Comparison of Naïve Bayes and Support Vector Machine for Sentiment Classification of Acne Skincare Reviews

Alti Arindika [1]*, Majid Rahardi [2]*
* Informatika, Universitas AMIKOM Yogyakarta
altiarindikaa@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-11-28
Revised 2025-12-29
Accepted 2026-01-07

Keyword:

Acne,
Classification, Naïve Bayes,
Sentiment,
Support Vector Machine.

ABSTRACT

The increasing popularity of skincare products for acne-prone skin had led to a surge in online consumer reviews, which are characterized by informal language, domain-specific terminology, and imbalanced sentiment distribution, posing challenges for sentiment classification tasks. This study aims not only to compare the performance but also to analyze the generalization behavior of two popular machine learning algorithms, Naïve Bayes and Support Vector Machine (SVM), for sentiment classification of skincare product reviews specifically targeting acne-prone skin. A comprehensive methodology was employed, including thorough text preprocessing, feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) with n-gram representation, and data balancing through Synthetic Minority Over-sampling Technique (SMOTE). The study utilized a dataset of 4,004 labeled reviews categorized into positive and negative sentiments. The models were evaluated using stratified 5-Fold cross-validation to ensure robust and fair assessment. Results indicate that Naïve Bayes slightly outperforms SVM on the testing set, achieving the highest accuracy of 91.14% compared to 90.64% for SVM. While SVM demonstrated higher performance during training, its testing performance suggested a tendency toward overfitting, whereas Naïve Bayes exhibited more stable generalization on unseen data. Further qualitative insight analysis revealed that product effectiveness and user experience are the primary drivers of consumer sentiment, while competitive analysis highlighted distinct brand perception patterns across skincare categories. These findings indicate that simpler probabilistic models such as Naïve Bayes can provide robust and reliable performance for sentiment analysis in specialized and imbalanced skincare review datasets.

I. Introduction

In this era of rapid digital economic growth, online reviews have become a key reference for consumers when making purchasing decisions[1]. *E-commerce* platforms such as Shopee in Indonesia have become one the main media for consumers to share their experiences and rate the products they buy. Reviews available in the form of text and star ratings not only help potential buyers[2] in choosing products, but also serve as a valuable source of information for sellers and manufacturers to improve the quality of their products and services[3]. With millions of transactions and reviews continuing to grow, review data analysis has become crucial in understanding consumer preferences and satisfaction in greater depth.

Although star ratings are a common indicator for evaluating products, this system is often ambiguous and does not reflect the overall quality of the product[4]. For example, a customer may give a 5-star rating simply because of fast delivery and neat packaging, even though the skincare product is not suitable for their skin and causes problems such as irritation or acne. This ambiguity is particularly critical for high-risk products like acne-prone skincare, where the product's effectiveness is highly subjective and depends on individual skin conditions, active ingredients, and specific usage cycles. Unclear interpretations of star ratings can mislead consumers and hinder manufacturers in identifying aspects of the product that need to be addressed[5].

Sentiment analysis has become a commonly used method to overcome the limitations of star ratings by extracting opinions from review texts in more detail[6]. Various studies have applied algorithms such as Naïve Bayes and Support Vector Machine (SVM) to classify product review sentiment[7]. For example, the study[8] compares Naïve Bayes and SVM on skincare reviews on the Female Daily platform and finds that SVM provides higher accuracy in capturing complex sentiment patterns. Additionally, [9] tests various SVM kernels with TF-IDF features and shows that the linear kernel excels in training time. The study[10] uses transformer-based models such as RoBERTa for sentiment analysis on YouTube educational videos, demonstrating the potential of modern methods in improving sentiment classification accuracy. However, most of these studies still combine product and service sentiment into one category without explicitly separating them, making it difficult to identify the actual product aspects that influence consumer satisfaction[11].

Reviews of acne-prone skincare products exhibit linguistic characteristics that differ substantially from those of general product reviews, particularly in Indonesian-language e-commerce platforms. User-generated texts frequently incorporate domain-specific dermatological terms (e.g., breakout, iritasi, purging) alongside informal and highly contextual beauty-related expressions (e.g., cocok di aku, bruntusan, nggak ngaruh), reflecting spontaneous, experience-driven modes of consumer communication[12]. Such non-standard vocabulary usage, lexical variation, and creative word forms increase semantic sparsity and lexical inconsistency within textual data, thereby complicating feature representation in bag-of-words-based models such as TF-IDF[12].

Moreover, skincare reviews are often conveyed through short evaluative phrases and domain-specific lexical items whose sentiment orientation is highly context-dependent [13]. Prior studies adopting lexicon-based sentiment analysis approaches have shown that general-purpose sentiment lexicons are insufficient for capturing informal expressions and domain-specific terminology commonly found in skincare reviews, limiting both interpretability and classification reliability in this domain [13]. These limitations highlight the challenges faced by conventional text representation techniques when applied to specialized consumer review data.

In addition, reviews within the acne skincare domain frequently contain ambiguous sentiment expressions in which positive user experiences coexist with negative outcomes, resulting in mixed or conflicting sentiment polarity within a single review instance [14]. This phenomenon challenges traditional sentiment classification models that assume a single dominant polarity and may contribute to misclassification, particularly in imbalanced datasets. Consequently, neglecting these domain-specific linguistic characteristics constrains the robustness, interpretability, and generalizability of sentiment classification models applied to acne-prone skincare reviews.

Therefore, this study fills these gaps by developing an integrated sentiment classification framework tailored for the acne skincare industry. This research goes beyond a simple algorithm comparison by: (1) incorporating an aspect-based analysis (Effectiveness, Experience, Service, Price) to address the specific linguistic nuances mentioned above; (2) evaluating the robustness of Naïve Bayes and SVM when trained on SMOTE-balanced data in a specialized domain [15];[16] and (3) integrating the classification results directly into a competitive brand mapping to validate how machine learning outcomes reflect real-world market perceptions of leading brands on Shopee. By utilizing Stratified 5-Fold Cross-Validation, this study provides a more reliable assessment of model stability and offers actionable strategic insights that bridge the gap between technical classification and industrial marketing applications.

