Institut Riset dan Publikasi Indonesia (IRPI)
**MALCOM: Indonesian Journal of Machine Learning and Computer Science**
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 3 July 2024, pp: 1077-1086
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

# *Comparison of Sentiment Analysis Algorithms with SMOTE Oversampling and TF-IDF Implementation on Google Reviews for Public Health Centers*

**I Gede Bintang Arya Budaya[1*], I Ketut Putu Suniantara[2]**

[1]Information Technology, Institute of Technology and Business STIKOM Bali, Indonesia
[2]Information System, Institute of Technology and Business STIKOM Bali, Indonesia

E-Mail: [1]bintang@stikom-bali.ac.id, [2]suniantara@stikom-bali.ac.id

*Received May 30th 2024; Revised Jun 28th 2024; Accepted Jul 5th 2024*
*Corresponding Author: I Gede Bintang Arya Budaya*

**Abstract**

*Sentiment analysis, or opinion mining, is a critical area of natural language processing that identifies sentiments in free text. As digital business services expand and user-generated content increases, analyzing sentiments in online reviews becomes crucial for enhancing business operations and customer satisfaction. This study focuses on sentiment analysis of user reviews from Google Reviews for Public Health Centers (PHCs) in Bali, Indonesia, utilizing five machine learning models: Logistic Regression, Support Vector Machine (SVM), XGBoost, Naive Bayes, and Random Forest. These models classify sentiments into positive and negative categories using a dataset balanced with SMOTE to address class imbalance and improve accuracy. Additionally, TF-IDF (Term Frequency-Inverse Document Frequency) is employed to highlight key sentiment indicators. We divided a total of 1,834 reviews, allocating 20% for testing and 80% for training, ensuring comprehensive evaluation under real-world conditions. Logistic Regression and Naive Bayes emerged as top performers, achieving an accuracy of 0.89, with Logistic Regression demonstrating balanced precision and recall. This study contributes to advancing academic understanding of sentiment analysis in healthcare and provides practical insights for business administrators on leveraging online customer feedback. The findings underscore the importance of integrating SMOTE for class balance and TF-IDF for identifying crucial sentiment markers, emphasizing the selection of appropriate machine learning techniques tailored to data characteristics and project objectives to optimize technological and business outcomes.*

*Keyword: Binary Classification, Imbalanced Classes, Sentiment Classification, Supervised Learning, User Feedback*

## 1. INTRODUCTION

Sentiment analysis, also known as opinion mining, is a branch of natural language processing focused on automatically classifying the sentiment contained within free text. This technique not only involves the classification of viewpoints and the prediction of the semantic orientation of adjectives but also the classification of subjectivity and other nuanced aspects. It has been applied to numerous languages, including Indonesian, as evidenced by references [1], [2], [3]. With the expansion of digital business services and the increase in user-generated content like product/services on business reviews [4] and social media interactions [5], [6], sentiment analysis has experienced significant growth. Google Business Profile (GBP) is one such platform that manages a vast array of business profiles, providing detailed information such as operational hours, address, contact numbers, and also integrated with google map that contain user reviews. In the healthcare sector, Public Health Centers (PHCs) in Indonesia have effectively utilized GBP to improve service accessibility and effectiveness. Profiles on GBP offer essential information, such as operating hours and locations, thus facilitating easier access to healthcare services. Reviews from visitors and patients are particularly valuable, offering a realistic view of service quality and fostering ongoing improvements [7]. Additionally, direct responses to reviews can significantly enhance user trust and satisfaction.

User reviews play a critical role as a source of information for the public, influencing decision-making processes across various businesses [8], [9]. Positive reviews often attract more customers, build trust, and enhance a business's reputation, while negative reviews can encourage business owners to evaluate and improve their offerings. The academic research community has extensively utilized sentiment analysis of digital platform reviews to gain deeper insights into public opinions and reactions to services or products. For example, research [10] employed machine learning algorithms like Naive Bayes and logistic regression to analyze sentiments from user reviews of a marketplace app, collected from Google Play Store reviews across two different datasets. Both algorithms demonstrated similarly strong performances. Furthermore, research

involved applying sentiment analysis techniques using a Support Vector Machine (SVM) [11], [12], which yielded favorable outcomes. Techniques such as XGBoost [13], [14] and Random Forest [15], [16] have also been employed in sentiment analysis which also yielded favorable outcomes.

