Journal of Engineering, Electrical and Informatics

E-ISSN: 2809-8706
P-ISSN: 2810-0557

Research Article

Application of TF-IDF and Xgboost Methods for Public Sentiment Analysis Towards Ozzaskin Skincare Brand on Social Media

Mesra Betty Yel [1], Elviwani [2], and Nova Dahliyanti [3], Ahmad Syahran Zidane [4]

1 Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika (STIKOM CKI) Jakarta; Email: mesrabettyyel@stikomcki.ac.id
2 Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika (STIKOM CKI) Jakarta; Email: elviwani@stikomcki.ac.id
3 Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika (STIKOM CKI) Jakarta; Email: nova.dahliyanti@stikomcki.ac.id
4 Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika (STIKOM CKI) Jakarta

* Corresponding Author : mesrabettyyel@stikomcki.ac.id

Abstract: Ozzaskin is a local skincare brand founded by Ustadzah Oki Setiana Dewi that targets Muslim women and focuses on reducing dark spots and acne scars. Over time, this domestic brand has attracted considerable public attention on social media—particularly among mothers—garnering both praise for its product efficacy and criticism regarding price and texture. This study aims to analyze public sentiment toward the Ozzaskin brand by performing web scraping on Instagram and TikTok data, employing TF-IDF for textual feature extraction and XGBoost as the classification algorithm. The findings are expected to provide a comprehensive overview of consumer perceptions of Ozzaskin and to assist the marketing team and product developers in formulating communication strategies and improving product formulas that more effectively address user needs. The novelty of this research lies in the comprehensive application of the TF-IDF + XGBoost framework for brand-related sentiment analysis on Indonesian-language social media.

Keywords: Sentiment Analysis; Natural Language Processing (NLP); Social Media; Ozzaskin Skincare; Public Opinion; TF-IDF.

Received: 28,June,2025;
Revised: 25,July,2025;
Received: 30,July,2025;
Published: 04,August,2025;
Current version: 04,August,2025;

1. Introduction

The beauty and personal care industry in Indonesia has experienced significant growth in recent years, driven by increasing public awareness of skincare and the rapid development of digital technology. According to Euromonitor International (2024), the demand for skincare products continues to increase as consumers become more concerned with maintaining healthy and attractive skin. Along with this growth, social media platforms have become important channels for consumers to exchange information, share experiences, and express opinions regarding products and brands. Social media enables users to create and distribute content interactively, making it a valuable source of consumer-generated information for businesses (Kaplan & Haenlein, 2010).

One local skincare brand that has attracted public attention is Ozzaskin, a skincare brand founded by Oki Setiana Dewi that focuses on products designed to address skin concerns such as dark spots and acne scars (Ozzaskin, n.d.). Through platforms such as TikTok and Instagram, consumers actively share their experiences, reviews, and evaluations of Ozzaskin products. These opinions provide valuable insights into consumer perceptions,

product satisfaction, and brand image. However, the increasing volume of user-generated content makes manual analysis difficult, time-consuming, and potentially subjective.

To address this challenge, Natural Language Processing (NLP) can be utilized to automatically process and analyze textual data. NLP is a branch of artificial intelligence that enables computers to understand, interpret, and process human language in a meaningful manner (Jurafsky & Martin, 2020). One of the most common applications of NLP is sentiment analysis, which aims to identify and classify opinions expressed in textual data into positive, neutral, or negative categories (Liu, 2012). Sentiment analysis allows organizations to gain a deeper understanding of public perceptions and supports data-driven decision-making processes.

Before sentiment classification can be performed, textual data must be transformed into numerical representations that can be processed by machine learning algorithms. One of the most widely used feature extraction techniques is Term Frequency–Inverse Document Frequency (TF-IDF), which evaluates the importance of a term within a document by considering both its frequency and rarity across the document collection (Manning et al., 2008). The resulting feature vectors can then be utilized as input for machine learning models to identify sentiment patterns within social media comments.

Among various machine learning algorithms, eXtreme Gradient Boosting (XGBoost) has gained considerable attention due to its high predictive performance, ability to handle sparse data, and built-in regularization mechanisms that help reduce overfitting (Chen & Guestrin, 2016). Previous studies have demonstrated that XGBoost performs effectively in sentiment classification tasks involving Indonesian-language text and frequently outperforms conventional classification methods such as Naïve Bayes (Alfian et al., 2024; Fathonah & Herliana, 2024). These characteristics make XGBoost a promising approach for analyzing consumer opinions expressed on social media platforms.

