Journal of Emerging Information Systems and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

Sentiment Analysis and Word Association Patterns in Skincare Product Customer Reviews

Aliyah Alifi[1], Wiyli Yustanti[2]

[1,2]*Universitas Negeri Surabaya, Surabaya, Indonesia*

aliyah.22064@mhs.unesa.ac.id, wiyliyustanti@unesa.ac.id

ABSTRACT

The growth of the skincare industry and the increasing activity of consumer reviews on e-commerce platforms have generated large text data containing customer opinions, experiences, and perceptions. This study aims to analyze sentiment and identify word association patterns in Indonesian-language customer reviews of skincare products. The literature review covers sentiment analysis, Natural Language Processing, the IndoBERT language model, Data Mining, Knowledge Discovery in Databases, as well as Association Rule Mining using the Apriori algorithm. The research method uses a quantitative approach based on KDD, which includes Data Selection, Preprocessing, Data Transformation, Data Mining, and interpretation and evaluation. Data was obtained through web scraping of skincare product reviews on the Shopee platform, resulting in 7,320 clean reviews. Sentiment analysis was conducted using IndoBERT with a Hybrid Linguistic approach to handle neutral rating ambiguities. The results of the sentiment classification were then used as the basis for analyzing word association patterns using the Apriori algorithm for each sentiment category. The findings indicate that IndoBERT is capable of classifying sentiment contextually, while Apriori successfully uncovers word patterns that represent product aspects such as quality, effectiveness, and user experience. This study concludes that the integration of sentiment analysis and word association patterns provides a more comprehensive understanding of consumer perceptions and can be utilized as a basis for strategic decision-making in the skincare industry.

Keyword: sentiment analysis, IndoBERT, association rule mining, Apriori, skincare reviews

Article Info:
Article history:
Received April 02, 2026
Revised April 10, 2026
Accepted June 02, 2026

Corresponding Author
Aliyah Alifi
Universitas Negeri Surabaya, Surabaya, Indonesia
aliyah.22064@mhs.unesa.ac.id

1. INTRODUCTION

The global beauty industry has shown significant development in recent years, particularly in the skincare products sector. In Indonesia, the cosmetics industry is projected to experience growth of 7.5% during the 2021–2027 period, making Indonesia one of the fastest-growing cosmetics markets in Asia [1]. Along with this development, e-commerce platforms have become the main means for consumers to purchase products as well as share their usage experiences through online reviews.

Customer reviews on e-commerce platforms contain important information regarding consumers' perceptions of product quality, effectiveness, price, and seller service. This information is part of electronic word of mouth, which can influence other consumers' purchasing decisions [2]. However, the very large number of reviews makes manual analysis inefficient. Moreover, the

star rating system often does not fully represent the content of the reviews because user comments contain more complex information regarding the product usage experience [3].

Sentiment analysis has become a widely used approach to identify users' opinions or attitudes towards a product based on text data. This approach is part of Natural Language Processing, which enables computers to understand and process human language automatically [4]. The development of deep learning technology has produced various language models capable of understanding sentence context more effectively, one of which is IndoBERT, specifically developed for the Indonesian language [5].

Previous research has shown that IndoBERT performs well in sentiment classification. Aras et al. reported that IndoBERT is capable of achieving accuracy of up to 93% in sentiment analysis of product reviews on the Shopee platform [6]. Another study by Jayadianti et al. indicated that a combination of IndoBERT and RCNN could achieve an accuracy of 95.16% on the IndoNLU dataset [7]. Nevertheless, most of these studies only focus on sentiment classification without exploring the relationships between words that appear in customer reviews.

To explore relationships between words in reviews, Association Rule Mining techniques can be used. This method aims to discover patterns of association between items in a dataset using support, confidence, and lift parameters [8]. The Apriori algorithm is one of the frequently used methods for finding frequent itemsets that form association rules [9]. Therefore, this study integrates sentiment analysis using IndoBERT with word association pattern analysis using the Apriori algorithm on Indonesian-language skincare product customer reviews to obtain a more comprehensive understanding of consumer perceptions.