In contrast to previous studies [17], which utilized a multi-label approach to sentiment analysis on Google reviews from GBP profiles without addressing imbalanced data, this research introduces several novel contributions. Firstly, it addresses the challenge of class imbalance within the dataset, which was found to adversely affect model performance [18]. By shifting to a binary classification system (positive and negative sentiments) and employing SMOTE Oversampling, this study aims to improve the accuracy and reliability of sentiment classification. Secondly, a comprehensive comparison of five different classification algorithms provides insights into which methods are most effective for sentiment analysis in this context, contributing to methodological advancements in the field. These steps collectively aim to provide a more robust framework for sentiment analysis of user-generated content, particularly in contexts where challenges in sentiment classification arise in the domain of public health centers.

## 2. MATERIALS AND METHOD

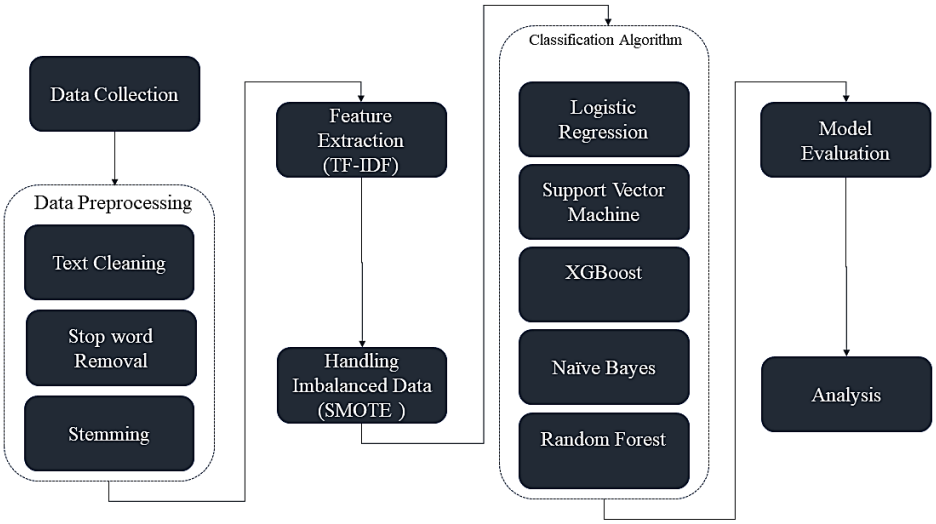


**Figure 1.** Research Methodology

### 2.1. Data Collection

The data were collected from various PHC's GBP reviews in Bali Province, Indonesia. The data were then labeled based on user ratings and the involvement of two experts. Reviews with 1 to 2 stars were initially assigned a negative label, and those with 3 to 5 stars were assigned a positive label. However, the final labels were determined after both experts reviewed the content and reached the same conclusion. There are a total of 1.995 reviews. After labeling, there are 834 negative reviews and 1.000 positive reviews, so the final dataset contains 1.834 reviews. Table 1 shows the sample dataset from labelling process and Figure 1 visualize the sentiment count for each class.

**Table 1.** Sample of Dataset

| No | Sentiment | User Star Rating | Expert 1 | Expert 2 | Conclusion |
|---|---|---|---|---|---|
| 1 | Tolong..! Untuk Petugas pendaftaran tolong belajar etika bicara dengan orang yg lebih tua. Saya hanya kasihan dengan orang tua yg lagi sakit saat daftar malah dibentak bentak | 1 | negative | negative | negative |
| 2 | Saya puas dalam penanganan nya terus ditingkatkan lagi biar semakin bagus | 5 | positive | positive | positive |
| 3 | CLEAN AND NEAT!. The staff especially in the administration and cashier are polite and communicative.Open at 8 am to 12.30 pm. If you want to go here, please bring your identity such as KTP or KK. Oh ya! There is a little playground too! | 5 | positive | positive | positive |
| ... | .... | .... | .... | .... | .... |
| 1995 | mau tanya yang ktp badung mau rapid test body brp ya biaya nya | 2 | No Determined Sentiment | No Determined Sentiment | Not Used |