Although sentiment analysis has been widely applied in various domains, research focusing on Indonesian local skincare brands remains relatively limited. In particular, studies examining public sentiment toward Ozzaskin using social media data are still scarce. Furthermore, the implementation of TF-IDF and XGBoost for sentiment classification in discussions related to local skincare products has not been extensively explored. Therefore, this study aims to analyze public sentiment toward the Ozzaskin skincare brand using comments collected from social media platforms. The proposed approach utilizes TF-IDF for feature extraction and XGBoost for sentiment classification to identify the distribution of positive, neutral, and negative sentiments expressed by consumers.

The results of this study are expected to provide a comprehensive overview of consumer perceptions toward Ozzaskin and serve as valuable information for marketing teams in developing communication strategies and product improvement initiatives. In addition, this study contributes to the application of sentiment analysis in the Indonesian skincare industry through the implementation of TF-IDF and XGBoost for analyzing consumer opinions derived from social media interactions.

2. Literature Review

Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence that combines computational linguistics, machine learning, and computer science to enable computers to understand, process, and generate human language (Jurafsky & Martin, 2020). NLP has been widely applied in various text mining tasks, including information retrieval, text classification, machine translation, and sentiment analysis. In social media analytics, NLP plays an important role in transforming unstructured textual data into structured representations that can be processed by computational models. Common NLP processes include text cleaning, tokenization, stop-word removal, and stemming, which help improve the quality of textual data before analysis.

Sentiment Analysis

Sentiment analysis is a text mining technique used to identify and classify opinions, emotions, or attitudes expressed in textual content toward a particular object, product, service, or topic (Liu, 2012). The primary objective of sentiment analysis is to determine whether a statement reflects a positive, negative, or neutral sentiment. In the context of social media, sentiment analysis provides valuable insights into public perceptions and consumer opinions by analyzing comments, reviews, and user-generated content. The results can support decision-making processes, marketing strategies, and product development initiatives based on consumer feedback.

Term Frequency–Inverse Document Frequency (TF-IDF)

Term Frequency–Inverse Document Frequency (TF-IDF) is one of the most widely used feature extraction techniques in text mining and information retrieval. TF-IDF measures the importance of a term within a document relative to a collection of documents by combining term frequency and inverse document frequency (Manning et al., 2008). Terms that appear frequently in a specific document but rarely across other documents receive higher weights, making them more influential in representing the document's content. In sentiment analysis, TF-IDF transforms textual data into numerical feature vectors that can be processed by machine learning algorithms for classification tasks.

eXtreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting (XGBoost) is an ensemble machine learning algorithm based on gradient boosting decision trees that is designed to improve predictive performance while maintaining computational efficiency (Chen & Guestrin, 2016). XGBoost incorporates regularization techniques to reduce overfitting and can effectively handle sparse and high-dimensional data, making it particularly suitable for text classification problems. Previous studies have demonstrated that XGBoost achieves competitive performance in sentiment analysis applications and often outperforms conventional machine learning algorithms when combined with appropriate feature extraction methods such as TF-IDF (Alfian et al., 2024; Fathonah & Herliana, 2024).

Social Media as a Source of Public Opinion

Social media platforms have become important channels for communication and information exchange among users. According to Kaplan and Haenlein (2010), social media enables individuals to create, share, and interact with content in virtual communities and networks. The large volume of user-generated content available on social media provides opportunities for researchers to analyze public opinions and consumer perceptions. In this study, social media comments related to Ozzaskin skincare products serve as the primary source of data for sentiment analysis, allowing the identification of consumer attitudes toward the brand based on naturally occurring online discussions

3. Proposed Method

Research Framework

This study employed a quantitative approach to analyze public sentiment toward the Ozzaskin skincare brand based on user comments collected from social media. The research process consisted of several stages, including data collection, text preprocessing, feature extraction using TF-IDF, sentiment classification using XGBoost, and performance evaluation. The overall workflow of the research is illustrated in Figure 1.

Table 1. Dataset Description

Attribute	Description
Platform	TikTok
Data Type	User Comments
Total Comments	500
Language	Indonesian
Domain	Skincare
Object of Study	Ozzaskin

Data Collection

The dataset used in this study consisted of comments related to Ozzaskin skincare products collected from the TikTok platform. Data collection was conducted through web scraping techniques to obtain user-generated comments discussing product experiences, opinions, and evaluations. A total of 500 comments were collected and used as the research dataset. Prior to analysis, the collected data were examined to remove duplicate entries, irrelevant content, spam comments, and comments containing only symbols or emojis.

