The 2nd International Conference on Multidisciplinary Engineering ( 2nd ICOMDEN 2024)

*The 2nd International Conference on Multidisciplinary Engineering: 00041* (2024) https://doi.org/10.29103/icomden.v2.xxxx
Volume 2, 2024
eISSN: xxxx-xxxx
*Research Original Article*

# Sentiment Analysis of Comments on Youtube Channel Beauty Vlogger in Indonesian Language Using Support Vector Machine Method

**Siti Chairani Siregar✉[1] Rizal Tjuet Adek[2] Zahratul Fitri[3]**
[1] Department of Informatics, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24353, Indonesia, siti.190170102@mhs.unimal.ac.id
[2] Department of Informatics, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24353, Indonesia, rizal@unimal.ac.id
[3] Department of Informatics, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24353, Indonesia, zahratulfitri@unimal.ac.id
✉Corresponding Author: **siti.190170102@mhs.unimal.ac.id | Phone: +6287862480315**

## Abstract
YouTube has become a major platform for beauty vloggers to share product reviews, where user comments provide valuable feedback. This research aims to analyze the sentiment in comments from Indonesian speaking users on beauty vlogger channels, focusing on reviews for powder and skincare products, to capture the positive or negative sentiments. The study utilizes the Support Vector Machine (SVM) method for classification and the TF-IDF weighting technique, analyzing 1,000 comments split into 800 training data and 200 testing data. Sentiment classification was performed post-text preprocessing. The results demonstrate a model accuracy of 97%, with a precision of 98% and recall of 96%, indicating that SVM effectively identifies sentiment in user comments. This system provides valuable insights for beauty vloggers to understand product feedback and contributes to the development of similar applications in other industries.
**Keywords:** YouTube, Beauty Vlogger, Sentiment Analysis, Support Vector Machine, TF-IDF

## Introduction
The rapid advancement of technology has transformed the way people communicate, share information, and express opinions. Among the many social media platforms, YouTube stands out as a leading platform for video content sharing. YouTube not only serves as a source of entertainment but also functions as a powerful tool for product reviews and tutorials, particularly in the beauty industry. Many content creators, commonly known as beauty vloggers, produce videos to provide makeup tutorials, product recommendations, and personal insights into various beauty products. Their reviews play a crucial role in influencing consumer decisions, making YouTube a dynamic space for branding and consumer interaction.

As consumers engage with beauty vlogs by leaving comments, these interactions provide invaluable feedback, reflecting public opinion on the products discussed. However, manually analyzing the large volume of user comments can be time consuming and challenging. To address this challenge, sentiment analysis offers a systematic approach to understanding user opinions by classifying comments as either positive or negative. Sentiment analysis can reveal trends in consumer perceptions, enabling beauty vloggers and companies to better understand market needs, improve product offerings, and enhance customer engagement.

This study focuses on analyzing user comments related to makeup and skincare products discussed on Indonesian beauty vlogger channels. The goal is to determine whether user sentiments expressed in the comments are positive or negative. Given the unstructured nature of user comments often filled with slang, abbreviations, and emoticons text preprocessing plays a critical role in standardizing the data. This study employs the Support Vector Machine (SVM) algorithm, a widely used classification method in text mining due to its ability to handle both linear and non-linear data. The TF-IDF (Term Frequency-Inverse Document Frequency) technique is used to convert text into numerical features, helping the SVM algorithm focus on the most relevant terms for classification.

Several previous studies have demonstrated the effectiveness of SVM in sentiment analysis tasks. For instance, Anshari et al. (2023) applied SVM to analyze sentiment from digital payment reviews, achieving over 90% accuracy. Similarly, Rizal et al. (2019) reported successful classification of facial images using SVM with a true detection rate of 90%. These findings highlight the potential of SVM for sentiment analysis across various domains, including social media and e-commerce. However, this study specifically explores the application of SVM in the context of beauty

product reviews, an area with significant commercial and social importance.

This research is motivated by the increasing need for automated tools to analyze public opinion in the beauty industry. While manual analysis of YouTube comments is impractical for large datasets, machine learning models like SVM offer a scalable and efficient solution. The primary objectives of this study are:

1. To design a sentiment analysis system capable of classifying comments into positive or negative categories.
2. To evaluate the performance of SVM in classifying sentiments related to beauty products.