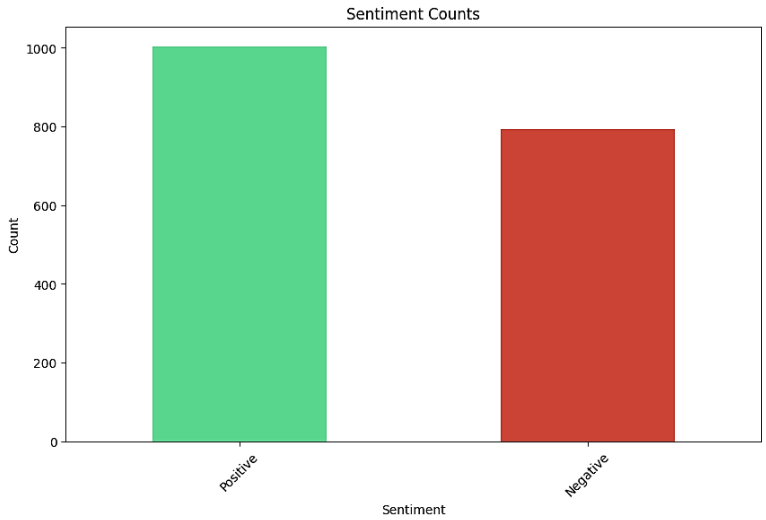


**Figure 2.** Sentiment Count

### 2.2. Data Preprocessing

The data preprocessing for our text analysis begins with text cleaning, which includes case folding and translating all English text into Indonesian and removing any numbers and symbols that could introduce noise into the dataset. After this initial cleanup, we proceed to remove stop words using Natural Language Toolkit, common Indonesian terms that appear frequently but do not contribute significantly to sentiment analysis, such as 'dan' (and), 'di' (in), and 'yang' (which). This step helps focus the analysis on more meaningful words. The final step in our preprocessing routine is stemming using the Sastrawi toolkit, which reduces words to their base or root form, thus simplifying the textual data and enhancing the performance of our sentiment classification by standardizing words to their core meanings. Table 2 shows the sample of preprocessed sentiment

**Table 2.** Sample of Preprocessed Sentiment

| No | Original Sentiment | Preprocessed Sentiment |
|---|---|---|
| 1 | Tolong..! Untuk Petugas pendaftaran tolong belajar etika bicara dengan orang yg lebih tua. Saya hanya kasihan dengan orang tua yg lagi sakit saat daftar malah dibentak bentak | tolong petugas daftar tolong ajar etik bicara orang tua kasih orang tua sakit daftar bentak bentak |
| 2 | Saya puas dalam penanganan nya terus ditingkatkan lagi biar semakin bagus | puas tangani tingkat bagus |
| ... | .... | .... |
| 1834 | CLEAN AND NEAT!. The staff especially in the administration and cashier are polite and communicative.Open at 8 am to 12.30 pm. If you want to go here, please bring your identity . Oh ya! There is a little playground too! | bersih rapi staf utama admin kasir sopan komunikatif buka jam pagi siang bawa identitas taman main kecil |

### 2.3. Feature Extraction using TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate how important a word is to a document within a collection of documents, or corpus, it is also important in the field of sentiment analysis [19], [20]. It is a crucial tool in the fields of text mining and information retrieval, especially for processing and analyzing large datasets of textual information. The TF component of TF-IDF emphasizes the frequency of a word in a specific document, highlighting its significance in that particular context.

$$\mathrm{TF}(t,d) = \frac{n\_t,d}{\sum t\prime \epsilon d n\_t\prime,d} \tag{1}$$

Term Frequency (TF) calculated as the number of times a term $t$, appears in a document $d$, divided by the total number of terms in that document. where $n_{t,d}$ is the term frequency, and the denominator is the sum of all term frequencies in the document.

$$\mathrm{IDF}(t,D) = \log(\frac{D}{|\{d \epsilon D : t \epsilon d\}|}) \tag{2}$$

Inverse Document Frequency (IDF) measured as the logarithm of the number of documents $D$ divided by the number of documents that contain the term $t$. The TF-IDF score is then calculated as:

$$\text{TFIDF}(t, d, D) = \text{TF}(t, d) \cdot \text{IDF}(t, D) \tag{3}$$

### 2.4. Handling Imbalanced Data using SMOTE

In dealing with imbalanced datasets, such as the case of 1.834 reviews comprising 1.000 positive and 834 negative ones, the Synthetic Minority Over-sampling Technique (SMOTE) proves to be a highly effective strategy [21], [22]. SMOTE addresses the imbalance by generating synthetic samples from the minority class—in this case, the negative reviews—to equalize the distribution between the classes. This is achieved by identifying feature space similarities among the minority class instances and synthesizing new samples that are interpolations of those that are closest. By balancing the dataset in this manner, SMOTE enhances the predictive performance of machine learning models, ensuring that they are not biased toward the majority class and can generalize better on unseen data. This approach is particularly valuable in sentiment analysis, where accurately predicting less frequent sentiments is often as crucial as identifying the majority sentiment.

