Journal of Electrical Engineering and Computer (JEECOM)
Vol. 7, No. 1 (2025), DOI: 10.33650/jeecom.v4i2
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Sentiment Modeling of Instagram Users Towards Traditional and Modern Body Scrubs Using the Naive Bayes Algorithm

Mardiana [1*], Kusrini [2]
[1,2] Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Ngringin, Condongcatur, Daerah Istimewa Yogyakarta 55281, Indonesia

Article Info

Article history:

Received Dec 16, 2024
Revised Dec 18, 2024
Accepted April 10, 2025

Keywords:

Sentiment Analysis
Naïve Bayes
Body Scrub
Consument Perception
Instagram

ABSTRACT

The aim of the research is to identify and compare consumer sentiments—positive, negative, or neutral—toward these two categories of skincare products between modern and traditional scrub using Naïve Bayes algorithm. The results indicate that neutral sentiment dominates, followed by negative and positive sentiments. The Naive Bayes algorithm demonstrated strong performance, particularly in detecting negative and neutral sentiments, but exhibited a lower recall rate for positive sentiments. The findings reveal that consumers value traditional body scrubs for their natural ingredients and cultural significance, while modern body scrubs are appreciated for their innovation. Additionally, K-Means Clustering was applied to analyze employee density data from a government agency, highlighting the influence of cluster determination (*k*) and initial centroid values on clustering outcomes. The analysis grouped 4788 employees into four clusters: very good (C1, 1995 employees), good (C2, 1936), fair (C3, 842), and poor (C4, 15). The Davies-Bouldin Index (DBI) produced a value of 1.89, indicating some overlap in cluster separation but acceptable performance. These findings emphasize the importance of optimizing clustering parameters and offer actionable insights for skincare brands to tailor marketing strategies while showcasing the value of data-driven approaches in organizational analysis.

Corresponding Author:
Mardiana,
Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Ngringin, Condongcatur, Daerah Istimewa Yogyakarta 55281, Indonesia
Email: mardiana@students.amikom.ac.id

1. INTRODUCTION

As of January 2023, there were 5.15 billion internet users worldwide, representing 64.4% of the global population of 8.01 billion. This figure marks an increase of 1.9% from the 5.01 billion users reported during the same period the previous year [1]. The evolution of social media has fundamentally transformed how consumers interact with brands, creating highly valuable channels for understanding consumer preferences and sentiments. Customer reviews play a crucial role in assessing product quality in today's digital era [2]. The widespread use of social media among college students is considered highly effective in spreading awareness about environmental sustainability [3]. Social media serves as a platform capable of disseminating information rapidly, but it also has the potential to be more misleading compared to traditional media or television broadcasts [4]. Among these platforms, Instagram has emerged as a highly influential space where users share their views on various products, including skincare items like body scrubs. Analyzing user-generated opinions provides valuable insights that enable companies to refine their product offerings and design more effective marketing strategies. Understanding customer perceptions is essential for determining whether a product or service delivers the desired value [5].

Skincare products are currently very popular, with both men and women showing a strong interest in purchasing them [6]. Body scrubs, an essential part of skincare routines, come in diverse formulations that can generally be categorized into traditional and modern products. Traditional body scrubs are valued for their use of natural ingredients and their deep-rooted connection to cultural practices, appealing to consumers seeking holistic and organic skincare solutions. In contrast, modern body scrubs, which emphasize innovation and ease of use, attract a market segment that values efficiency and the latest advancements in skincare technology. For brands, understanding how these two product categories are perceived by consumers is crucial to staying competitive in a crowded market.

Sentiment analysis, a technique within the domain of Natural Language Processing (NLP), offers a powerful way to extract and measure consumer sentiment from textual data. By applying machine learning algorithms such as Naive Bayes classification, researchers can analyze large datasets of text from platforms like Instagram to uncover underlying sentiments—positive, negative, or neutral—expressed by consumers about specific products. Sentiment analysis enables a deeper understanding of user responses and provides actionable insights for brands to optimize their strategies [7]. Several methods can be used to perform sentiment analysis, one of which is through machine learning [8]. The Naive Bayes classification method is a technique used in sentiment analysis, offering theoretical advantages related to data consistency and the efficiency of classification computations [9]. Naive Bayes also applies probability and the concept of likelihood in the classification process for sentiment analysis [10]. This approach not only provides a systematic method for interpreting consumer feedback but also helps identify sentiment trends that can influence purchasing behavior. Through sentiment analysis, promotions and marketing can be tailored via television ads, street banners, and digital campaigns based on sentiment data. Consumers can now view, select, and purchase products using smartphones, laptops, and computers without the need for physical travel [11].