2. METHODS

This study uses the Knowledge Discovery in Databases (KDD) approach to extract knowledge from customer review data of skincare products. The KDD model consists of several main stages, namely data collection, preprocessing, transformation, data mining, and interpretation. This approach is widely used in text mining research because it can process unstructured text data into information that can be analyzed systematically [10].

The research framework used in this study is shown in Figure 1. The research stages begin with the process of collecting customer review data, followed by text data preprocessing, sentiment classification using the IndoBERT model, and word association pattern analysis using the Apriori algorithm.

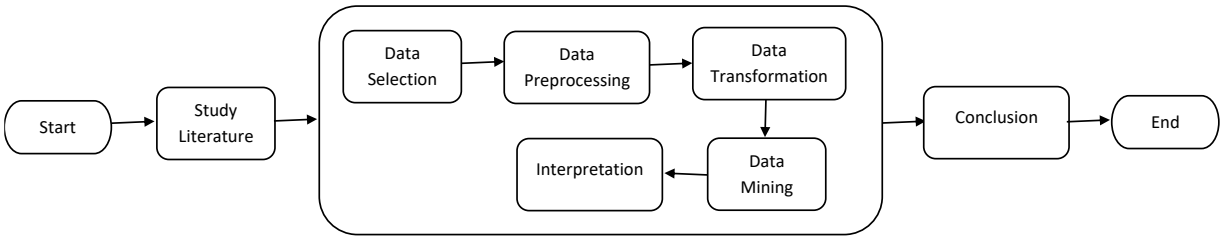

Figure 1. Research Framework using KDD

2.1 Data Collection

The research data was obtained from customer reviews of skincare products on the Shopee e-commerce platform. Data collection was carried out using web scraping techniques to

automatically extract review data from web pages. This technique is widely used in sentiment analysis research because it is capable of collecting large amounts of text data efficiently [11].

The obtained dataset includes several important attributes such as username, review time, product rating, and customer comments. Customer reviews on e-commerce platforms are an important data source because they reflect consumers' experiences and perceptions of the product directly. The collected data is then stored in CSV (Comma Separated Values) format so that it can be further processed using the Python programming language.

2.2 Data Preprocessing

The preprocessing stage aims to clean the dataset of various irrelevant elements so that it can improve the quality of sentiment analysis. Customer review data generally contains various unstructured characters such as symbols, emojis, as well as non-standard language that can interfere with the text analysis process.

The preprocessing stages conducted in this study include:

1. Text Cleaning
   Removing irrelevant characters such as symbols, URLs, numbers, and punctuation marks.
2. Case Folding
   Convert the entire text to lowercase to maintain data consistency.
3. Tokenization
   Breaking down text into word units so that it can be analyzed individually.
4. Normalization
   Converting non-standard or slang words into standard words using an Indonesian language dictionary.
5. Stopword Removal
   Removing common words that do not have a significant contribution in text analysis.

The preprocessing stages are very important in sentiment analysis because they can improve the quality of text representation and the performance of the classification model[12].

2.3 Hybrid Linguistic Labeling

In this study, a hybrid linguistic labeling approach was used to determine sentiment labels on customer reviews. This approach combines product rating information with review text content analysis, enabling it to address ambiguities that often arise in neutral ratings.

For example, a review with a high rating but containing negative words can be re-categorized based on the context of the text. The hybrid labeling approach has been proven to improve the quality of datasets in sentiment analysis based on transformers such as IndoBERT [13].

2.4 Sentiment Classification Using IndoBERT

After the preprocessing process is completed, the next stage is sentiment classification using the IndoBERT model. IndoBERT is a transformer-based language model specifically developed for the Indonesian language, allowing it to understand local linguistic contexts better compared to multilingual models.

This model uses the Bidirectional Encoder Representations from Transformers (BERT) architecture, which enables the system to understand the context of words from both directions

within a sentence. This approach has been proven to improve sentiment classification accuracy compared to conventional machine learning methods [14], [15].

The stages of sentiment classification are carried out through the following steps:
1. Text tokenization using the IndoBERT tokenizer
2. Conversion of text into numerical representations
3. Sentiment label prediction by the model
4. Grouping of sentiment results into positive, neutral, and negative categories

The results of this classification are then used as the basis for analyzing word association patterns in the next stage.