### 2.5. Classification Algorithm

For the classification purpose the dataset split 80% for testing and 20% In the classification algorithm section are exploring the application and performance of five distinct machine learning models: Logistic Regression [10], Support Vector Machine (SVM) [4], [11], [12], [23], XGBoost [13], [14], Naive Bayes [5], [10], and Random Forest [15], [16]. Each of these models has been chosen due to their proven capabilities in handling classification tasks, particularly in the realm of sentiment analysis. Logistic Regression is utilized as a baseline model due to its simplicity and effectiveness in binary classification tasks. It models the probabilities for classification problems with two possible outcomes and is especially useful for understanding the impact of several independent variables. Support Vector Machine (SVM) is employed for its robustness and effectiveness in high-dimensional spaces, which is ideal for text classification tasks. SVM is known for its ability to create the optimal boundary between the possible outputs, which can be particularly useful in distinguishing between positive and negative sentiments.

XGBoost stands out for its speed and performance, particularly in structured or tabular data. As an implementation of gradient boosted decision trees designed for speed and performance, XGBoost is particularly adept at managing imbalances in data, making it a strong candidate for our dataset of sentiment-laden text. Naive Bayes is chosen for its efficiency and suitability for large datasets. Given its assumption of independence among predictors, Naive Bayes is particularly fast, making it ideal for scenarios where computational efficiency is crucial, such as processing large volumes of review data. Random Forest is included for its high accuracy and robustness, stemming from its ensemble approach, which combines multiple decision trees to produce a more effective overall model. This model is less likely to overfit compared to some other models and is good at handling both binary and multiclass classification tasks.

### 2.6. Model Evaluation

Each model's performance is tested using metrics such as accuracy, precision, recall, and F1-score, which are critical for evaluating the effectiveness of each classifier in accurately predicting sentiments. The comparative analysis of these models helps in identifying which algorithm performs best under specific conditions of our dataset, considering factors such as class imbalance, the complexity of the feature space, and the nuances of textual data in sentiment analysis. This methodical approach ensures a comprehensive understanding of each model's strengths and limitations in the context of sentiment classification.

$$\text{Accuracy} = \frac{\text{True Positives+True Negatives}}{\text{Total Observations}} \tag{4}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}} \tag{5}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}} \tag{6}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision*Recall}}{\text{Precision+Recall}} \tag{7}$$

### 2.7. Analysis

This section conducts an analysis of the outcomes produced by each of the five classification models: Logistic Regression, SVM, XGBoost, Naive Bayes, and Random Forest. The analysis aims to elucidate the comparative strengths and weaknesses of these models in the context of sentiment analysis.

## 3. RESULTS AND DISCUSSION

In this section present the findings from our sentiment analysis of Google reviews from GBP profiles using the methodologies outlined in the previous section. The results are discussed in terms of their implications for understanding public perceptions and sentiments towards public health centers. In this section also analyze the performance of different classification algorithms and evaluate the effectiveness of strategies implemented to address data imbalance.

### 3.1. Result

The result of the implementation using the five algorithms, TF-IDF for feature extraction, and SMOTE oversampling can be shown in Table 3 and Figure 3.

**Table 3.** Performance Comparison

| No | Model | Class | Accuracy | Precision | Recall | F1-Scores |
|---|---|---|---|---|---|---|
| 1 | Logistic Regression | Positive | 0,89 | 0,92 | 0,87 | 0,89 |
| | | Negative | | 0,86 | 0,91 | 0,88 |
| 2 | SVM | Positive | 0,89 | 0,88 | 0,90 | 0,89 |
| | | Negative | | 0,89 | 0,87 | 0,88 |
| 3 | XGBoost | Positive | 0,86 | 0,84 | 0,90 | 0,87 |
| | | Negative | | 0,87 | 0,81 | 0,84 |
| 4 | Naive Bayes | Positive | 0,89 | 0,89 | 0,90 | 0,89 |
| | | Negative | | 0,88 | 0,88 | 0,88 |
| 5 | Random Forest | Positive | 0,86 | 0,83 | 0,92 | 0,88 |
| | | Negative | | 0,90 | 0,79 | 0,84 |