II. Methodology

This research was conducted systematically through several stages. The methodology included data collection and semi-automatic data labeling, extensive text pre-processing, feature extraction, data balancing, and finally, model training and evaluation. These stages are described in more detail below, and the research stages are illustrated in Figure 1.

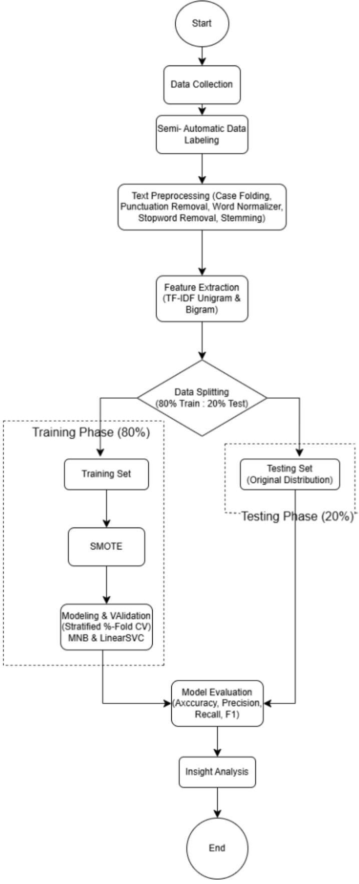
Start
Data Collection
Semi- Automatic Data Labeling
Text Preprocessing (Case Folding, Punctuation Removal, Word Normalizer, Stopword Removal, Stemming)
Feature Extraction (TF-IDF Unigram & Bigram)
Data Splitting (80% Train : 20% Test)
Training Phase (80%)
Training Set
SMOTE
Modeling & VAlidation (Stratified %-Fold CV) MNB & LinearSVC
Testing Set (Original Distribution)
Testing Phase (20%)
Model Evaluation (Axccuracy, Precision, Recall, F1)
Insight Analysis
End

Figure 1. Research Flow

A. *Data Collection*

The dataset used in this study consists of *ratings, usernames, brands,* types of skincare, and reviews of *skincare* products for acne-prone skin[12], which were collected from the Shopee marketplace using a JavaScript console scraping method. The initial collection yielded a total of 7,231 Indonesian-language reviews representing three prominent local skincare brands Npure, Wardah, and Emina, across four essential product categories: face wash, toner, moisturizer, and sunscreen. These specific brands and categories were selected based on their high transaction volumes and significant consumer engagement within the acne-prone skincare segment on the platform.

After the initial collection, a rigorous preprocessing and filtering phase was conducted to ensure data quality, resulting in 5,531 reviews that contained valid textual content. The sentiment distribution of this filtered data comprised 3,498 positive, 506 negative, and 3,228 neutral reviews. To focus the model's performance on clear polarity discrimination, neutral reviews were excluded from the classification phase. Consequently, the final labeled dataset used for model training and evaluation consists of 4,004 instances (Positive and Negative), providing a robust foundation for comparing the performance of Naïve Bayes and Support Vector Machine (SVM) algorithms.

B. *Semi-Automatic Data Labeling*

The semi-automatic data labeling process is carried out using a product-centric framework. First, reviews aspects including Effectiveness, Experience, Service, Price or Irrelevant are identified using a rule-based method with a keyword dictionary. Second, review sentiment is classified as Positive, Negative, or Neutral using the RoBERTa AI model[17]. At the modelling stage, only data labelled Positive and Negative is used, while data labeled Neutral is filtered out.

C. *Text Preprocessing*

This preprocessing stage aims to clean and standardize data from text analysis. This process is very important in removing noise and converting text into a structured format[18], thereby facilitating the assessment of model classification accuracy. In the study, four main steps were applied sequentially in the preprocessing of each review. The steps are described as follows:

1. *Case Folding* aims to covert all text to lowercase[19]. This step is crucial to ensure data consistency. Even if there are capital letters, words can appear identical due to variations in letter usage. For example, “Cocok” and “cocok” will be processed identically.
2. *Punctuation Removal* involves removing all non-alphanumeric characters, including punctuation marks, numbers, and symbols[20]. The goal is to focus the analysis solely on words.
3. *Word Normalizer* is used to correct words with typos, abbreviations, and non-standard language[21].
4. *Stopword Removal* is the process of removing common words that appear frequently but do not have significant sentiment weight[22], such as ‘yang’, ‘di’, ‘juga’. This process uses a predetermined list of Indonesian *stopwords* to reduce noise in the data.
5. *Stemming* is a process that converts each word to its root form, for example ‘mencerahkan’ becomes ‘cerah’. This is done reduce the number of unique features and ensure that words with the same root are treated as entities. This step is implemented using the Sastrawi library[23].

After going through the five preprocessing steps, the clean and normalized review text is ready for the next stage, which is feature extraction using the TF-IDF method.

D. *Feature Extraction*

After pre-processing, the cleaned was transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) vectorization technique. TF-IDF assigns weights to terms based on their frequency within individual documents relative to their distribution across the entire corpus, enabling the identification of words that are characteristic of specific reviews [24]. In this study, the vectorizer was configured to incorporate both unigrams and bigrams (ngram_range = (1, 2)) [25], with the feature space limited to a maximum of 5,000 features to maintain computational efficiency.

The selection of TF-IDF with n-gram representation was motivated by its interpretability, robustness, and proven effectiveness for short-text classification tasks in Indonesian-language datasets. This configuration is specifically designed to address the linguistic challenges and contextual nuances of skincare reviews discussed in Section I. By utilizing bigrams, the model can capture critical negation patterns (e.g., "tidak cocok", "nggak ngefek") and specific dermatological emphatic expressions (e.g., "iritasi parah", "breakout parah") that unigrams often fail to distinguish. Furthermore, the inherent transparency of TF-IDF weights allows for a clearer analysis of lexical contributions. This supports the methodology for identifying product aspects such as Effectiveness and Experience by enabling the extraction of high-weight terms that define each category, thus bridging the gap between technical classification and qualitative consumer insights. Prior studies have demonstrated that TF-IDF with optimized n-gram configurations remains a strong baseline for classical machine learning classifiers, particularly when combined with feature tuning and data balancing strategies on limited datasets [26]. While embedding-based approaches may capture deeper semantic relationships, TF-IDF supports the analytical objectives of this study by maintaining high interpretability for competitive brand mapping.