Through this research, we aim to provide insights into public perceptions of makeup and skincare products reviewed on Indonesian beauty vlogger channels. These insights are expected to help content creators understand user feedback more effectively and contribute to the development of similar sentiment analysis applications in other industries.

In the following sections, we discuss the literature that supports our study, the methodology employed for data collection and analysis, the results obtained, and the implications for future research and practical applications.

## Literature Review

Sentiment analysis, a branch of Natural Language Processing (NLP), focuses on extracting opinions or sentiments from text data, making it a critical tool in applications such as product reviews, customer feedback, and social media analysis. In this study, sentiment analysis is applied specifically to analyze user comments on Indonesian beauty vlogger channels, targeting reviews of powder and skincare products to assess public perception.

### 1. Sentiment Analysis on Social Media

Social media platforms like YouTube serve as rich data sources for understanding public opinion due to the vast and diverse nature of user comments. Sentiment analysis on these platforms can reveal user attitudes toward content or products, offering actionable insights for brands, content creators, and marketers. Munthe et al. (2021) emphasized that sentiment analysis enables a structured approach to understanding opinion patterns, highlighting how digital platforms facilitate nuanced audience insights. This capability supports strategic decision-making, allowing brands and influencers to refine content strategies based on audience sentiment and trends in user feedback.

### 2. TF-IDF for Text Weighting

Term Frequency-Inverse Document Frequency (TF-IDF) is a widely adopted technique for weighting text in sentiment analysis. By assigning numerical values to words based on their occurrence within a specific document relative to other documents, TF-IDF converts text data into a structured format that enhances the relevance of unique words. High TF-IDF values are given to words that appear frequently in individual documents but are rare across the corpus, thus emphasizing sentiment-bearing terms and reducing the weight of commonly used words. Sya'bani and Umilasari (2018) demonstrated that TF-IDF effectively highlights significant terms, making it an ideal feature extraction method for sentiment analysis. This ensures the model focuses on words that are contextually relevant, optimizing performance by enhancing its sensitivity to sentiment-laden terms.

### 3. Support Vector Machine (SVM) in Sentiment Analysis

Support Vector Machine (SVM) is a robust machine learning algorithm commonly used for text classification due to its efficacy in handling high dimensional data and achieving high accuracy. SVM operates by identifying an optimal hyperplane that separates data into different classes, making it highly suitable for binary classification tasks like sentiment analysis. The hyperplane concept in SVM plays a crucial role, as it maximizes the margin between data points from different classes, enhancing the classifier's accuracy and robustness.

1. The optimal hyperplane in SVM is defined by the following equations:
   Primary Equation of the Hyperplane:
   $(x) = w \cdot x + b$ (1)
   where w is the weight vector orthogonal to the hyperplane, x represents the input data point, and b is the bias term that adjusts the position of the hyperplane.
2. Alternate Equation using Kernel Functions:
   $(x) = \sum m\ aiyik(x, xi) + b$ (2)
   Where:
   αi represents the weight assigned to each support vector,
   yi is the label of each training sample,
   K(x,xi) is the kernel function, mapping data to a higher-dimensional space if needed, and
   b is the bias parameter.
   Including an illustration of the SVM hyperplane below will clarify the separation between positive and negative classes. This visual representation, along with the equations, helps to understand how SVM effectively divides sentiment data into distinct categories, ensuring accurate classification.

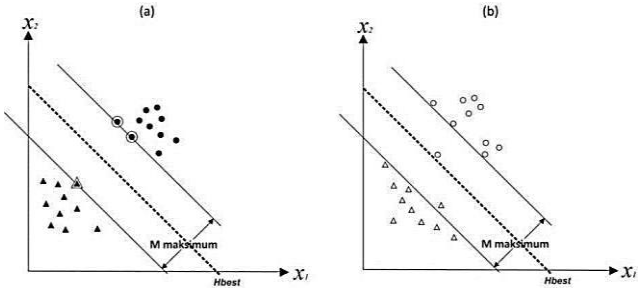


**Figure 1.** Best hyperlane and maximum margin

**4. Related Work on YouTube Comment Sentiment Analysis**

Previous research has explored sentiment analysis on YouTube comments, though specific studies on beauty vlogger content remain limited. Rifqi (2016) conducted a comparative study between Naive Bayes and SVM for analyzing YouTube comments, finding that SVM with unigram features provided higher accuracy than Naive Bayes. This study underscores the suitability of SVM for analyzing sentiment in YouTube comments, especially when combined with TF-IDF weighting, as this pairing effectively isolates relevant sentiment-related terms. The success of SVM in this context suggests it can handle the unique linguistic structures and emotive expressions often found in social media comments, particularly in the beauty and lifestyle domains.