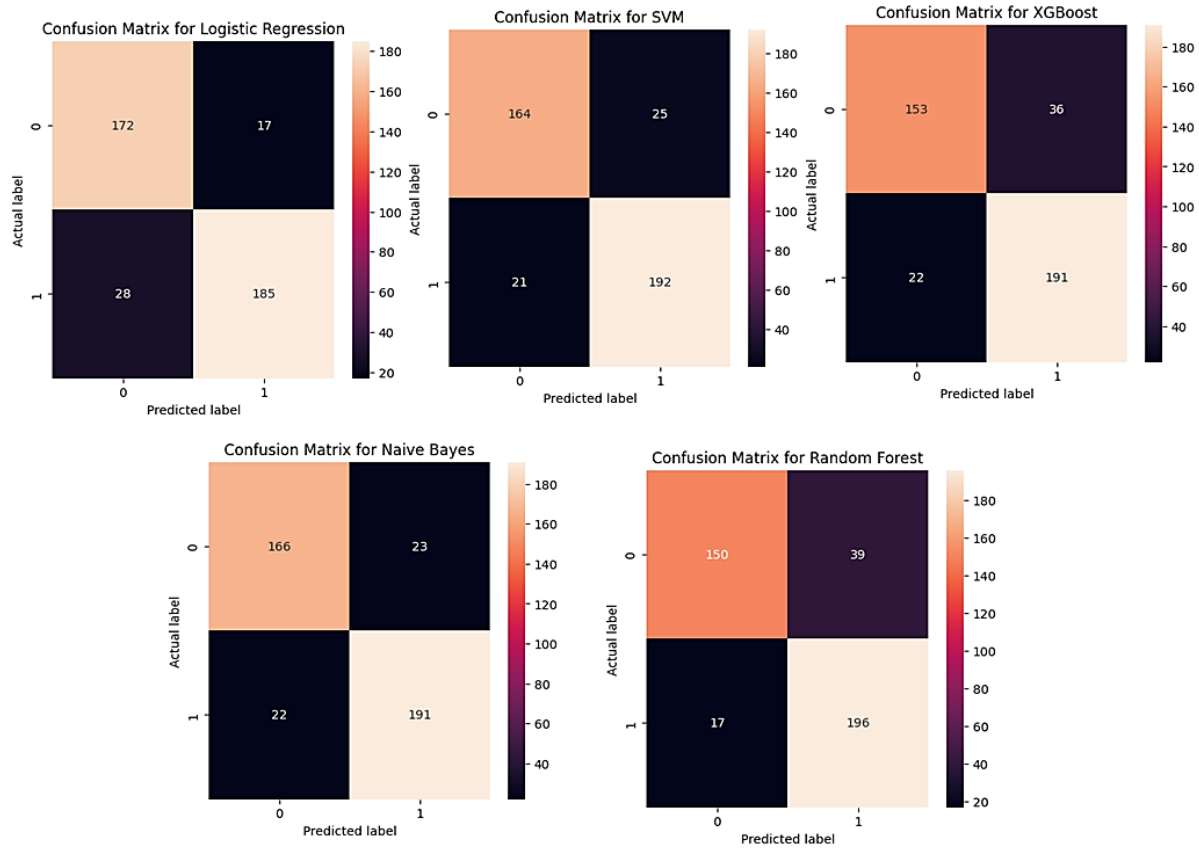


**Figure 3.** Confusion Matrix of Each Algorithm

The performance of the Logistic Regression model in classifying sentiments into negative (0) and positive (1) categories is captured comprehensively in the classification report, with an overall accuracy of 0.89. This suggests that the model correctly predicts the sentiment 89% of the time across the dataset consisting of 402 reviews. For the negative class, the model achieves a precision of 0.86, indicating that 86% of the reviews classified as negative are accurately identified, with a recall of 0.91, which reflects that 91% of actual negative reviews were correctly captured by the model. Conversely, in the positive class, the precision is higher at 0.92, showing that 92% of reviews classified as positive were truly positive, but with a slightly lower recall of 0.87, meaning the model captured 87% of all actual positive reviews. The F1-scores, which are balanced measures of precision and recall, stand at 0.88 for negative and 0.89 for positive reviews, indicating robust performance. The macro and weighted averages for precision, recall, and F1-score all align closely at around 0.89, underscoring a consistent performance across both sentiment classes.

The classification report for the SVM model provides detailed metrics for its performance in sentiment classification, with a dataset composed of 402 reviews. The overall accuracy of the SVM model stands at 0.89, indicating that the model accurately predicts the sentiment in 89% of cases, which demonstrates a high level of effectiveness in sentiment analysis tasks. For the negative class (0), the model shows a precision of 0.89, which means that 89% of the reviews predicted as negative are indeed negative, coupled with a recall of 0.87, indicating that the model successfully identifies 87% of all actual negative reviews. The corresponding F1-score of 0.88 suggests a balanced accuracy in terms of both precision and recall for the negative reviews. In the positive class (1), the SVM model exhibits a slightly lower precision of 0.88, indicating that 88% of the reviews classified as positive were correctly identified, and a recall of 0.90, which shows that 90% of all actual positive reviews were accurately captured by the model. This class achieves an F1-score of 0.89, slightly higher than that of the negative class, reflecting effective identification and classification of positive sentiments. The macro and weighted averages for precision, recall, and F1-score across both classes are consistent, each standing at around 0.89, which highlights the SVM model's reliable performance across different sentiment categories without significant bias toward either class. This balanced performance makes it a robust tool for analyzing sentiments, capable of handling the nuances of both positive and negative reviews effectively.