E. SMOTE

To address the imbalance issue, the *Synthetic Minority Over-Sampling Technique* (SMOTE) was employed[27]. SMOTE generates synthetic samples of the minority class through interpolation in the feature space, thereby improving class balance during model training. However, in text classification tasks based on TF-IDF representations, synthetic instances are created in a high-dimensional vector space rather than at the lexical level, which may limit their direct linguistic interpretability.

Despite this limitation, previous studies have demonstrated that SMOTE remains effective for improving classification performance on imbalanced textual datasets when applied at the feature level [28]. To prevent data leakage and preserve evaluation validity, the dataset was first split into 80% training data and 20% testing data. SMOTE was then applied exclusively to the training set, while the testing set was retained in its original distribution as an unbiased benchmark. This methodological choice represents a trade-off between enhancing class balance and maintaining linguistic realism, which is commonly adopted in imbalanced text classification studies.

F. Modeling and Evaluation

After data balancing, the study compared two machine learning models to determine the most effective approach for sentiment classification on the acne-prone skincare dataset:

1. Multinomial Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem with the assumption of conditional independence among features. This model is particularly well suited for discrete feature representations such as term frequency or TF-IDF vectors and is known for its computational efficiency and robustness when handling sparse text data [29]. In this study, MNB was implemented using default Laplace smoothing to address zero-frequency issues in the feature space.
2. Linear Support Vector Machine is a margin-based, non-probabilistic classifier that constructs an optimal separating hyperplane between classes. The linear kernel was selected due to its proven effectiveness in high-dimensional and sparse feature spaces generated by TF-IDF representations, as well as its lower computational complexity compared to non-linear kernels [30]. The regularization parameter was set to its default value (C = 1.0) to ensure a fair comparison with MNB under consistent experimental conditions.

To ensure robust performance analysis, the model evaluation was conducted in two distinct phases:

1. Validation Phase: Stratified 5-Fold Cross-Validation scheme[31] was applied exclusively to the training data. This scheme ensures that each fold maintains the same class distribution as the original dataset, providing a reliable assessment of model stability.
2. Testing Phase: The final performance was measured using the held-out 20% testing data. This phase evaluated accuracy, precision, recall, and F1-Score, ensuring that the reported results reflect the model's ability to generalize to new, unseen data.

$$Accuracy = \frac{TP + TN}{TP+FP+TN+FN} \quad (1)$$

Accuracy measures the proportion of correct predictions out of all observations. It is easy to interpret and is a common initial benchmark, but it can be misleading on imbalanced data. Accuracy can be deceptive because the model can "win" simply by guessing the majority class. Therefore, accuracy should be interpreted alongside other metrics for a comprehensive view.

$$Precision_{macro} = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FP_i} \quad (2)$$

Precisionmacro assesses the accuracy of predictions for each class and then averages them across classes (K). This is because each class has an equal weight, rather than a weight based on the number of samples. This metric is sensitive to performance in minority classes, which is very useful when sentiment distribution is unbalanced.

$$Recall_{macro} = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i} \quad (3)$$

Recallmacro measures the completeness of detection for each class across all classes. The macro average ensures that each class influences the score equally. A high recall value is important when false negatives (FN) are more risky, for example, when failing to capture negative reviews impacts customer service.

$$F1_{macro} = \frac{1}{K}\sum_{i=1}^{K}\frac{2\, Precision_i \,.\, Recall_i}{Precision_i + Recall_i} \quad (4)$$

F1macro is the average of F1-per-class. F1 is a harmonic average that balances precision and recall, useful when you want to take the middle ground between accuracy and completeness. Because it is macro, this metric penalizes poor performance on minority classes even if the global accuracy appears high.

$$\overline{\mathrm{Score}}=\frac{1}{\mathrm{K}}\sum_{\mathrm{i=1}}^{\mathrm{K}} \mathrm{Score}^{(\mathrm{j})} \quad (5)$$

Formula 5 is the average of Stratified 5-Fold Cross-Validation. The dataset is divided into k folds with StratifiedKFold, so that the proportion of classes is maintained in each fold. The model is trained on k – 1 folds (approximately 80% of data) and tested on the remaining fold

(20% of data) iteratively. The accuracy, precision-macro, recall-macro, and f1- macro metrics are calculated per fold as Score((j)),then averaged to obtain a more stable and less biased performance estimate than a single hold-out.

G. Insight Analysis

Following the sentiment classification stage, an insight analysis was conducted to extract deeper and actionable information from the model-labeled reviews. This analysis was implemented through two complementary approaches. First, aspect-level analysis was performed using a rule-based keyword matching method In this process, product aspects were identified by extracting the most frequent keywords associated with specific categories from the dataset. For instance, the 'Effectiveness' aspect was identified through keywords such as *'sembuh'*, *'jerawat hilang'*, *'cocok'*, and *'ampuh'*. Meanwhile, the 'User Experience' aspect focused on keywords related to texture and application sensations, such as *'lengket'*, *'cepat meresap'*, and *'bau'*. The implementation of this keyword dictionary ensures that aspect grouping is conducted consistently and objectively based on the context of consumer reviews. Second, a competitive analysis by product category was carried out by aggregating sentiment labels predicted by the classification models at the brand level. This approach enabled the identification of competitive strengths and weaknesses across brands within the same skincare category, ensuring that brand positioning was derived from inferred sentiment patterns rather than raw review frequencies alone.

III. RESULTS

After the data collection and *pre-processing* stages, the feature extraction stage was carried out to convert text data into numerical vector representations using the TF-IDF method[33]. Then, before the TF-IDF feature representation was applied to train the two *machine learning* algorithms, Multinomial Naïve Bayes and Support Vector Machine (SVM), SMOTE was first applied to overcome the imbalance between positive and negative data[34]. The final stage is to evaluate and compare the performance of the two models using the *5-Fold Cross-Validation* method to find the most appropriate model, with the results of the comparison based on the metrics of Accuracy, Precision, Recall, and F1- Score.