This study aims to perform sentiment analysis on Instagram comments related to traditional and modern body scrub products using the Naive Bayes algorithm. By examining a large number of comments, this research seeks to identify and compare the dominant sentiments associated with each category of body scrub. The results of this analysis are expected to provide valuable insights for skincare brands regarding consumer preferences, thus supporting the development of marketing strategies that are more aligned with their target audience. Additionally, these findings will contribute to a broader understanding of consumer perceptions of traditional versus modern products within the beauty and skincare industry in the digital age.

2. METHOD

2.1. Data Source

In this study, Instagram will serve as the primary data source. Instagram is a widely used social media platform where users frequently share their opinions and experiences about various products, including skincare items. With its broad reach and high user engagement, this platform is an ideal choice for collecting data related to consumer sentiment towards body scrub products.

2.2. Scope of Data

The scope of data collection will focus on posts and comments on Instagram related to specific beauty brands. We will concentrate on comments and reviews about body scrub products from several leading beauty brands, namely Sari Ayu Marta Tilaar, Marina, The Body Shop, Purbasari, and Mabello. These brands represent a blend of traditional and modern skincare solutions, providing a comprehensive overview of consumer sentiment towards various product categories.

2.3. Data Preprocessing

In this stage, the collected data will undergo a cleaning and pre-processing process. This pre-processing phase is crucial to prepare the data for analysis, ensuring that it is in the appropriate format for further examination [12]. First step is to clean the documents by removing numbers, usernames, websites, and hashtags [13]. Before sentiment analysis is performed, the raw data collected from Instagram comments will undergo a thorough cleaning process. This step is crucial to ensure that the text data is in a suitable format for analysis. The text cleaning process includes the following steps:

1. **Removing Irrelevant Content:** All promotional content, advertisements, and irrelevant links will be removed to focus solely on genuine user comments about body scrub products.
2. **Removing Special Characters and Emojis:** Special characters, emojis, and unnecessary spaces will be eliminated from the text. This step helps standardize the data and prevents potential disruptions during analysis.
3. **Correcting Typos and Normalizing Text:** Typographical errors will be corrected, and the text will be normalized to ensure consistency throughout the dataset.

2.4. Tokenization

After the cleaning process is complete, the next step is tokenization, which involves breaking the text into smaller units known as tokens. Tokens can be individual words or phrases, and this process helps convert the text into a structured format that can be analyzed. The tokenization process includes the following steps:

1. **Splitting the Text:** The cleaned text will be broken into tokens by separating words and phrases based on spaces and punctuation.
2. **Creating Tokens:** Each token will represent a distinct meaningful word or phrase, which will then be used for further analysis.

2.5. Stopwords Removal

Removing stopwords and minimizing noise in the text data is crucial for improving the accuracy and quality of sentiment analysis [14]. Stopwords are common words that do not carry significant meaning in sentiment analysis and can be removed to allow the analysis to focus more on relevant terms. In the context of Indonesian language comments, stopwords include words such as: (yang, dan, di, untuk, dari, ada, ini, saya, itu, sebuah, akan, atau, dengan, kamu, jika, pada, oleh, and kami). These stopwords will be removed from the text data to eliminate common words that do not provide meaningful insights into consumer sentiment.

2.6. Stemming

The final step in pre-processing is stemming, which is the process of reducing words to their base form. Stemming helps consolidate word variations into the same root form, thus improving consistency and accuracy in sentiment analysis. By applying stemming, the dataset becomes simpler, and word variations are treated as equivalent, making sentiment analysis of consumer opinions more accurate. For stemming example:

- "memilih" (choosing) and "pilihan" (choice) will be reduced to the base form "pilih" (choose).
- "membeli" (buying) and "beli" (buy) will be reduced to "beli" (buy).

2.7. Sentiment Analysis

Sentiment Analysis, also known as opinion mining, is a research field within Natural Language Processing (NLP) [15]. Sentiment analysis is a crucial part of this study, aiming to extract and measure consumer sentiment from Instagram comments regarding traditional and modern body scrub products. This analysis involves several key steps, including feature extraction, model training, and classification.