**5. Summary**

The literature review affirms that sentiment analysis using TF-IDF and SVM is a powerful approach for classifying social media text. Prior studies have shown that TF-IDF effectively emphasizes meaningful words, providing a structured and weighted representation of text that enhances model focus on sentiment driven terms. Meanwhile, SVM consistently achieves high accuracy and is well suited to the classification of sentiment laden content, especially within social media and user generated comments. By applying these methods to YouTube comments on beauty vlogger channels, this study aims to capture public sentiment around beauty products, contributing insights that are valuable for content creators and brands aiming to engage more effectively with their audience.

## Materials & Methods

This study employs a systematic methodology encompassing data collection, preprocessing, and sentiment classification using Support Vector Machine (SVM) to analyze comments on Indonesian beauty vlogger YouTube channels. The primary goal is to accurately classify the sentiment of these comments, specifically those related to powder and skincare products.

Figure 1 below outlines the flow of the methodological steps followed in this study.

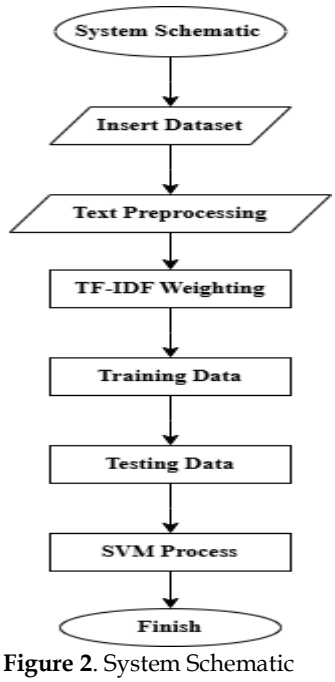


**Figure 2**. System Schematic

**1. Data Collection**

Data collection was conducted by gathering relevant comments from videos posted by Indonesian beauty vloggers, with a specific focus on videos reviewing powder and skincare products. The dataset consists of 1,000 user comments, providing a balanced representation of positive and negative sentiments. 800 comments were designated as training

data, and 200 comments as testing data, following standard practices in supervised machine learning for effective model evaluation and training. Each comment underwent manual labeling to assign it a sentiment label positive or negative. This manual labeling process was essential to establish a high quality ground truth, allowing the model to learn sentiment characteristics specific to Indonesian language patterns and terms common in beauty related discussions.

**2. Text Preprocessing**

Text preprocessing is essential for converting raw, unstructured text data into a clean and structured format that is suitable for machine learning models. The steps performed in the preprocessing phase include:

a. Case Folding All text was converted to lowercase to maintain uniformity throughout the dataset. This step ensures that differences in capitalization do not affect the model's understanding, enabling consistent word recognition.
b. Tokenizing Comments were split into individual words or "tokens," isolating each word or symbol for independent analysis. Tokenizing enables the model to analyze each word separately, simplifying the feature extraction process.
c. Stopword Removal: Commonly used words such as "and," "or," "the," "in," and similar words that do not carry significant sentiment are removed. These "stopwords" often add noise to the dataset and can reduce model performance by diluting the influence of more sentimentally relevant words.
d. Stemming Words were reduced to their base or root forms to standardize variations, making different forms of a word equivalent. For instance, "liked" and "likes" were both converted to "like." This step reduces vocabulary size and ensures that variations of the same word are treated as a single feature, thus improving model consistency and reducing redundancy.

Through these preprocessing steps, the dataset is transformed into a cleaner and more concise format, emphasizing sentiment-bearing terms and eliminating unnecessary noise.

**3. TF-IDF Weighting**

After text preprocessing, the Term Frequency-Inverse Document Frequency (TF-IDF) method is applied to convert the text data into a numerical format suitable for SVM classification. This method assigns a weight to each word based on two factors:

Term Frequency (TF): This measures the frequency of a word within a specific comment. Higher TF values indicate that a word appears frequently within a comment, potentially signaling relevance to the sentiment.