Table 2. Dataset Description

Action	Expectation	Results
Gather data comment netizen from TikTok	Data succeed collected around 500 entries that representative For analysis sentiment	Data collected 500 comments
Apply TF-IDF + XGBoost for classification sentiment	Model NLP can classify sentiment with accuracy ≥ 85 %, precision & recall ≥ 85 %	Model reach accuracy 89.3 %, precision 88.1 %, recall 87.5 %
Analyze distribution sentiment	Seen pattern domination sentiment positive (> 50 %) And main issues (price, texture, aroma)	Distribution: 60 % positive, 25 % neutral, 15 % negative; issue dominant: effectiveness, price, aroma
Compile recommendation repair product & communication	Obtained recommendation concrete (varian price, aroma, education (use) relevant to analysis results sentiment	Recommendation arranged: package value pack, variants aroma soft, content timeline education usage, monitoring

real-time

To facilitate sentiment classification, a subset of comments was manually labeled into three sentiment categories: positive, neutral, and negative. The labeling process was performed to provide reference labels for the machine learning model during training and evaluation.

Text Preprocessing

Text preprocessing was performed to improve data quality and prepare textual information for further analysis. Several preprocessing stages were applied to the collected comments.

The first stage was text cleaning, which aimed to remove unnecessary elements such as URLs, special characters, punctuation marks, numbers, and emojis. The second stage involved stopword removal to eliminate commonly used words that did not contribute significantly to sentiment identification. Subsequently, stemming was performed to convert words into their root forms, thereby reducing variations in word representation. Finally, tokenization was applied to split each comment into individual tokens that could be processed computationally.

These preprocessing steps were intended to reduce noise within the dataset and improve the effectiveness of feature extraction and classification processes.

Feature Extraction Using TF-IDF

After preprocessing, textual data were transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) method. TF-IDF assigns weights to terms based on their frequency within a document and their distribution across the entire dataset. Terms that frequently appear in a particular comment but rarely occur in other comments receive higher weights and are considered more informative for sentiment classification.

The resulting TF-IDF vectors were used as input features for the machine learning classification model.

Sentiment Classification Using XGBoost

Sentiment classification was conducted using the eXtreme Gradient Boosting (XGBoost) algorithm. XGBoost was selected due to its capability to handle high-dimensional data efficiently and its strong performance in classification tasks. The TF-IDF feature vectors generated during the previous stage were used as input to the classifier.

The dataset was divided into training and testing subsets to evaluate model performance. During the training phase, the model learned sentiment patterns from labeled comments. Subsequently, the trained model was used to predict sentiment categories for unseen testing data. The classification results were categorized into positive, neutral, and negative sentiment classes.

Performance Evaluation

The performance of the proposed model was evaluated using standard classification metrics. Accuracy was used to measure the proportion of correctly classified instances relative to the total number of instances. Precision was used to assess the correctness of positive

predictions, while recall measured the model's ability to identify relevant sentiment classes. In addition, the F1-score was calculated to provide a balanced evaluation of precision and recall.

These evaluation metrics were used to assess the effectiveness of the TF-IDF and XGBoost combination in classifying public sentiment toward the Ozzaskin skincare brand.

4. Results and Discussion

Sentiment Distribution

A total of 500 comments related to Ozzaskin skincare products were collected from the TikTok platform and analyzed using the proposed sentiment analysis framework. After completing the preprocessing stages, each comment was classified into one of three sentiment categories: positive, neutral, or negative.

The sentiment distribution results indicate that the majority of comments belong to the neutral category, accounting for approximately 60.36% of the dataset. This finding suggests that many users expressed opinions, questions, or general statements without showing strong emotional tendencies toward the product. Positive comments represented the second largest category, indicating that a considerable number of consumers were satisfied with their experiences using Ozzaskin products. Meanwhile, negative comments constituted the smallest proportion of the dataset.

The dominance of neutral and positive sentiments reflects a generally favorable public perception of the Ozzaskin brand. However, the relatively small number of negative comments may also contribute to class imbalance, which can influence the performance of machine learning models during classification.

Figure 2. Distribution of Sentiment Classes

XGBoost Classification Performance

Table 3. Performance Evaluation of TF-IDF and XGBoost Model.

Sentiment	Precision	Recall	F1-Score
Negative	1.00	1.00	1.00
Neutral	0.84	0.87	0.85
Positive	0.78	0.74	0.76
Accuracy	colspan	colspan	82%

The sentiment classification process was performed using TF-IDF as the feature extraction method and XGBoost as the classification algorithm. The model was trained and

evaluated using labeled sentiment data to determine its capability in identifying sentiment categories from TikTok comments.