2.8. Feature Extraction

The first step in sentiment analysis is feature extraction, where textual data is transformed into a numerical format that can be processed by machine learning algorithms. This process includes:

- **Text Representation:** Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and the bag-of-words model will be used to convert textual data into numerical features. TF-IDF measures the importance of each word within the context of a document relative to the entire dataset, while the bag-of-words model counts the frequency of each word in the text. These representations capture critical information from the comments while maintaining the relevance of words for sentiment analysis.
- **Vectorization:** The cleaned and tokenized text will be visualized as feature vectors. This step enables the application of machine learning algorithms to the data.

2.9. Model Training

After feature extraction, the next step is training the Naive Bayes classifier. The Naive Bayes algorithm is well-suited for sentiment analysis due to its simplicity and effectiveness in handling large datasets. The training process involves:

- **Dataset Preparation:** A labeled dataset will be used to train the Naive Bayes model. This dataset consists of comments classified into sentiment categories: positive, negative, or neutral. The model will learn how to associate specific features (words or phrases) with each sentiment category based on this labeled data.
- **Naive Bayes Classifier:** Naive Bayes is a probabilistic classifier based on Bayes' Theorem, with an assumption of feature independence. Despite its simplicity, this algorithm performs well in text classification tasks. The algorithm calculates the probability of a comment belonging to a specific sentiment category based on the extracted features. The formula used by the Naive Bayes model to predict the probability of each class is shown in the accompanying image.

$$P(c|d) = \frac{P(d|c).P(c)}{P(d)} \tag{1}$$

- **P(c|d)**: The probability that document *d* belongs to sentiment category *c*, based on the features within the document.
- **P(d|c)**: The probability of observing document *d* given that it belongs to category *c*.
- **P(c)**: The prior probability for category *c*, representing the likelihood that a document belongs to category *c* before considering specific features in the document.

- **P(d)**: The total probability of document *d* appearing in the dataset, regardless of its category.

2.10. Classification

After training, the Naive Bayes classifier will be used to classify the sentiment of comments in the Instagram dataset. The classification process includes the following steps:

- **Applying the Model:** The trained Naive Bayes model will classify each comment into one of the sentiment categories: positive, negative, or neutral. The model will use the probabilities it has learned to make predictions based on the features extracted from each comment.
- **Interpreting Results:** The classified comments will be analyzed to determine the overall sentiment distribution towards traditional and modern body scrub products. This analysis will provide insights into consumer attitudes toward these products and help identify preferences and trends.

3. RESULTS AND DISCUSSION

3.1. Dataset

Table 1 presents the dataset utilized in this study, which was gathered through web scraping from Instagram. The dataset specifically focuses on user comments regarding body scrub products. The dataset consists of 123 entries, where each entry represents an individual comment made by a user on various posts. The dataset is organized into three main columns:

1. **Date:** This column records the timestamp of each comment, indicating when users posted their comments. The date and time format used is "YYYY-MM-DD HH," allowing for more accurate temporal analysis.
2. **User:** This column contains the usernames of the individuals who posted the comments. These usernames serve as unique identifiers for each commenter, but personal data has been anonymized to protect user privacy.
3. **Text:** The text column includes the content of the user comments. This unstructured data encompasses various sentiments, opinions, questions, and feedback related to the discussed body scrub products. The comments are written in an informal language, reflecting the natural conversational style commonly found on social media platforms.

Table 1. Raw Dataset

date	user	text
2020-11-05 23:32:36	aisha.beautyskin	Pepaya emg the best fruit ever!!
2020-10-25 08:06:44	filzaalrn	Di masih minyak zaitun lg ga ka?
2020-10-22 14:27:11	dwisetyowati86	Kenapa si kak yg tanjung stoknya dimana mana gaada
2020-10-22 14:03:23	dwisetyowati86	Herannn
2020-10-22 14:00:43	dwisetyowati86	Kenapa si kak yg tanjung stoknya dimana mana gaada
2020-10-15 03:43:18	aprillianimeri	Enak nih pke ini
2020-10-12 12:00:02	arisa_baruany	Kak aku mau tanya, sari ayu body scrub isinya memang encer ya ?
2020-10-11 08:14:49	sa.samman	Aku suka varian yang warnanya pink tapi lupa namany 😳