2.5 Association Rule Mining Using Apriori

To find the relationships between words that frequently appear in customer reviews, the Association Rule Mining method is used with the Apriori algorithm. This method is employed to identify patterns of association between items in large datasets through the identification of frequent itemsets.

Association rules are usually expressed in the form of :

$X \rightarrow Y$

Where X and Y are sets of items that have an association relationship in the dataset [16]. The Apriori algorithm uses three main parameters, namely support, confidence, and lift, to evaluate the strength of association rules.

Support

Support indicates how frequently an itemset appears in the dataset.

$$Support\ (A => B) = \frac{The\ number\ of\ transactions\ that\ contain\ (A \cup B)}{Total\ transactions} \quad (1)$$

Support is used to determine whether an itemset is considered frequent or not [17].

Confidence

Confidence indicates the probability of item Y appearing when item X appears.

$$Confidence\ (A => B) = \frac{Support\ (A \cup B)}{Support\ (\mathrm{A})} = \frac{The\ number\ of\ transactions\ that\ contain\ (A \cup B)}{The\ number\ of\ transactions\ that\ contain\ (A)} \quad (2)$$

A high confidence value indicates a strong relationship between two items [18].

Lift

Lift is used to measure the strength of association between two items compared to their random occurrence.

$$Lift\ (A => B) = \frac{Confidence\ (A =>B)}{Support\ (B)} = \frac{Support\ (A \cup B)}{Support\ (A)\ X\ Support\ (B)} \quad (3)$$

A lift value greater than 1 indicates the presence of a strong associative relationship between two items [19].

2.6 Apriori Parameters

The parameters used in this study are shown in the following table.

Table 1. Apriori Parameters

Sentiment	Minimum Support	Minimum Confidence
Positive	0.10	0.80
Negative	0.02	0.90
Neutral	0.10	0.80

The determination of these parameter values aims to produce association rules that have strong and relevant relationships within the dataset.

3. RESULTS AND DISCUSSION

This section presents the analysis results based on the research stages described in the methodology. The analysis process begins with the collection and selection of customer review datasets for skincare products, followed by text preprocessing, data transformation to form ground truth labels using a hybrid linguistic approach, sentiment classification using IndoBERT, and word association pattern analysis using the Apriori algorithm. All these stages are conducted to obtain a more comprehensive understanding of consumer perceptions of skincare products.

3.1 Data Selection

Data collection was conducted through a web scraping process on the e-commerce platform Shopee using the Shopee Reviews Extractor extension on the Chrome browser. The data collected included customer reviews on various skincare products with ratings ranging from one star to five stars.

The crawling process produced an initial dataset of 10,067 customer reviews. This dataset then underwent an initial cleaning stage by removing data that did not have a rating or comment. After the initial cleaning process was carried out, the number of datasets used in the study became 7,933 reviews.

The dataset used in this study consists of several main attributes such as product name, username, review time, rating, and the content of customer comments. The dataset variables used in the study are shown in the following table.

Based on the identification results in the Item column, the research dataset includes 133 types of skincare products from various brands available on the Shopee platform. All of these reviews were then combined into a single dataset for overall analysis.

Table 2. Dataset Variables

Variable	Description
Item	Product name on the review page
Username	Reviewer's username
Time	Time of review upload
Rating	User satisfaction rating on a scale of 1–5
Comment	Review/comment content in text form

3.2 Preprocessing

The customer review dataset obtained from the scraping process still contains various unstructured characters such as symbols, punctuation marks, emojis, and non-standard words. Therefore, a preprocessing stage is required to clean the data before performing sentiment analysis.

An example of the preprocessing results is shown in the following table.