The classification report for the Random Forest model provides a detailed assessment of its performance in sentiment classification across a dataset of 402 reviews. The model achieves an overall accuracy of 0.86, which signifies that it correctly predicts sentiment 86% of the time, a strong indication of its capability in handling sentiment analysis. For the negative class (0), the Random Forest model demonstrates a high precision of 0.90, meaning that 90% of the reviews predicted as negative were correctly identified as such. However, the recall for this class is 0.79, indicating that the model successfully identifies 79% of all actual negative reviews. The F1-score for this class is 0.84, which reflects a somewhat balanced performance between precision and recall, although the lower recall suggests some negative reviews are being missed. In contrast, for the positive class (1), the model exhibits a slightly lower precision of 0.83, meaning that 83% of reviews classified as positive were indeed positive. The recall, however, is higher at 0.92, indicating that the model captures 92% of all actual positive reviews effectively. The F1-score for this class is 0.88, higher than that of the negative class, indicating a better balance between precision and recall for the positive reviews. The macro average and weighted average for precision, recall, and F1-score are both reported as 0.86, reflecting consistent performance across the two classes. This demonstrates that the Random Forest model, while slightly more conservative in predicting negative sentiments, performs well in recognizing positive sentiments, making it a reliable tool for sentiment analysis, particularly when it is crucial to capture as many positive instances as possible.

The classification report for the XGBoost model offers a nuanced view of its performance in the sentiment classification task, analyzing a dataset comprising 402 reviews. The model achieves an overall accuracy of 0.86, indicating its proficiency in accurately classifying both positive and negative sentiments. For the negative class (0), the XGBoost model achieves a precision of 0.87, meaning that 87% of the reviews predicted as negative are correctly identified as such. The recall for this class stands at 0.81, suggesting that the model captures 81% of all actual negative reviews, while the F1-score is 0.84, reflecting a fairly balanced measure of precision and recall, though indicating a slight room for improvement in capturing negative sentiments more completely. In the case of the positive class (1), the model displays a precision of 0.84, with 84% of reviews classified as positive being correct. The recall is higher at 0.90, demonstrating that the model is effective at identifying 90% of the true positive cases in the dataset. The corresponding F1-score of 0.87 is indicative of a strong balance between precision and recall, highlighting the model's effectiveness in handling positive sentiments. The macro average and weighted average values for precision, recall, and F1-score are approximately 0.86, underscoring a consistent and balanced performance across both sentiment categories. This consistency shows that XGBoost, while slightly better at detecting positive sentiments, maintains a reliable performance across both classes, making it an effective tool for comprehensive sentiment analysis tasks.

The classification report for the Naive Bayes model details its effectiveness in sentiment analysis within a dataset of 402 reviews. The model exhibits an overall accuracy of 0.89, indicating a high level of precision in predicting both negative and positive sentiments accurately. For the negative class (0), Naive Bayes shows

a precision of 0.88, indicating that 88% of the instances it predicts as negative are accurately identified. The recall also stands at 0.88, meaning the model correctly identifies 88% of all actual negative reviews, which suggests a strong alignment between what the model predicts as negative and the true negative cases in the dataset. The F1-score, which harmonizes precision and recall, is similarly 0.88, underlining a balanced performance in identifying negative sentiments. Turning to the positive class (1), the precision is slightly higher at 0.89, with 89% of reviews classified as positive confirmed as true positives. The recall is also commendable at 0.90, showing that the model captures 90% of the actual positive sentiments, effectively ensuring that positive reviews are seldom overlooked. The F1-score for this class is 0.89, echoing the model's competency in accurately and reliably classifying positive reviews. Both the macro and weighted averages across precision, recall, and F1-scores are consistent at 0.89, demonstrating the model's uniform efficiency across sentiment categories. This uniformity in scores across both negative and positive classes indicates that the Naive Bayes model provides a dependable and balanced approach to sentiment classification, making it particularly useful for applications where equal importance is placed on accurately detecting both sentiment polarities.

