggravate or relieve acne. The presence of words like "tarik" (0.0079), “muncul” (0.0065), "beruntus" (0.0151), and "parah" (0.0074) further emphasizes concerns regarding undesirable side effects such as excessive dryness or worsening acne conditions.

Furthermore, the probability distribution of words in Topic 1 is consistent with recent research on skincare customer behavior, which has identified product compatibility and perceived effectiveness as important decision-making considerations (Sujith 2024). The presence of "cocok" in the word rating promotes the idea that people should evaluate if the product suits their individual skin needs. This suggests that subjective experiences rather than objective product characteristics, have a significant impact on buyer satisfaction.

By examining the visual representation in Figure 5, it is clear that the product's ability to enhance or exacerbate acne-prone skin determines consumer preference. Users evaluate a product adversely when it causes dryness or oil imbalance, resulting in fewer recommendations. This information can help manufacturers improve formulas for acne-prone skin while eliminating unwanted effects.

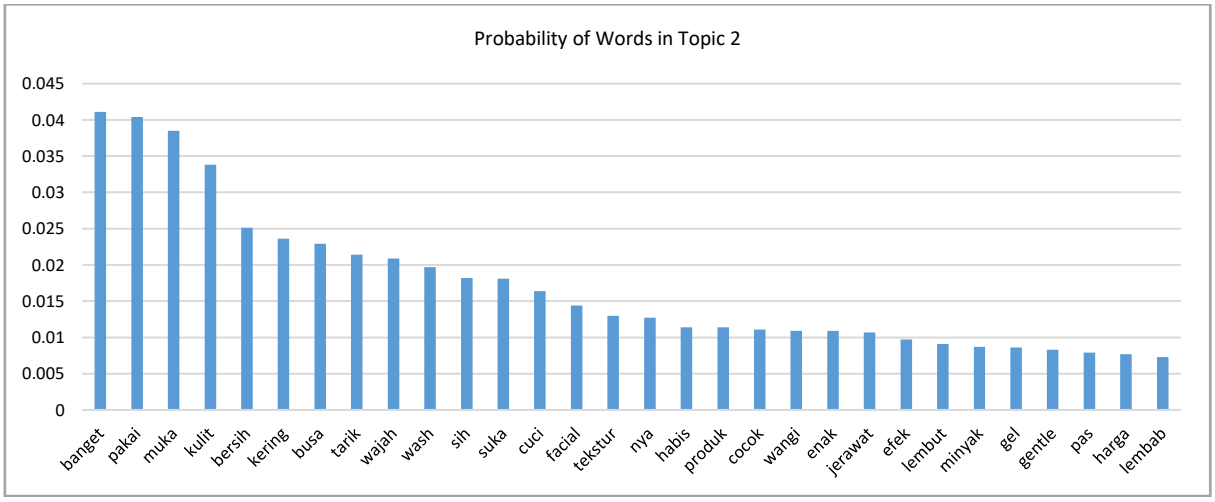

Figure 6 Word distribution for Topic 2

Figure 6 presents key terms such as "bersih" (0.0251), "kering" (0.0236), "busa" (0.0229), "tarik" (0.0214), "suka" (0.0181), and "tekstur" (0.013). These words suggest that Topic 2 represents user discussions on preferred product advantages, including texture, price, and the perceived effects after use.

A more in-depth look at word distribution demonstrates that users base their evaluations of product quality on sensory experience and performance. Words like "bersih", "lembut" (0.0091), and "wangi" (0.0109) clearly demonstrate the importance of cleanliness, tenderness, and smell in influencing consumer preference. Similarly, terminology like "kering" and "minyak" (0.0087) imply that cutaneous reactions, such as dryness or oil control are taken into account when determining product effectiveness.

Furthermore, the prevalence of "harga" (0.0077) in the distribution suggests that affordability influences consumer pleasure. Consumers appear to value a balance between performance and cost, which is indicating that a product’s texture, foaming ability and skin effects are weighed against its price point.