A. Data Collection

The dataset collected for this study contains four types of skincare products specifically for acne- prone skin from three different brands, namely Npure, Wardah, and Emina, with a total of 7,231 reviews taken from the Shopee marketplace as the data source using the JavaScript console scraping method. For each brand, the researchers took four types of skincare products, namely face wash, toner, moisturizer, and sunscreen. Then, for each review, the researcher collected information such as ratings, usernames, and review content.

B. Semi-Automatic Labelling

Semi-automatic labeling was applied to a dataset consisting of two sources: 575 manually labeled reviews and 6,657 new reviews to be processed automatically. After the manual labeling process was completed, the automatic sentiment labeling process was continued using standard RoBERTa. All 6,657 data that have been successfully labeled automatically for their aspects and sentiments are combined with 575 manually labeled data, resulting in a final combined dataset ready for analysis with a total of 7,232 data with a data distribution of 5,531 reviews that have non-null reviews, a total of 3,498 positive sentiments, a total of 506 negative sentiments, and 3,228 neutral sentiments. Neutral-labeled data was filtered out and not included in the modeling stage.

C. Text Pre-processing

From the results of data collection and data labeling, the data will then proceed to the text pre- processing stage. The results of data pre-processing, ranging from case folding, punctuation removal, word normalizer, stopword removal, and stemming.

D. Extraction Features

The cleaned reviews were then converted into numerical vector representations using TF-IDF. This process produced a feature matrix measuring 4004 x 5000. These features were extracted not only from single words such as "absorb," but also from two-word phrases such as "quickly absorb," allowing the model to capture a richer context. This method also ensures that the most informative words for distinguishing sentiment, such as 'acne', are given the highest weight, creating a robust data representation for the classification stage.

E. SMOTE

After pre-processing and filtering, the total consisted of 4,004 reviews. To prevent data leakage, the data first stratified into a training set (80%) containing 3,203 samples and a held-out testing set (20%) containing 801 samples.

SMOTE was applied exclusively to the training set. This technique generated synthetic samples for the minority class (negative) to match the majority class (positive), resulting in 2,798 samples for each class. Consequently, the final training dataset ready for 5-Fold Cross-Validation increased to 5,596 balanced samples. Meanwhile, the testing set remained in its original distribution to serve as an unbiased benchmark for final evaluation.

F. Modelling and Evaluation

The performance of the proposed models was evaluated on the held-out testing data, which constituted 20% of the original dataset. This phase is critical to measure the models' ability to generalize to new, unseen reviews. As visualized in the Confusion Matrices (Figure 2), the Multinomial Naïve Bayes (MNB) model demonstrated a slightly higher overall prediction capability compared to Linear SVM. Specifically, MNB correctly classified a total of 730 instances (comprising 65 negative and 665 positive samples), whereas SVM correctly classified 726 instances. While the SVM model showed a marginal advantage in detecting negative sentiments (68 correct predictions versus 65 for MNB), the MNB model proved more effective and robust in correctly identifying the majority positive class, resulting in fewer false negatives for positive reviews.

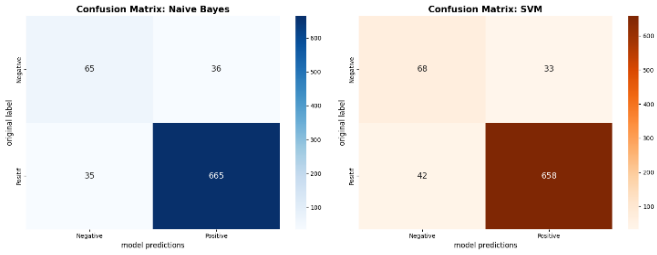

Figure 2. Confusion Matrix

Further quantitative comparison is summarized in Table I and illustrated in Figure 3. The results confirm that Multinomial Naïve Bayes outperformed Linear SVM across most key metrics. MNB achieved the highest Accuracy at 91.14% and a superior F1-Score of 79.80%. In comparison, Linear SVM recorded an Accuracy of 90.64% and an F1-Score of 79.53%. Although SVM exhibited a higher Recall of 80.66%, indicating better sensitivity towards the minority class its lower Precision (78.52%) negatively impacted its overall harmonic mean score. Consequently, due to its balanced performance between precision and recall and higher overall accuracy, Multinomial Naïve Bayes is identified as the most suitable model for sentiment classification in this study.

TABEL I
COMPARISON OF PERFORMANCE METRICS

Model	Accuracy	Precision	Recall	F1-Score
MNB	91.14%	79.93%	79.68%	79.80%
SVM	90.64%	78.52%	80.66%	79.53%

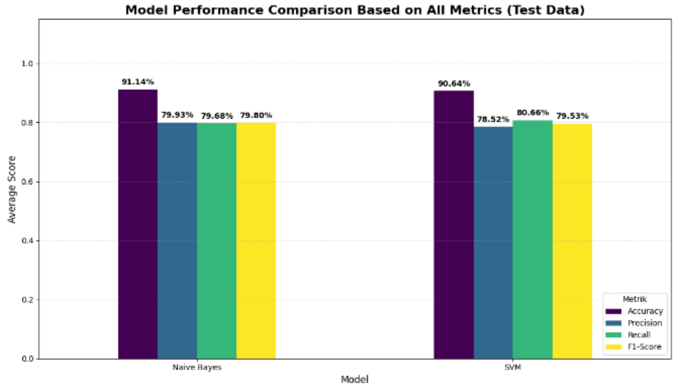

Figure 3. Comparison of NB and SVM

G. Insight Analysis

Analysis in Figure 4 highlights 'Experience' as the leading topic, surpassing 'Effectiveness'. This suggests that while consumers value tangible results, their positive sentiment is most heavily influenced by sensory factors and ease of use, including texture, aroma, and packaging practicality.