### 3.2. Discussion

Based on shown in Figure 4, the implementation of TF-IDF in this study has influenced the accuracy of sentiment classification. Key terms such as "good" (*baik*), "excellent" (*bagus*), and "friendly" (*ramah*), which received high TF-IDF scores, consistently align with positive sentiment across the dataset. These terms are pivotal indicators of positivity, reflecting their frequent occurrence and importance in distinguishing positive sentiments from neutral or negative expressions. Conversely, the term "bad" (*buruk*), which also received a high TF-IDF score, serves as a crucial indicator of negative sentiment. Its prominence in TF-IDF rankings highlights its significance in identifying and distinguishing negative sentiments within the analyzed texts. Additionally, other words in the top 20 TF-IDF list similarly relate directly to the main sentiment categories to which the sentiment refers, further demonstrating TF-IDF's effectiveness in capturing and prioritizing terms critical for accurate sentiment classification. This analysis underscores TF-IDF's role in enhancing the overall performance and reliability of sentiment analysis models.

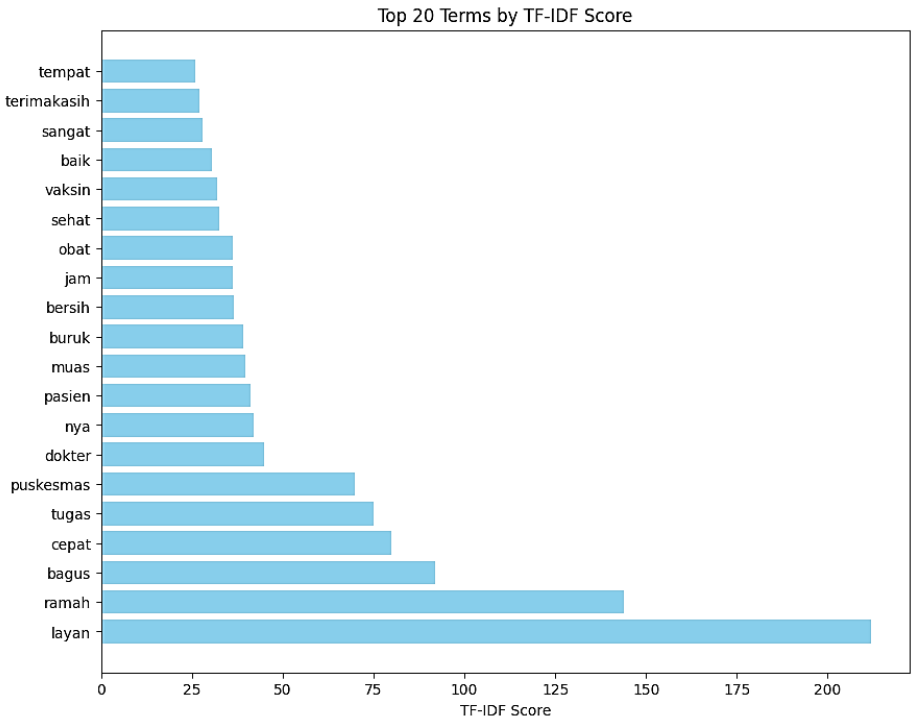


**Figure 4.** Top 20 Terms by TF-IDF Scores

In addressing the initial class imbalance where the dataset consisted of 800 instances of the negative class and 1000 instances of the positive class, applying SMOTE effectively balanced the class distribution. By generating synthetic samples for the minority negative class, SMOTE increased its representation until both classes reached an equal count of 1000 instances each as shown in Figure 5. This transformation mitigates the risk of biased model predictions that may favor the majority class, ensuring a more robust training process. By enhancing the dataset's balance, SMOTE enables the model to learn from and accurately classify instances across both classes, thereby improving overall prediction performance and reliability. The implementation of

SMOTE appears to have effectively balanced the performance of the models across both positive and negative classes. Analysis of the precision, recall, and F1 scores for various models, including Logistic Regression, SVM, XGBoost, Naive Bayes, and Random Forest, indicates that the metrics for the positive and negative classes are closely aligned. For instance, in Logistic Regression, the positive class exhibits a precision of 0.92, recall of 0.87, and F1-score of 0.89, while the negative class shows a precision of 0.86, recall of 0.91, and F1-score of 0.88. Similar trends are observed across the other models. These results suggest that SMOTE has successfully mitigated the class imbalance, resulting in more equitable model performance metrics, which is crucial for applications requiring balanced sensitivity and specificity.