The presented dataset has undergone a preprocessing phase, during which each comment was cleaned and labeled according to its sentiment using a pipeline developed by Wilson Wongso [16]. The cleaned data is organized into two main columns:

1. **Cleaned_Text:** This column contains comments after undergoing a text-cleaning process, such as the removal of irrelevant characters, punctuation marks, and stopwords. The resulting text is a more concise version of the original comments, focusing on the most relevant words for sentiment analysis.
2. **Sentiment:** This column represents the sentiment label assigned to each comment. Sentiment labels are categorized into three classes: positive, negative, and neutral. These labels indicate the overall sentiment expressed in each comment. For example, a comment like *"Papaya is the best fruit ever!!"* is labeled as positive, reflecting a favorable opinion, whereas a comment like *"Why is the tanjung stock unavailable everywhere?"* is labeled as negative, indicating dissatisfaction. Neutral comments, such as *"I want to ask, is the Sari Ayu body scrub normally runny?"*, do not strongly express positive or negative sentiment.

Table 2 presents a labeled dataset that distinguishes between neutral, positive, and negative sentiments. This labeled dataset serves as the foundation for training and evaluating the sentiment analysis model. The distribution of sentiment labels provides insights into overall consumer attitudes toward the discussed body scrub products. The dataset structure, with clearly defined features and labels, facilitates the application of machine learning algorithms, particularly for tasks such as classification and prediction in sentiment analysis.

Table 2. Labeled Dataset

cleaned_text	sentiment
Pepaya emg the best fruit ever!!	positive
Di masih minyak zaitun lg ga ka?	negative
Kenapa si kak yg tanjung stoknya dimana mana gaada	negative
Herannn	negative
Kenapa si kak yg tanjung stoknya dimana mana gaada	negative
Enak nih pke ini	negative
Kak aku mau tanya, sari ayu body scrub isinya memang encer ya ?	neutral
Aku suka varian yang warnanya pink tapi lupa namany 😳	positive

3.2. Sentiment Distribution

The sentiment analysis model was evaluated based on three key metrics: precision, recall, and F1-score, across the categories of negative, neutral, and positive sentiment. Table 3 presents the sentiment analysis results of this study, including precision, recall, and F1-score for each sentiment category. Below is a detailed explanation of each metric and the insights provided by the results:

1. **Negative Sentiment:**
   - **Precision (0.74):** Among all comments predicted as negative, 74% were correctly identified as negative. This indicates that while the model performs well in identifying negative sentiment, there is still a 26% chance that comments classified as negative may not actually be negative.
   - **Recall (1.00):** The model achieved a perfect recall score of 1.00, meaning it successfully identified all truly negative comments. This demonstrates the model's high sensitivity in detecting negative sentiment, ensuring that no negative comments were missed.
   - **F1-Score (0.85):** The F1-score, which balances precision and recall, is 0.85. This high score indicates that the model performs reliably overall in classifying negative comments, making it dependable for identifying negative sentiment.
2. **Neutral Sentiment:**
   - **Precision (0.94):** The model achieved very high precision of 0.94 for neutral comments, meaning 94% of the comments predicted as neutral were correctly classified. This demonstrates the model's high accuracy in identifying neutral sentiment, with only a small margin for error.
   - **Recall (1.00):** Similar to negative sentiment, the model achieved a perfect recall score of 1.00 for neutral comments. This indicates that the model successfully identified all neutral comments in the dataset.
   - **F1-Score (0.97):** With an F1-score of 0.97, the model shows excellent performance in classifying neutral comments. This suggests the model is nearly flawless in detecting neutral sentiment, combining high precision and perfect recall.
3. **Positive Sentiment:**
   - **Precision (1.00):** The model achieved perfect precision for positive comments, meaning that every comment classified as positive was indeed positive. This represents an ideal scenario where there are no false positives (i.e., no comments were incorrectly labeled as positive).
   - **Recall (0.57):** However, the recall for positive comments is 0.57, indicating that the model only identified 57% of the actual positive comments. This reveals that while the model is highly precise, it is less effective at capturing all positive sentiments, potentially missing a significant portion of them.
   - **F1-Score (0.73):** An F1-score of 0.73 reflects a balance between perfect precision and lower recall. This score indicates that while the model is excellent at accurately identifying positive sentiment when it does so, there is room for improvement in ensuring all positive sentiments are detected.