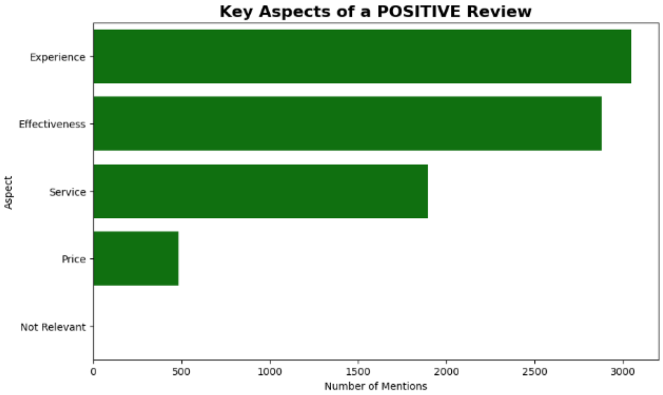

Figure 4. Analysis of positive reviews based on aspects

Conversely, Figure 5. Analysis of negative reviews based on aspects shows a change in order, with the aspect of "Effectiveness" becoming the main complaint, followed by "*Experience*." These findings indicate that the main trigger for consumer disappointment is when the product fails to deliver the expected results, such as not working or causing breakouts. Poor "Experience" aspects, such as stinging or stickiness, are also the second most important factor driving negative reviews.

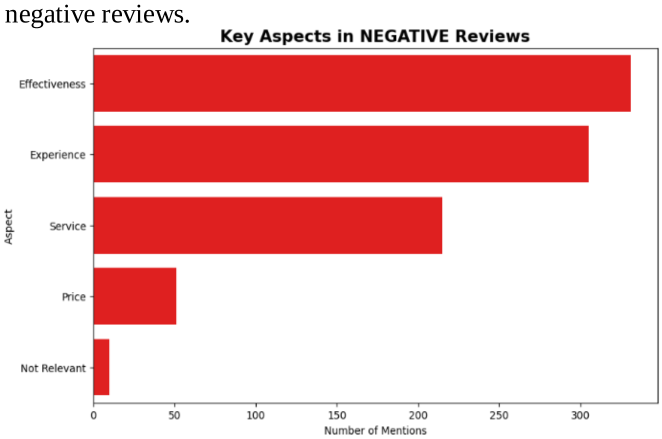

Figure 5. Analysis of negative reviews based on aspects

Figure 6 presents the sentiment distribution of toner products across different brands. The results indicate that certain brands receive both a high volume of positive and negative reviews, reflecting diverse consumer responses. This pattern suggests that higher market exposure tends to amplify heterogeneous user experiences rather than uniformly positive sentiment.

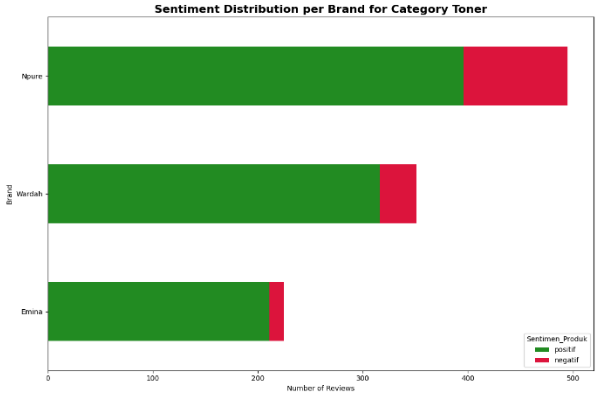

Figure 6. Sentiment Distribution of Toner Products by Brand

As shown in Figure 7, Wardah dominates positive sentiment in the face wash category, while Npure accounts for the highest proportion of negative reviews. This pattern suggests greater polarization of consumer responses in cleansing products.

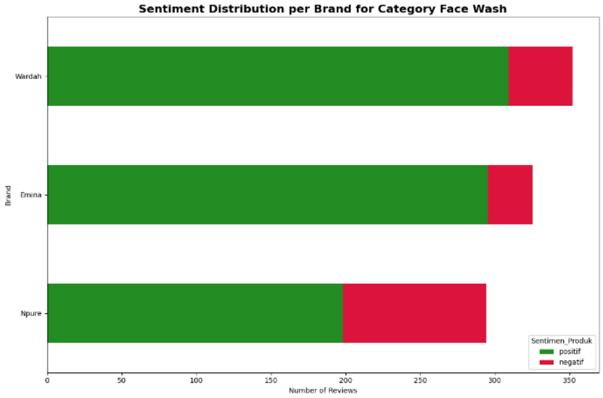

Figure 7. Sentiment Distribution of Face Wash Products by Brand

Figure 8 presents the competitive sentiment distribution for moisturizer products. Similar to the toner category, Npure emerges as both the most positively and negatively discussed brand, reflecting heterogeneous user experiences.

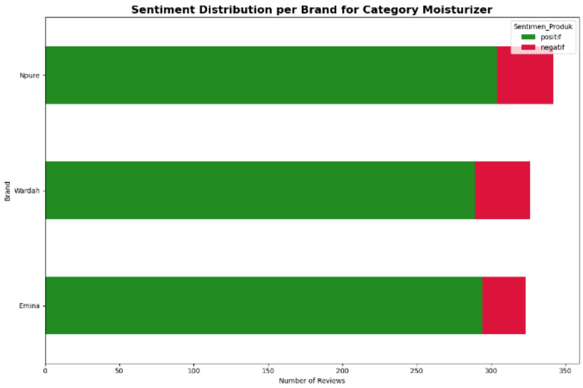

Figure 8. Sentiment Distribution of Moisturizer Products by Brand

Figure 9 depicts the sentiment distribution for sunscreen products by brand. The results show pronounced variability in consumer sentiment, suggesting that factors such as skin sensitivity, texture feel, and product formulation play a crucial role in shaping user satisfaction with sunscreen products.

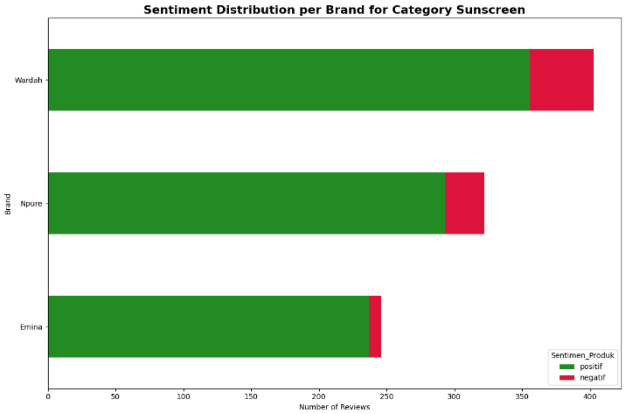

Figure 9. Sentiment Distribution of Sunscreen Products by Brand

The subsequent analysis examined competitive sentiment patterns among brands within specific skincare product categories to better understand their relative market positioning. As illustrated in Figures 6, 8, and 9, an interesting pattern emerges in the toner, moisturizer, and sunscreen categories. Brands such as Npure (toner and moisturizer) and Wardah (sunscreen) simultaneously appear as the most positively and negatively discussed brands within their respective categories. This pattern indicates that brands with broader consumer exposure tend to generate a wider range of user experiences, resulting in higher volumes of both favorable and unfavorable sentiment as reflected in the model-derived sentiment distributions.