Table 3. Example of Preprocessing Results

Comment	Tekstur:👎👎👎👎 Kandungan:✨✨ Efektivitas:🤦🤦🤦🤦🤦 Alhamdulillah barang sudah sampe rumah dengan selamat, packing nya aman banget.. Sudah jadi langganan disini 😘
cleaning	Tekstur Kandungan Efektivitas Alhamdulillah barang sudah sampe rumah dengan selamat packing nya aman banget Sudah jadi langganan disini
case_folding	tekstur kandungan efektivitas alhamdulillah barang sudah sampe rumah dengan selamat packing nya aman banget sudah jadi langganan disini
tokenization	['tekstur', 'kandungan', 'efektivitas', 'alhamdulillah', 'barang', 'sudah', 'sampe', 'rumah', 'dengan', 'selamat', 'packing', 'nya', 'aman', 'banget', 'sudah', 'jadi', 'langganan', 'disini']
normalization	['tekstur', 'kandungan', 'efektivitas', 'alhamdulillah', 'barang', 'sudah', 'sampai', 'rumah', 'dengan', 'selamat', 'packing', 'nya', 'aman', 'banget', 'sudah', 'jadi', 'langganan', 'disini']
final_clean_text	tekstur kandung efektivitas alhamdulillah barang rumah selamat packing nya aman banget langgan

After the preprocessing stage is completed, the text dataset becomes cleaner and is ready to be used in the data transformation process.

3.3 Data Transformation and Ground Truth Labeling

The data transformation stage aims to create ground truth labels used in the sentiment classification model training process. At this stage, initial labeling is carried out based on product rating values.

The initial labels are determined using the following rules:

Table 4. Initial Distribution Label

Rating	Label	Data Quantity
4 – 5	Positive	3961
3	Neutral	2080
1 – 2	Negative	1279
Total		7320

This distribution shows that the majority of customer reviews have a positive sentiment based on product ratings. The use of ratings as sentiment labels has limitations because they do not always accurately represent the content of the reviews. For example, some reviews with a rating of 3 may contain ambiguous positive or negative opinions.

Therefore, this study employs a Hybrid Linguistic approach to handle reviews with neutral ratings. This approach combines rating information with text content analysis using the IndoBERT model to determine the final sentiment label.

This process produces two types of datasets, namely:

1. Strong labels, which are data with clear ratings representing sentiment
2. Weak labels, which are data with ambiguous ratings that require model classification

The dataset with strong labels is then used as training data in the sentiment classification process using IndoBERT.

3.4 Sentiment Classification Using IndoBERT

The next stage is sentiment classification using the IndoBERT model. This model is employed because it has a better capability to understand Indonesian language context compared to multilingual language models.

The IndoBERT model is trained using a dataset of strong labels that was generated in the data transformation stage. The dataset is then divided into training data and test data to evaluate the model's performance.

Table 5. Training Data and Test Data Distribution (Strong labels)

Type of Data	Data Quantity
Training Data	4.832
Test Data	1.209
Total	**6.041**

The model evaluation results are presented using a confusion matrix.

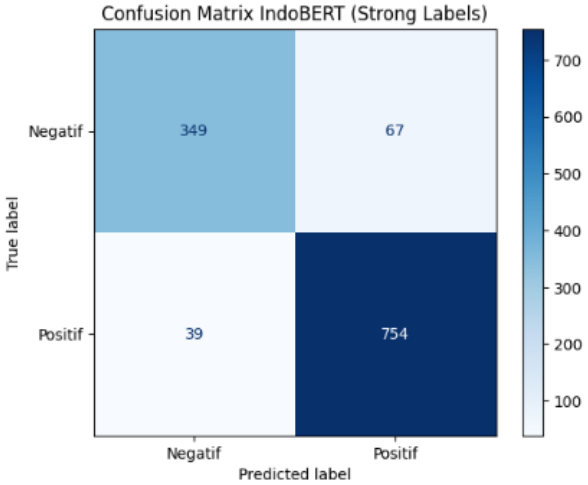

Figure 2. Confusion Matrix Strong Labels

Based on the evaluation results, the IndoBERT model is capable of classifying customer review sentiments contextually. The hybrid linguistic approach also improves the quality of data labeling, especially for reviews with neutral ratings that were previously ambiguous.

The final sentiment distribution after the classification process is shown in the following graph.