A juxtaposition of Topics 1 and 2 discloses that Topic 1 centers on acne-prone skin concerns, while Topic 2 accentuates positive attributes that influence purchasing decisions. The observed inclination towards specific product attributes indicates that manufacturers have the potential to enhance formulations by improving texture and ensuring the achievement of desirable skin effects, thereby aligning with consumer demand. This critical interpretation of Topic 2 facilitates a more profound comprehension of user behavior and the factors that influence their selection of facial wash products.

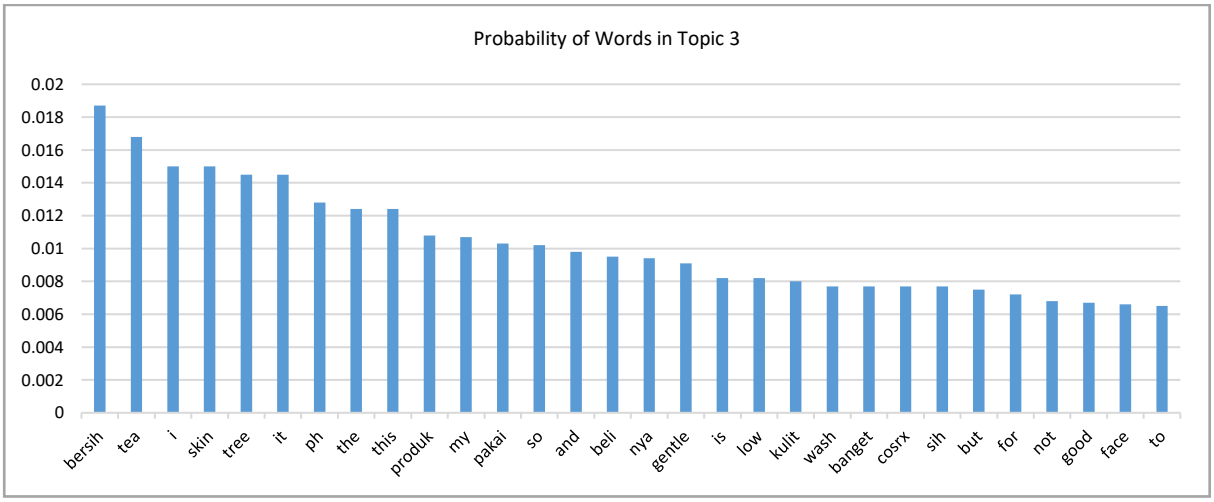

Figure 7 Word distribution for Topic 3

As illustrated in Figure 7, the keywords "tea" (0.0168), "tree" (0.0145), "pH" (0.0128), "gentle" (0.0091), "low" (0.0082), and "good" (0.0067) collectively indicate that Topic 3 focuses on user discussions concerning products with natural ingredients.

A more thorough examination of these terms indicates that consumers place a higher value on formulations that are gentle and have low pH levels. These formulations are frequently associated with skincare products that are less irritating and more appropriate for individuals with sensitive skin. The presence of tea tree, a well-known ingredient recognized for its antibacterial and anti-inflammatory properties, serves to reinforce the notion that users are interested in natural solutions for skin health and acne treatment.

Furthermore, the prevalence of terms such as "gentle" and "good" signifies a user preference for products that offer mild cleansing effects, thus avoiding the use of harsh formulations that may result in the stripping of the skin's natural moisture. In comparison with Topic 1 which is centers on reactions to acne-prone skin, Topic 3 emphasizes the proactive user choices in selecting skincare products with natural components that are in accordance with their skin needs.

This interpretation implies that producers could improve product appeal by using more botanical extracts and maintaining pH balance in facial wash compositions. Furthermore, users' emphasis on these characteristics suggests a growing understanding of constituent safety and effectiveness, emphasizing the necessity of transparency in product labeling and marketing techniques.