In contrast, a different pattern is observed in the face wash category (Figure 7), where Wardah consistently receives the highest proportion of positive sentiment, while Npure accounts for the largest share of negative reviews. This suggests that consumer responses to face wash products may be more polarized and sensitive to individual skin conditions and usage preferences.

IV. Discussions

Based on the evaluation results presented in Table 1, both models demonstrated strong performance, achieving accuracy scores exceeding 90%. However, contrary to the initial hypothesis, Multinomial Naïve Bayes (MNB) slightly outperformed Linear SVM, with a testing accuracy of 91.14% compared to 90.64% for SVM.

Although accuracy provides a general overview of model performance, relying solely on accuracy can be misleading when dealing with imbalanced datasets. Therefore, this study also considers precision, recall, and F1-score to provide a more comprehensive evaluation of the proposed models. As shown in the evaluation results, both Multinomial Naïve

Bayes and Linear SVM demonstrate relatively balanced precision and recall values, indicating that the models are capable of identifying both positive and negative sentiment classes without excessive bias toward the majority class. The F1-score further confirms that the classification performance is not solely driven by accuracy, but also reflects a balanced trade-off between precision and recall.

An analysis conducted during the training phase shows that SVM achieved a higher validation accuracy (97.41%) than Naïve Bayes (96.62%). These training-phase validation values are reported solely to illustrate relative learning behavior and are not used as the primary evaluation metrics, which are based on the testing results presented in Table 1. The larger discrepancy observed between training-phase validation accuracy and testing accuracy in the SVM model suggests a tendency toward overfitting under the current experimental configuration. This behavior may be influenced by the interaction between high-dimensional TF-IDF features and the synthetic samples introduced through SMOTE. In contrast, Naïve Bayes exhibited a smaller gap between training and testing performance, indicating more stable generalization behavior for this dataset.

Although Multinomial Naïve Bayes achieved slightly higher testing accuracy than Linear SVM, the observed performance difference is marginal (less than 1%). As no statistical significance testing was conducted, this difference should be interpreted as an indicative trend rather than a definitive claim of model superiority.

These findings contrast with several prior studies that consistently reported SVM as outperforming Naïve Bayes [35], [36], and [37]. The divergence observed in this study suggests that model performance is highly dependent on dataset characteristics, feature representation, and data balancing strategies. In the context of acne-prone skincare reviews, which contain informal, domain-specific, and highly varied vocabulary, Naïve Bayes appears to offer more consistent performance under imbalanced data conditions.

Despite the observed performance differences, achieving accuracy levels above 90% for both classifiers indicates that the proposed preprocessing pipeline was effective in capturing sentiment-related patterns within acne-prone skincare reviews. Normalization and stemming contributed to standardizing informal and colloquial vocabulary, while the application of SMOTE helped mitigate class imbalance and reduce bias toward the majority sentiment class. Although oversampling may introduce additional variability for certain classifiers, its application played an important role in enabling balanced learning across sentiment categories under the current experimental setting.

Beyond classification performance, the aspect-based analysis provides deeper insights into consumer perceptions of skincare products. The results indicate that Experience is the most dominant contributor to positive sentiment, followed by Effectiveness. This finding suggests that user satisfaction is strongly influenced by sensory comfort and ease of use, such as texture, aroma, and packaging, in addition to perceived functional benefits. Furthermore, competitive analysis reveals that brands such as Npure (toner and moisturizer) and Wardah (sunscreen) appear as both highly favored and frequently criticized within their respective categories. This duality reflects the impact of high market penetration, where diverse consumer expectations and subjective experiences play a significant role in shaping sentiment outcomes.

V. Conclusion

This study successfully implemented a sentiment classification framework for acne-prone skincare products reviews on the Shopee platform. By employing TF-IDF for feature extraction and applying Synthetic Minority Over-sampling Technique (SMOTE) exclusively to the training set, the research ensured a rigorous evaluation on a separate held-out testing set.

Based on the final evaluation, the study concludes that Multinomial Naïve Bayes (MNB) is the most effective algorithm for this specific domain. Contrary to the initial hypothesis, MNB demonstrated superior generalization capabilities with a testing accuracy of 91.14% and an F1-Score of 79.80%. In contrast, while the Support Vector Machine (SVM) achieved high stability during the training phase (97.41%), it suffered from mild overfitting, resulting in a lower accuracy of 90.64% on unseen testing data. This finding suggests that for high-dimensional and sparse text data heavily processed with synthetic oversampling, the probabilistic nature of Naïve Bayes offers more robustness compared to the decision boundary approach of SVM.

The success of the classification framework was significantly bolstered by the comprehensive pre-processing pipeline. Case folding, slang normalization, and stemming played a critical role in standardizing informal Indonesian text, making the features extracted via TF-IDF more representative. Furthermore, the strategic application of SMOTE on the training data proved crucial in mitigating class imbalance, preventing the models from being biased toward the majority class.

Beyond technical performance, this study provides valuable actionable insights through aspect and competitive analysis. The discovery that 'Experience' (sensory attributes like texture and smell) is the dominant driver of positive sentiment, surpassing 'Effectiveness' indicates that consumer satisfaction is multidimensional. Additionally, the competitive mapping of brands such as Npure and Wardah offers industry players a strategic overview of their market position, highlighting specific areas for product differentiation and marketing improvement.