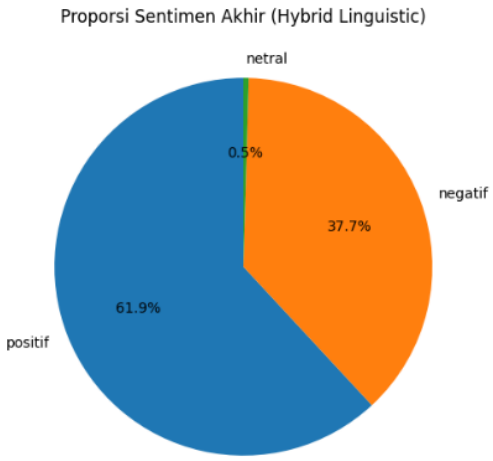

Figure 3. Sentiment Distribution Hybrid Linguistic

These results indicate that the hybrid linguistic approach produces a more representative sentiment distribution compared to rating-based labeling alone.

3.5 Word Association Pattern Using Apriori

After the final sentiment labels are obtained, the dataset is then grouped based on sentiment categories, namely positive, neutral, and negative. Each sentiment group is then analyzed using the Apriori algorithm to find association patterns of words that frequently appear in customer reviews.

The analysis process uses the minimum support and confidence parameters that were determined in the research method stage.

Table 6. Results of Rules per sentiment

Label	antecedents	consequent	support	confidence	lift
Positive	moga	cocok	0.199	0.812	1.67
Negative	ada	enggak	0.022	1.938	14.4
Neutral	barang	bagus	0.121	0.8	5.28

The results of the association rules analysis indicate that each sentiment category has a different word pattern. In positive sentiment, the association rule formed is "moga → cocok" with a support value of 0.199, confidence of 0.81, and lift of 1.67. This pattern shows that positive reviews often contain expressions of hope regarding the suitability of the product for the user's skin condition.

In negative sentiment, the rule "ada → enggak" has a support value of 0.022, confidence of 0.94, and lift of 14.37. This pattern indicates the presence of word associations that describe discrepancies or dissatisfaction with a product or service.

Meanwhile, in neutral sentiment, the rule "barang → bagus" appears with a support value of 0.12, confidence of 0.8, and lift of 5.28. Although the word "bagus" is generally associated with positive sentiment, in the context of neutral reviews, it is often used descriptively without indicating strong satisfaction.

CONCLUSION

This study successfully analyzed customer sentiment in skincare product reviews using the IndoBERT model and identified word association patterns using the Apriori algorithm. The hybrid linguistic approach applied during the labeling stage was able to overcome ambiguity in neutral ratings, thereby producing a more representative ground truth before sentiment classification was performed.

The classification results indicate that IndoBERT is capable of contextually identifying customer review sentiments. Furthermore, the association rule mining analysis successfully discovered patterns of word relationships within each sentiment category, reflecting users' experiences with skincare products. These findings suggest that the integration of sentiment analysis and word association can provide a more comprehensive understanding of consumer perceptions of products.

Future research can develop this approach by using larger datasets or review sources from various digital platforms so that the analysis results have a broader scope.

REFERENCES

[1] M. Ferdinand and W. S. Ciptono, "Indonesia's Cosmetics Industry Attractiveness, Competitiveness and Critical Success Factor Analysis," *Jurnal Manajemen Teori dan Terapan | Journal of Theory and Applied Management*, vol. 15, no. 2, pp. 209–223, Aug. 2022, doi: 10.20473/jmtt.v15i2.37451.

[2] E. H. Muktafin, K. Kusrini, and E. T. Luthfi, "Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing," *Jurnal Eksplora Informatika*, vol. 10, no. 1, pp. 32–42, Sep. 2020, doi: 10.30864/eksplora.v10i1.390.

[3] N. Intan, S. Nabila, and D. Putra, "Consumer Sentiment Analysis Of Viral Skincare Brands: A Literature Review," *The 1st International Student Conference on Economics and Business Excellence (ISCEBE)*, 2024.

[4] W. W. Kamal and C. I. Ratnasari, "Analisis Sentimen Ulasan Produk: Kajian Pustaka," *AUTOMATA*, vol. 2, no. 1, 2021.