Table 3. Sentiment Result

Sentiment	Precision	Recall	F1-Score
Negative	0.74	1.00	0.85
Neutral	0.94	1.00	0.97
Positive	1.00	0.57	0.73

Figure 1 illustrates a word cloud that highlights the terms commonly associated with positive sentiment. In this visualization, the size of each word reflects its frequency of occurrence, where larger words signify higher usage. Notably, words such as *natural*, *scrub*, *lembut* (soft), and *menawan* (charming) stand out as some of the most frequently mentioned terms. These words indicate a recurring emphasis on qualities that convey a sense of gentleness, appeal, and natural attributes within the positive sentiment. This pattern suggests a preference for describing experiences or products in a way that highlights their soothing, attractive, and natural characteristics.

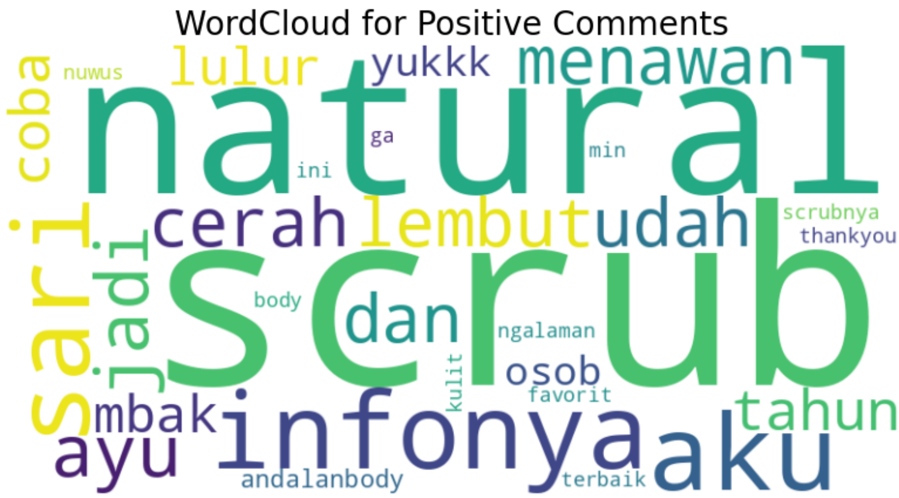

Figure 1. Word Cloud of Positive Sentiment

4. CONCLUSION

Based on the results of this study, it can be concluded that clustering using the K-Means Clustering method is highly influenced by the determination of the number of clusters (k) and the initial centroid values. These factors affect the calculation of the nearest distance, the number of iterations, and the accuracy of the group membership data. The author performed clustering on employee density data from a government agency using the K-Means Clustering algorithm, resulting in four clusters (k=4) with the following cluster statuses: very good (C1), good (C2), fair (C3), and poor (C4). From the 4788 employee data analyzed between January and July 2024, the following results were obtained: 1995 employees were in Cluster 1 with a very good status, 1936 employees in Cluster 2 with a good status, 842 employees in Cluster 3 with a fair status, and 15 employees in Cluster 4 with a poor status. The accuracy test using the Davies-Bouldin Index (DBI) in Google Colab showed that the clustering of employee density from 4788 employee data into 4 clusters achieved a PerformanceVector accuracy of 1.89. The Davies-Bouldin Index (DBI) value for K = 4 of 1.89 indicates that there is some level of overlap or imperfection in the cluster separation, although this value is not too bad. However, it cannot be considered an optimal clustering result, as the ideal value should approach 0.

REFERENCES

[1] A. Erfina and M. R. N. Ramdani Alamsyah, "Implementation of Naive Bayes classification algorithm for Twitter user sentiment analysis on ChatGPT using Python programming language," *Data and Metadata*, vol. 2, p. 45, May 2023, doi: 10.56294/dm202345.

[2] R. Rahmawati, W. Fuadi, and Y. Afrillia, "Cosmetic Shop Sentiment Analysis on TikTok Shop Using the Support Vector Machine Method," *Int. J. Eng. Scie. and Inform. Technology.*, vol. 4, no. 2, pp. 31–37, Jul. 2024, doi: 10.52088/ijesty.v4i2.498.