Overall, this research contributes to both academic understanding and practical industry applications. It demonstrates that simpler models like Naïve Bayes can outperform more complex ones in specific sentiment analysis contexts. For future research, it is recommended to explore hyperparameter tuning or ensemble methods to further mitigate overfitting in SVM and enhance the predictive performance of both models.

VI. Limitations And Future Work

This study has several limitations. First, the comparison between Naïve Bayes and SVM did not incorporate statistical significance testing to formally assess whether the observed performance differences are statistically meaningful. Second, the use of classical machine learning models, while interpretable and computationally efficient, may not capture deeper semantic relationships present in complex textual data. Additionally, the dataset was collected from a single e-commerce platform, which may limit cross-platform generalizability.

Future research may address these limitations by incorporating repeated experiments or cross-validation-based statistical tests, exploring embedding-based representations, and extending the dataset across multiple platforms and product categories to improve robustness and generalization.

References

[1] Dwi Tiyas Novitasari, M. A. Barata, and P. E. Yuwita, “Analisis Sentimen Pengguna Twitter Terhadap Skincare Dengan Metode Support Vector Machine (Svm),” *INTI Nusa Mandiri*, vol. 19, no. 2, pp. 325–332, 2025, doi: 10.33480/inti.v19i2.6297.

[2] F. Khoirunisa and S. Nurhayati, “Pengaruh Customers Online Review, Customers Online Rating, dan Harga Produk terhadap Keputusan Pembelian Melalui Marketplace Shopee,” *INOBIS J. Inov. Bisnis dan Manaj. Indones.*, vol. 7, no. 4, pp. 456–469, 2024, doi: 10.31842/jurnalinobis.v7i4.336.

[3] Restuti Nunik and Kurnia Marlina, “Pengaruh Harga, Ulasan Produk, Kemudahan Transaksi, Kualitas Informasi dan Kepercayaan Terhadap Keputusan Pembelian Produk Kecantikan Secara Online Pada Marketplace Shopee,” *Borobudur Manag. Rev.*, vol. 2, no. 1, pp. 24–40, 2022, doi: 10.31603/bmar.v2i2.6817.

[4] V. G. Shintarani, R. Mayasari, and ..., “Analisis Sentimen Ulasan Konsumen Pada Produk Ponsel Pintar Menggunakan Metode Naïve Bayes,” *... Mandalika ISSN 2721 ...*, pp. 771–781, 2023, [Online]. Available: https://ojs.cahayamandalika.com/index.php/JCM/article/view/2101%0Ahttps://ojs.cahayamandalika.com/index.php/JCM/article/download/2101/1662

[5] B. Z. Ramadhan, R. I. Adam, and I. Maulana, “Analisis Sentimen Ulasan pada Aplikasi E-Commerce dengan Menggunakan Algoritma Naïve Bayes,” *J. Appl. Informatics Comput.*, vol. 6, no. 2, pp. 220–225, 2022, doi: 10.30871/jaic.v6i2.4725.

[6] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiasari, “Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE,” *Aiti*, vol. 18, no. 2, pp. 173–184, 2021, doi: 10.24246/aiti.v18i2.173-184.

[7] I. Kurniawan, A. Lia Hananto, S. Shofia Hilabi, A. Hananto, B. Priyatna, and A. Yuniar Rahman, “Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter,” *J. Tek. Inform. dan Sist. Inf.*, vol. 10, no. 1, pp. 731–740, 2023, [Online]. Available: http://jurnal.mdp.ac.id

[8] R. F. Rahmadzani, R. W. Pratiwi, and A. N. Paradita, “Comparison of Naive Bayes and Support Vector Machine (SVM) Methods in Female Daily Skincare Sentiment Analysis,” *Radiant*, vol. 6, no. 2, pp. 99–108, 2025, doi: 10.52187/rdt.v6i2.316.

[9] I. Bazar, F. Wajidi, and A. A. Asnan Cirua, “Analisis Sentimen Ulasan Aplikasi Wondr By BNI Menggunakan Algoritma Svm Dengan Optimasi Kernel Trick,” *STORAGE J. Ilm. Tek. dan Ilmu Komput.*, vol. 4, no. 2, pp. 69–81, 2025, doi: 10.55123/storage.v4i2.5178.

[10] U. I. Shabrina, M. I. Java, and S. Rochimah, “Optimizing Sentiment Analysis in Educational Youtube Videos: a Comparative Study of Roberta and Multinomial Naive Bayes,” *JUTI J. Ilm. Teknol. Inf.*, pp. 83–90, 2024, doi: 10.12962/j24068535.v22i2.a1204.

[11] K. S. Putri, I. R. Setiawan, and A. Pambudi, “Analisis Sentimen Terhadap Brand Skincare Lokal Menggunakan Naïve Bayes Classifier,” *Technol. J. Ilm.*, vol. 14, no. 3, p. 227, 2023, doi: 10.31602/tji.v14i3.11259.

[12] W. Clarisha, “Analisis Sentimen Sunscreen Lokal Skintific , Somethinc , dan Avoskin dengan Naive Bayes dan SVM Sentiment Analysis of Local Sunscreen Skintific , Somethinc , and Avoskin with Naive Bayes and SVM,” vol. 7, pp. 264–271, 2025.

[13] L. S. Lestari, T. Sutrisno, and I. Lewenusa, “Sentiment Analysis on Skincare Product Reviews Using Lexicon-Based and Comparison of SVM Kernel,” vol. 5, no. 11, pp. 5250–5259, 2024.

[14] S. C. Jenkins, R. F. Lachlan, and M. Osman, “An integrative framework for mapping the psychological landscape of risk perception,” *Sci. Rep.*, pp. 1–17, 2024, doi: 10.1038/s41598-024-59189-y.

[15] Jasmarizal, Junadhi, Rahmaddeni, and M. Khairul Anam, “Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Terhadap Produk Skincare,” *Indones. J. Comput. Sci.*, vol. 13, no. 1, pp. 1438–1450, 2024, doi: 10.33022/ijcs.v13i1.3654.

[16] H. Harnelia, “Analisis Sentimen Review Skincare Skintific Dengan Algoritma Support Vector Machine (Svm),” *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 2, 2024, doi: 10.23960/jitet.v12i2.4095.