The experimental results show that the initial XGBoost model achieved an overall classification accuracy of approximately 82%. The evaluation results indicate that the model performed well in identifying negative comments, achieving perfect precision, recall, and F1-score values. However, this result should be interpreted carefully because the number of negative samples in the dataset was relatively small.

For the neutral sentiment class, the model achieved a precision of 0.84, recall of 0.87, and F1-score of 0.85. These results demonstrate that the model was effective in recognizing neutral comments, which constituted the majority of the dataset. For the positive sentiment class, the model obtained a precision of 0.78, recall of 0.74, and F1-score of 0.76. Although the performance remained acceptable, positive comments were more difficult to classify correctly compared to neutral comments.

Discussion

The findings demonstrate that the combination of TF-IDF and XGBoost is capable of performing sentiment classification on Indonesian-language social media comments related to skincare products. The achieved accuracy of approximately 82% indicates that the proposed approach can effectively identify sentiment patterns within user-generated content.

The predominance of neutral sentiments suggests that many users engage with Ozzaskin content primarily for information exchange rather than explicitly expressing satisfaction or dissatisfaction. This behavior is commonly observed in social media environments where users frequently ask questions, seek recommendations, or provide short responses without strong emotional indicators.

The classification results also reveal the impact of dataset imbalance on model performance. The limited number of negative comments contributed to exceptionally high classification scores for the negative class while reducing the model's ability to generalize across all sentiment categories. This issue is further reflected in the slightly lower performance observed for positive comments.

Hyperparameter tuning using RandomizedSearchCV produced an accuracy of approximately 81%, which was slightly lower than the baseline model. This outcome suggests that the original model parameters were already close to optimal for the available dataset. Nevertheless, the tuned model is expected to provide better generalization performance because the optimization process considered multiple validation folds during training.

Overall, the results indicate that TF-IDF and XGBoost provide an effective framework for sentiment analysis of social media comments and can support marketing teams in understanding consumer perceptions toward Ozzaskin products. The generated insights can be utilized to improve communication strategies, monitor customer feedback, and support future product development initiatives.

6. Conclusions

This study applied the TF-IDF and XGBoost methods to analyze public sentiment toward the Ozzaskin skincare brand based on comments collected from the TikTok platform. The preprocessing stages, including text cleaning, stopword removal, stemming, and tokenization, successfully transformed raw textual data into structured data suitable for machine learning analysis. TF-IDF was utilized to extract textual features, while XGBoost was employed to classify comments into positive, neutral, and negative sentiment categories.

The experimental results demonstrated that the proposed model achieved an accuracy of approximately 82% in sentiment classification. The findings also revealed that neutral sentiment dominated the dataset, followed by positive sentiment, while negative sentiment

represented the smallest proportion of comments. These results indicate that public perceptions of Ozzaskin on social media tend to be neutral to positive, reflecting generally favorable consumer responses toward the brand.

Overall, the combination of TF-IDF and XGBoost proved effective for sentiment analysis of Indonesian-language social media comments. The generated sentiment insights can support marketing teams in understanding consumer perceptions, evaluating customer feedback, and developing more targeted communication and product improvement strategies. Future studies may improve model performance by increasing the dataset size, balancing sentiment classes, and exploring more advanced feature extraction or deep learning approaches.

References

Alfian, F., dkk. (2024). Klasifikasi sentimen review health-care di Indonesia: Perbandingan XGBoost dan Naïve Bayes. *Jurnal Teknologi Informasi, 15*(1), 23–38.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

Euromonitor International. (2024). *Beauty and personal care in Indonesia*.

Fathonah, F., & Herliana, A. (2024). Analisis sentimen e-commerce Indonesia. *Jurnal Sistem & Informatika, 10*(2), 45–61.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing* (3rd ed.). Prentice Hall.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool.

Macdonald, A., & O'Brien, L. (2014). Measuring public sentiment in social media. *Journal of Information Science, 40*(5), 606–617.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mitchell, R. (2015). *Web scraping with Python: Collecting data from the modern web*. O'Reilly Media.

Ozzaskin. (2025, Juni 15). *About Us*. https://www.ozzaskin.co.id/about

Pappas, N., & Popović, M. (2021). IndoBERT: A pre-trained BERT for Bahasa Indonesia. *ACL Anthology*.