Inverse Document Frequency (IDF): IDF reduces the weight of words that commonly appear across many comments, highlighting terms unique or sentimentally significant to individual comments. Thus, words that occur frequently in the dataset but lack sentiment value (e.g., "also," "because") receive a lower weight.

**4. Sentiment Classification with SVM**

SVM is used to classify the comments as positive or negative based on their TF-IDF values. The model learns to distinguish sentiment by finding an optimal hyperplane that maximally separates the classes. The training set is used to build the model, and the testing set is used to evaluate its accuracy.

**5. Evaluation Metrics**

The model's performance is evaluated using:

a. Accuracy: The percentage of correctly classified comments.
b. Precision: The ratio of correctly predicted positive comments to total predicted positives.
c. Recall: The ratio of correctly predicted positive comments to all actual positives.

## Results and Discussion

The sentiment analysis conducted on 1,000 comments from Indonesian beauty vlogger YouTube channels, specifically focusing on reviews of powder and skincare products, leveraged a Support Vector Machine (SVM) model with Term Frequency-Inverse Document Frequency (TF-IDF) weighting. The model's evaluation was based on key metrics of accuracy, precision, and recall, which are summarized in Table 1.

**Table 1**. Model Performance Metrics

| Metric | Score |
|---|---|
| Accuracy | 97% |
| Precision | 98% |
| Recall | 96% |

The accuracy of 97% indicates that the model can consistently classify comments as positive or negative with minimal error. This high accuracy underscores the effectiveness of SVM in identifying sentiment-related patterns within a relatively diverse dataset**.** precision, recorded at 98%, suggests that when the model labels a comment as positive or negative, it is highly likely to be correct, thus minimizing false positives. Likewise, the recall rate of 96% confirms that the model is successful in identifying relevant comments within each category, indicating a low occurrence of false negatives.

**1. TF-IDF Formula and Importance in Classification**

To enhance classification accuracy, the model used TF-IDF weighting to assign importance to words based on their relevance in each comment. The TF-IDF formula combines Term Frequency (TF), which reflects how often a term appears within a comment, with Inverse Document Frequency (IDF), which reduces the weight of terms that appear frequently across many comments but are less specific to individual sentiments. This ensures that the model can effectively focus on sentiment rich words. The formulas used are as follows:

Term Frequency (TF): This measures the frequency of a word t within a document d:

$$TF(t,d)= \frac{f(t,d)}{\sum t' \in d f(t',d)} \quad (3)$$

Inverse Document Frequency (IDF): This decreases the weight of words that are commonly found across documents:

$$IDF(t)= \log\left(\frac{N}{df\,(t)}\right) \quad (4)$$

where N is the total number of documents, and df(t)is the count of documents containing the term t.

TF-IDF Calculation: The final TF-IDF weighting is a product of the two metrics:

$$TF\text{-}IDF(t,d)=TF(t,d)\times IDF(t) \quad (5)$$

This approach allows the model to identify and prioritize sentiment-bearing terms, ultimately increasing the classification performance.

**2. SVM Formula**

Support Vector Machine (SVM) is employed here to maximize separation between positive and negative comments. The model uses a linear kernel to establish a hyperplane that effectively divides data points into two sentiment classes. The SVM decision function is represented as follows:

$$f(x)=sign(i=1\sum n\alpha iyiK(xi,x)+b) \quad (6)$$

where:

x is the input vector,

xi are the support vectors,

yi are the labels of the support vectors (+1 for positive, -1 for negative),

αi are the weights, determined by the SVM algorithm,

K(xi,x) is the kernel function, and

b is the bias term

In this study, a linear kernel $K(xi,x)=xi\cdot x$ was used, as it provides an effective and efficient classification boundary for sentiment data.

**3. System Testing Results**

System testing involved inputting comments to evaluate their sentiment. The model's high performance across key metrics accuracy, precision, and recall demonstrates its robustness in identifying sentiment in diverse comment data. These findings validate the suitability of SVM for sentiment analysis tasks, highlighting its potential applications across other domains where consumer feedback analysis is valuable. The model thus offers beauty vloggers and related content creators an insightful tool to gauge public opinion on their reviewed products accurately.