[17] R. Durgam, N. B. Pamula, N. Dharani, B. V. N. S. Kumar, P. Durga, and V. Balaji, “AI-Powered Empathy: Sentiment Analysis In Personal Care Using RoBERTa And XLNet,” *J. Theor. Appl. Inf. Technol.*, vol. 103, no. 8, pp. 3455–3470, 2025.

[18] S. K. Wardani and Y. A. Sari, “Analisis Sentimen menggunakan Metode Naïve Bayes Classifier terhadap Review Produk Perawatan Kulit Wajah menggunakan Seleksi Fitur N-gram dan Document Frequency Thresholding,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 12, pp. 5582–5590, 2021.

[19] R. Nurhidayat and K. E. Dewi, “Penerapan Algoritma K-Nearest Neighbor Dan Fitur Ekstraksi N-Gram Dalam Analisis Sentimen Berbasis Aspek,” *Komputa J. Ilm. Komput. dan Inform.*, vol. 12, no. 1, pp. 91–100, 2023, doi: 10.34010/komputa.v12i1.9458.

[20] N. Babanejad, H. Davoudi, A. Agrawal, A. An, and M. Papagelis, “The Role of Preprocessing for Word Representation Learning in Affective Tasks,” *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 254–272, 2024, doi: 10.1109/TAFFC.2023.3270115.

[21] S. K. Narayanasamy, Y. C. Hu, S. M. Qaisar, and K. Srinivasan, “Effective Preprocessing and Normalization Techniques for COVID-19 Twitter Streams with POS Tagging via Lightweight Hidden Markov Model,” *J. Sensors*, vol. 2022, 2022, doi: 10.1155/2022/1222692.

[22] S. Sarica and J. Luo, “Stopwords in technical language processing,” *PLoS One*, vol. 16, no. 8 August, pp. 1–13, 2021, doi: 10.1371/journal.pone.0254937.

[23] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, “Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.

[24] R. L. Musyarofah, E. U. Utami, and S. R. Raharjo, “Analisis Komentar Potensial pada Social Commerce Instagram Menggunakan TF-IDF,” *J. Eksplora Inform.*, vol. 9, no. 2, pp. 130–139, 2020, doi: 10.30864/eksplora.v9i2.360.

[25] Y. Sulistiyo Wibowo *et al.*, “Journal of Data Science and Software Engineering Performance Analysis Of Classifier Of Facebook Data Using Unigram & Bigram Combinations (Yudha) | 63,” vol. 01, no. 2, pp. 63–72, 2020.

[26] M. F. Ramadhan, “Klasifikasi Topik dan Sentimen Judul Berita dengan Augmentasi dan,” vol. 4, no. 2, pp. 6732–6741, 2025.

[27] Normah, B. Rifai, S. Vambudi, and R. Maulana, “Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE,” *J. Tek. Komput. AMIK BSI*, vol. 8, no. 2, pp.

174–180, 2022, doi: 10.31294/jtk.v4i2.

[28] I. Technology, K. Utama, J. G. Kelang, W. Persekutuan, and K. Lumpur, “A Comparative Study Of Machine Learning Algorithms For Sentiment Analysis,” vol. 2021, no. Icdxa 2021, pp. 63–68, 2022.

[29] R. A. Fauzan and M. Mufti, “Analisis Sentimen Komentar Youtube Program Kampus Merdeka Berbasis Web Menggunakan Algoritma Multinomial Naïve Bayes,” *Semin. Nas. Mhs. Fak. Teknol. Inf.*, vol. 2, no. 2, pp. 864–871, 2023, [Online]. Available: https://senafti.budiluhur.ac.id/index.php/senafti/article/view/929/563

[30] A. R. Makhtum and M. Muhajir, “Sentiment Analysis of Omnibus Law Using Support Vector Machine (Svm) With Linear Kernel,” *Barekeng*, vol. 17, no. 4, pp. 2197–2206, 2023, doi: 10.30598/barekengvol17iss4pp2197-2206.

[31] J. R. Almonteros and J. B. Matias, “Integration of Stratified KFold Cross Validation to Enhance Prediction Accuracy: A Comparison Study,” *2024 5th Int. Conf. Data Anal. Bus. Ind. ICDABI 2024*, no. October 2024, pp. 81–85, 2024, doi: 10.1109/ICDABI63787.2024.10800425.

[32] D. Soyusiawaty and F. G. Putra, “Pengembangan Chatbot Untuk Layanan Pimpinan Daerah Muhammadiyah Kota Yogyakarta Menggunakan Metode Rule-based,” *J. Penerapan Sist. Inf. (Komputer Manajemen)*, vol. 4, no. 2, pp. 354–363, 2023.

[33] D. Pakpahan, V. Siallagan, and S. Siregar, “Classification of E-Commerce Product Descriptions with The Tf-Idf and Svm Methods,” *Sinkron*, vol. 8, no. 4, pp. 2130–2137, 2023, doi: 10.33395/sinkron.v8i4.12779.

[34] S. Wang, Y. Dai, J. Shen, and J. Xuan, “Research on expansion and classification of imbalanced data based on SMOTE algorithm,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-03430-5.

[35] N. Z. B. Jannah and K. Kusnawi, “Comparison of Naïve Bayes and SVM in Sentiment Analysis of Product Reviews on Marketplaces,” *Sinkron*, vol. 8, no. 2, pp. 727–733, 2024, doi: 10.33395/sinkron.v8i2.13559.

[36] U. Kusnia and F. Kurniawan, “Analisis Sentimen Review Aplikasi Media Berita Online Pada Google Play menggunakan Metode Algoritma Support Vector Machines (SVM) Dan Naive Bayes,” *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 4, no. 36, pp. 222–231, 2022, [Online]. Available: https://jurnal.yudharta.ac.id/v2/index.php/EXPLORE-IT/article/view/3116

[37] Ketut Mediana Ayu Candrayani, I Made Agus Dwi Suarjaya, and Anak Agung Ketut Agung Cahyawan Wiranatha, “Analisis Sentimen Pembelajaran Daring Era Pandemi COVID-19 Menggunakan Naive Bayes Dan SVM,” *Tematik*, vol. 10, no. 1, pp. 47–53, 2023, doi: 10.38204/tematik.v10i1.1274.