CONCLUSION

This study successfully applied LDA for topic modeling on consumer reviews of facial wash products, which is identifying three key discussion topics. The topic of preferred product advantages emerged as the most frequently discussed topic (48.5%), followed by the impact of products on acne-prone skin (38%) and products with natural ingredients (13.5%). These results demonstrate that rather than objective product parameters, user assessments are largely influenced by perceived benefits and skin compatibility. The study's conclusions are useful to manufacturers and consumers alike, as they assist firms in improving product offers, marketing tactics, and formulas while giving potential customers data-driven insights to help them make

well-informed selections. Despite its contribution, this research has some limitations in the form of a normalization dictionary that is still incomplete and not well structured. There are still many foreign words that have not been converted to Indonesian, thus creating too many word variations but actually have the same meaning. Some slang words are still there. This normalization dictionary should be more complete because it can affect the results of topic modelling. The normalization dictionary should be updated to match the problem domain. Future research should focus on improving the text normalization process, exploring alternative topic modeling approaches such as BERT-based models or Non-Negative Matrix Factorization (NMF), and integrating sentiment analysis to further refine the interpretation of consumer preferences and product attributes.

REFERENCES

Al-Badrashiny M, Pasha A, Diab M, Habash N, Rambow O, Salloum W, Eskander R. 2016. SPLIT: Smart Preprocessing (Quasi) Language Independent Tool. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*. 4055–4060.

Alpiana V, Salam A, Alzami F, Rizqa I, Aqmala D. 2024. Analisis Topic-Modelling Menggunakan Latent Dirichlet Allocation (LDA) pada Ulasan Sosial Media Youtube. *Jurnal Media Informatika Budidarma*. [online] 8(1): 332. https://doi.org/10.30865/mib.v8i1.7127.

Alzami F, Megantara RA, Prabowo DP, Sulistiyawati P, Pramunendar RA, Dewi IN, Rizqa I, Ritzkal. 2023. LDA Topic Analysis for Product Reviews in Social Media Platform. *Moneter: Jurnal Keuangan Dan Perbankan*. [online] 11(2): 277–83. https://doi.org/10.32832/moneter.v11i2.402.

Ameko MK, Baee S, Barnes LE. 2021. LonelyText: A Short Messaging Based Classification of Loneliness. [online] (January 2025). Available at: https://doi.org/10.48550/arXiv.2101.09138.

Axelborn H, Berggren J. 2023. Topic Modeling for Customer Insights: A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls [master's thesis]. Umeå (SE): Umeå University. https://umu.diva-portal.org/smash/get/diva2:1763637/FULLTEXT01.pdf.

Baranowski M. 2022. Epistemological Aspect of Topic Modelling in the Social Sciences: Latent Dirichlet Allocation. *Przegląd Krytyczny*. [online] 4(1): 7–16. https://doi.org/10.14746/pk.2022.4.1.1.

Barde BV, and Bainwad AM. 2017. An Overview of Topic Modeling Methods and Tools. *Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems, ICICCS 2017*. 745–750. https://doi.org/10.1109/ICCONS.2017.8250563.

Chundi P, Go S. 2015. Latent Dirichlet Allocation Approach for Analyzing Text Documents. *Encyclopedia of Information Science and Technology* Third Edition. Hershey(AS): PA: IGI Global. https://doi.org/10.4018/978-1-4666-5888-2.ch175.

Gagan SB, Chirag GA. 2022. Topic Modeling By Using LDA. *International Journal of Scientific Research in Engineering and Management*. [online] 06(06): 1–6. https://doi.org/10.55041/ijsrem14663.

Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L. 2019. Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. *Multimedia Tools and Applications*. [online] 78(11): 15169–15211. https://doi.org/10.1007/s11042-018-6894-4.

Jivani AG. 2011. A Comparative Study of Stemming Algorithms. *Int. J. Comp. Tech. Appl.* [online] 2(6): 1930–1938. https://doi.org/10.1093/oxfordhb/9780195396430.013.0038.

Ladani DJ, Desai NP. 2020. Stopword Identification and Removal Techniques on TC and IR Applications: A Survey. *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*. 466–472. https://doi.org/10.1109/ICACCS48705.2020.9074166.

Miao W, Lin KC, Wu CF, Sun J, Sun W, Wei W, Gu C. 2023. How Could Consumers' Online Review Help Improve Product Design Strategy? *Information*. [online] 14(8): 434. https://doi.org/10.3390/info14080434.