[3] S. Hamid, M. T. Ijab, H. Sulaiman, R. Md. Anwar, and A. A. Norman, "Social media for environmental sustainability awareness in higher education," *IJSHE*, vol. 18, no. 4, pp. 474–491, May 2017, doi: 10.1108/IJSHE-01-2015-0010.

[4] I. Irawanto, C. Widodo, A. Hasanah, P. A. Dharma Kusumah, K. Kusrini, and K. Kusnawi, “Sentiment Analysis and Classification of Forest Fires in Indonesia,” *Ilk. J. Ilm.*, vol. 15, no. 1, pp. 175–185, Apr. 2023, doi: 10.33096/ilkom.v15i1.1337.175-185.

[5] Sukesh Marla, “SENTIMENT ANALYSIS OF CUSTOMER FEEDBACK ON APPAREL AND COSMETICS PURCHASE IN E-COMMERCE: A TAKE ON CUSTOMER SATISFACTION,” *ITII*, vol. 9, no. 1, pp. 786–794, Mar. 2021, doi: 10.17762/itii.v9i1.199.

[6] M. Habibi and K. Kusumaningtyas, “Customer Experience Analysis Skincare Products Through Social Media Data Using Topic Modeling and Sentiment Analysis,” *j. of sci. and appl.eng.*, vol. 6, no. 1, p. 1, Jun. 2023, doi: 10.31328/jsae.v6i1.4169.

[7] S. A. Ashari, M. W. A. Saputra, E. Larosa, and B. S. Rijal, “Analisis Sentimen pada Aplikasi Translate Google Menggunakan Metode SVM (Studi Kasus: Komentar Pada Playstore),” *jt*, vol. 21, no. 2, pp. 168–182, Dec. 2023, doi: 10.37031/jt.v21i2.412.

[8] Parasian D.P Silitonga, Mitra Hasibuan, Z Situmorang, and Desinta Purba, “Comparison of Tiktok User Sentiment Analysis Accuracy with Naïve Bayes and Support Vector Machine,” *IJATCSE*, vol. 12, no. 1, pp. 11–15, Feb. 2023, doi: 10.30534/ijatcse/2023/031212023.

[9] S. A. H. Bahtiar, C. K. Dewa, and A. Luthfi, “Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling,” *J. Inf. Syst. Informatics*, vol. 5, no. 3, Art. no. 3, Aug. 2023, doi: 10.51519/journalisi.v5i3.539.

[10] I. Irvandi, B. Irawan, and O. Nurdiawan, “Naive Bayes and Wordcloud for Sentiment Analysis of Halal Tourism in Lombok Island Indonesia,” *iiroi*, vol. 5, no. 1, May 2023, doi: 10.37058/innovatics.v5i1.6675.

[11] P. S. Hutapea and W. Maharani, “Sentiment Analysis on Twitter Social Media towards Shopee E-Commerce through Support Vector Machine (SVM) Method,” *jinav j. inf. vis.*, vol. 4, no. 1, pp. 7–17, Jan. 2023, doi: 10.35877/454RI.jinav1504.

[12] I. N. Switrayana, D. Ashadi, H. Hairani, and A. Aminuddin, “Sentiment Analysis and Topic Modeling of Kitabisa Applications using Support Vector Machine (SVM) and Smote-Tomek Links Methods,” *IJECSA*, vol. 2, no. 2, pp. 81–91, Sep. 2023, doi: 10.30812/ijecsa.v2i2.3406.

[13] W. B. Zulfikar, A. R. Atmadja, and S. F. Pratama, “Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes,” *SJI*, vol. 10, no. 1, pp. 25–34, Jan. 2023, doi: 10.15294/sji.v10i1.39952.

[14] K. Mouyassir, A. Fathi, and N. Assad, “Elevating Aspect-Based Sentiment Analysis in the Moroccan Cosmetics Industry with Transformer-based Models,” *ijacsa*, vol. 15, no. 6, 2024, doi: 10.14569/IJACSA.2024.0150654.

[15] H. Benarafa, M. Benkhalifa, and M. Akhloufi, “An Enhanced SVM Model for Implicit Aspect Identification in Sentiment Analysis,” *IJACSA*, vol. 14, no. 5, 2023, doi: 10.14569/IJACSA.2023.0140505.

[16] Wilson Wongso, “indonesian-roberta-base-sentiment-classifier”, doi: 10.57967/HF/0644.