[5] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.13357

[6] S. Aras, M. Yusuf, R. Y. Ruimassa, E. A. B. Wambrauw, and E. B. Pala'langan, "Sentiment Analysis on Shopee Product Reviews Using IndoBERT," *Journal of Information Systems and Informatics*, vol. 6, no. 3, pp. 1616–1627, Sep. 2024, doi: 10.51519/journalisi.v6i3.814.

[7] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[8] T. Slimani and A. Lazzez, "Efficient Analysis of Pattern and Association Rule Mining Approaches," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 6, 2014, doi: https://doi.org/10.5815/ijitcs.2014.03.09.

[9] M. Fitriani, G. F. Nama, and M. Mardiana, "Implementasi Association Rule Dengan Algoritma Apriori Pada Data Peminjaman Buku UPT Perpustakaan Universitas Lampung Menggunakan Metodologi CRISP-DM," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 10, no. 1, Jan. 2022, doi: 10.23960/jitet.v10i1.2263.

[10] N. N. Fauziah, "A Multi-Label Text Classification and Association Rule Mining Framework for… A Multilabel Text Classification and Association Rule Mining Framework for Data-Driven Skincare Product Name Generation in E-Commerce," 2025. [Online]. Available: www.ascendiumglobal.org

[11] S. Mahmudah, P. Yanna, and W. Yustanti, "Sentiment Analysis and Topic Modeling Using BERT And LDA Methods (Case Study of Free Meal Program on Twitter)," *Journal of Emerging Information Systems and Business Intelligence*, vol. 7, no. 1, pp. 144–158, 2026.

[12] F. Rafiandi Andhika, W. Witanti, and P. N. Sabrina, "Analisis Sentimen Menggunakan Metode IndoBERT pada Ulasan Aplikasi Zoom Menggunakan Fitur Ekstrasi GloVe," *METIK Jurnal*, vol. 9, no. Vol. 9 No. 2 (2025): METIK Jurnal, p. 2025, 2025, doi: 10.47002/metik.v9i2.1098.

[13] N. Fauzil Adhim and N. Cahyono, “Optimization of IndoBERT for Sentiment Analysis of FOMO on Social Media Through Fine-Tuning and Hybrid Labeling,” 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[14] I. Nyoman Saputra Wahyu Wijaya, K. Agus Seputra, and N. Putu Novita Puspa Dewi, “FINE TUNNING MODEL INDOBERT UNTUK ANALISIS SENTIMEN BERITA PARIWISATA INDONESIA,” *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 22, no. 2, 2025, [Online]. Available: https://www.detik.com/search/searchall?query=wisata&siteid=3&sortby=time&fromdatex=01/01/2022&

[15] S. Apriliani, A. Erfina, and C. Warman, “Fine-Tuned IndoBERT for Aspect-Based Sentiment Analysis of Indonesian Five-Star Hotel Reviews,” *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 14, no. 4, pp. 437–445, Oct. 2025, doi: 10.32736/sisfokom.v14i4.2491.

[16] S. Wu, “An association rule-based approach for frequent item mining of multi-stage access data,” *Discover Computing*, vol. 28, no. 1, Dec. 2025, doi: 10.1007/s10791-025-09644-9.

[17] H. Essalmi and A. El Affar, “Dynamic Algorithm for Mining Relevant Association Rules via Meta-Patterns and Refinement-Based Measures,” *Information (Switzerland)*, vol. 16, no. 6, Jun. 2025, doi: 10.3390/info16060438.

[18] R. Maneiro, M. Amatria, J. L. Losada, G. K. Jonsson, A. Ardá, and I. Iván-Baragaño, “Application of association rules to ball possessions in professional men’s football,” *Front. Psychol.*, vol. 16, 2025, doi: 10.3389/fpsyg.2025.1527437.

[19] B. Sowan, L. Zhang, N. Matar, J. Zraqou, F. Omar, and A. Alnatsheh, “A novel lift adjustment methodology for improving association rule interpretation,” *Decision Analytics Journal*, vol. 15, Jun. 2025, doi: 10.1016/j.dajour.2025.100582.