Patil HB, Pawar B.V, and Patil AS. 2016. A Comprehensive Analysis of Stemmers Available for Indic Languages. *International Journal on Natural Language Computing*. [online] 5(1): 45–55. https://doi.org/10.5121/ijnlc.2016.5104.

Phanidhar A, Kowshick A, Mahendrareddy A, Nikitha A, Hariharan G. 2023. Topic Modelling of Web Pages with Latent Dirichlet Allocation Methods. *International Journal of Scientific Research in Engineering and Management*. [online] 07(11): 1–11. https://doi.org/10.55041/ijsrem27350.

Puspita E, Shiddieq DF, Roji FF. 2024. Pemodelan Topik Pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc). *MALCOM: Indonesian Journal of Machine Learning and Computer Science*. [online] 4(2): 481–489. https://doi.org/10.57152/malcom.v4i2.1204.

Rajesh A, Hiwarkar T. 2023. Exploring Preprocessing Techniques for Natural LanguageText: A Comprehensive Study Using Python Code. *International Journal of Engineering Technology and Management Sciences*. [online] 7(5): 390–399. https://doi.org/10.46647/ijetms.2023.v07i05.047.

Röder M, Both A, Hinneburg A. 2015. Exploring the Space of Topic Coherence Measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. 399–408. https://doi.org/10.1145/2684822.2685324.

Salsabila KDA, Trianasari N. 2021. Analisis Persepsi Produk Kosmetik Menggunakan Metode Sentiment Analysis Dan Topic Modeling (Studi Kasus: Laneige Water Sleeping Mask). *Jurnal Teknologi Dan Manajemen Informatika*. [online] 7(1): 1–9. https://doi.org/10.26905/jtmi.v7i1.5593.

Sharma A, Parmar M. 2021. A Survey on Text Pre-Processing and Feature Extraction Techniques for Sentiment Analysis of Twitter Data. *International Research Journal of Computer Science*. [online] 8(12): 271–278. https://doi.org/10.26562/irjcs.2021.v0812.001.

Sharma H, Sharma AK. 2017. Study and Analysis of Topic Modelling Methods and Tools – A Survey. *American Journal of Mathematical and Computer Modelling*. [online] 2(2): 84–87. https://doi.org/10.11648/j.ajmcm.20170203.12.

Sievert C, Shirley KE. 2014. LDAvis: A Method for Visualizing and Interpreting Topics. *Proceedings Ofthe Workshop on Interactive Language Learning, Visualization, and Interfaces*. 63–70. https://doi.org/10.3115/v1/w14-3110.

Sujith. 2024. A Study on Factors Influencing Purchase Behaviour of Skin Care Products. *International Journal of Future Management Research*. [online] 6(4): 1-8. https://doi.org/10.36948/ijfmr.2024.v06i04.26112

Sutherland I, Sim Y, Lee SK, Byun J, Kiatkawsin K. 2020. Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation. *Sustainability*. [online] 12(5): 1–15. https://doi.org/10.3390/su12051821.

Swain K, Nayak AK. 2018. A Review on Rule-Based and Hybrid Stemming Techniques. *Proceedings - 2nd International Conference on Data Science and Business Analytics, ICDSBA 2018*. 25–29. https://doi.org/10.1109/ICDSBA.2018.00012.

Tabassum A, Patil RR. 2020. A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. *International Research Journal of Engineering and Technology*. [online] 07(06): 4864–4867. https://www.irjet.net/archives/V7/i6/IRJET-V7I6913.pdf

Wang C, Shao Q. 2023. LDA-Based Cosmetic Satisfaction Factors Mining. *Proceedings of the 7th International Conference on Information Systems Engineering*. 46–51. https://doi.org/10.1145/3573926.3573935.

Wang W, Feng Y, Dai W. 2018. Topic Analysis of Online Reviews for Two Competitive

Products Using Latent Dirichlet Allocation. *Electronic Commerce Research and Applications*. [online] 29: 142–156. https://doi.org/10.1016/j.elerap.2018.04.003.

Xiao-bo L. 2011. Use of LDA Model in Topic Tracking. *Computer Science*. https://api.semanticscholar.org/CorpusID:201897233.