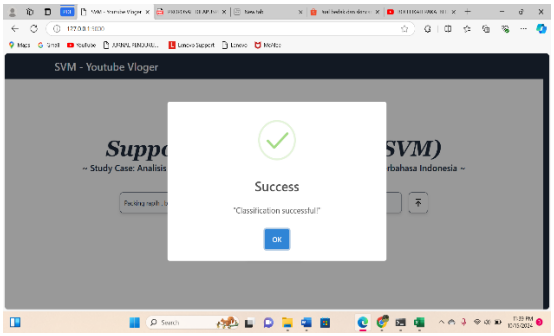


**Figure 3**. system test result image

## Conclusions

This study demonstrates the effectiveness of using Support Vector Machine (SVM) with TF-IDF weighting for sentiment analysis on YouTube comments about beauty products. Using a dataset of 1,000 comments (800 training and 200 testing), the model achieved an accuracy of 97%, precision of 98%, and recall of 96%, proving that SVM is highly effective in distinguishing positive and negative sentiments in Indonesian-language comments.

The high accuracy and reliability of this model provide valuable insights for beauty vloggers and brands, enabling them to understand user opinions on powder and skincare products more accurately. This approach not only enhances

brand responsiveness to consumer feedback but also shows potential for similar applications in other industries requiring sentiment analysis.

Future research may explore the application of other machine learning techniques or hybrid approaches to further enhance sentiment detection accuracy and adapt the system to various social media platforms.

## References

[1] Adi, S., & Wintarti, A. (2022). Comparison Of Support Vector Machine (Svm), K-Nearest Neighbors (Knn), And Random Forest (Rf) For Heart Disease Prediction. *Mathunesa: Jurnal Ilmiah Matematika*, 10(2), 258–268. Https://Doi.Org/10.26740/Mathunesa.V10n2.P258-268.

[2] Anshari, R. A. L., Alam, S., & Hafid, M. T. (2023). Digital Payment Sentiment Analysis On Google Play Store Reviews Using Support Vector Machine. *Jurnal Ilmiah Teknik Dan Ilmu Komputer*, 2(3), 118–128.

[3] Bei, F., & Sudin, S. (2021). Sentiment Analysis Of Online Ticket Applications In The Play Store Using Support Vector Machine (Svm). *Sismatik*, 1(1), 91–97.

[4] Munthe, M. P., Ansori, A. S. R., & Fauziyah, N. (2021). Sentiment Analysis On Indonesian Language Food Vlogger Comments On Youtube Using Naïve Bayes Algorithm. *Eproceedings Of Engineering*, 8(6), 11909–11916. Https://Openlibrarypublications.Telkomuniversity.Ac.Id/Index.Php/Engineering/Article/View/16897.

[5] Rizal, R. A., Girsang, I. S., & Prasetiyo, S. A. (2019). Face Classification Using Support Vector Machine (Svm). *Remik (Riset Dan E-Jurnal Manajemen Informatika Komputer)*, 3(2), 1. Https://Doi.Org/10.33395/Remik.V3i2.10080.

[6] Sakti, S. P. H., & Adam, M. (2020). Application Of Svm Algorithm In Data Mining For Learning Outcome Analysis (Case Study: Software Engineering). *Seminar Nasional Informatika*, 2020, 291–300.

[7] Switrayana, I. N., Ashadi, D., Hairani, H., & Aminuddin, A. (2023). Sentiment Analysis And Topic Modeling Of Kitabisa Application Using Support Vector Machine (Svm) And Smote-Tomek Links Methods. *International Journal Of Engineering And Computer Science Applications*, 2(2), 81–91. Https://Doi.Org/10.30812/Ijecsa.V2i2.3406.

[8] Sya'bani, M. M., & Umilasari, R. (2018). Application Of Cosine Similarity And Tf/Idf Weighting For Synopsis Classification In Jember District Library. *Justindo (Jurnal Sistem Dan Teknologi Informasi Indonesia)*, 3(1), 31–42. Http://Jurnal.Unmuhjember.Ac.Id/Index.Php/Justindo/Article/View/2345.

[9] Mustakim, H., & Priyanta, S. (2022). Aspect-Based Sentiment Analysis Of Kai Access Reviews Using Nbc And Svm. *Ijccs (Indonesian Journal Of Computing And Cybernetics Systems)*, 16(2), 113. Https://Doi.Org/10.22146/Ijccs.68903.

[10] Permana, D. S., & Silvanie, A. (2021). Heart Disease Prediction Using Support Vector Machine And Python Based On Cleveland Database. *Junif: Jurnal Nasional Informatika*, 2(1), 29–34.