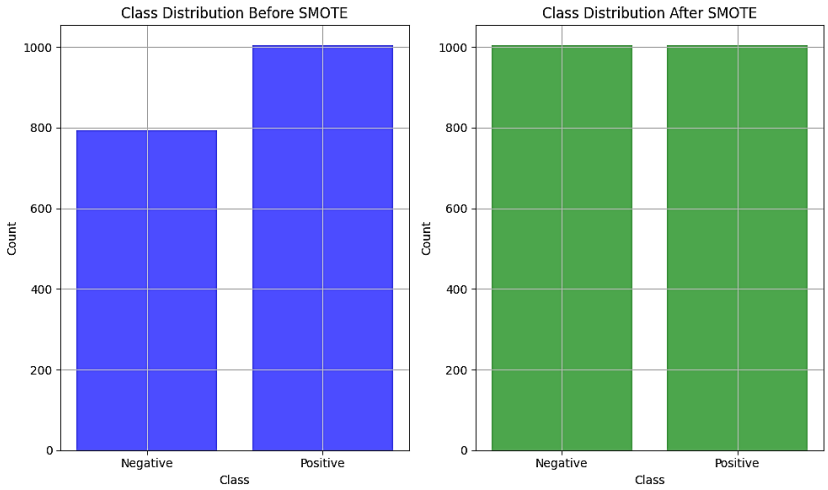


**Figure 5.** Class distribution before and after SMOTE

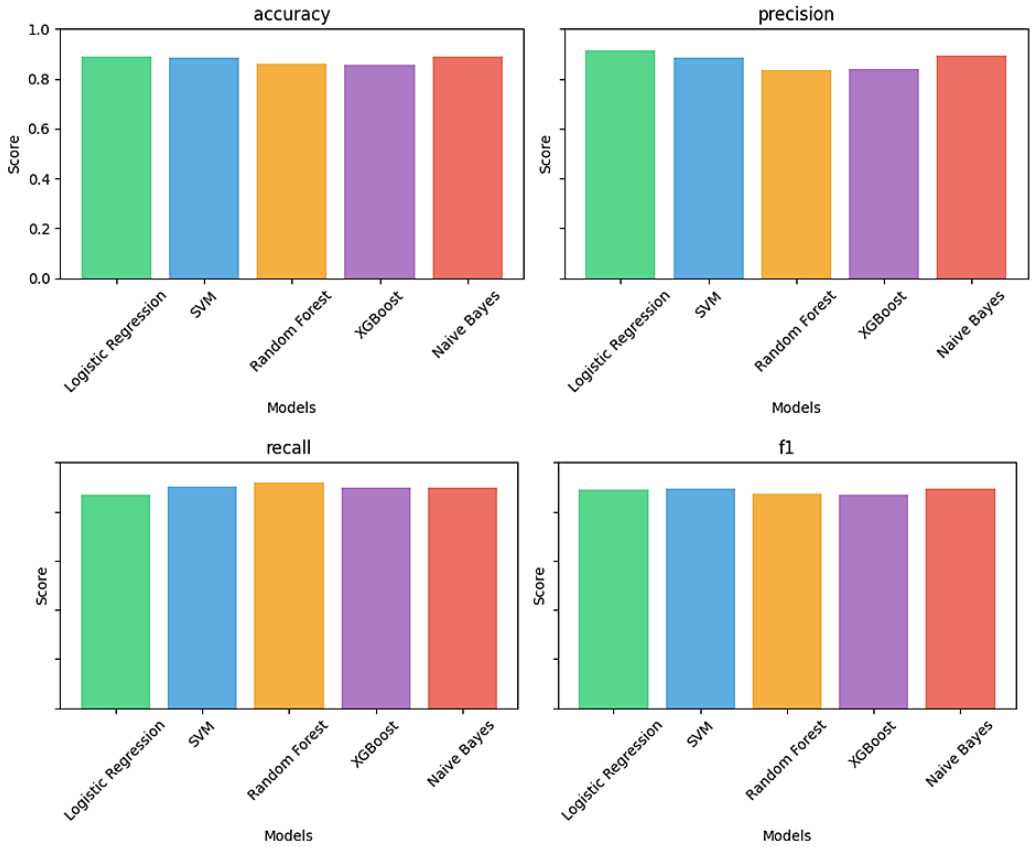


**Figure 6.** Algorithm Model Comparison

Based on shown in Figure 6 overall accuracy, precision, recall, computational efficiency, and the specific needs of the dataset. Logistic Regression and Naive Bayes both excelled with the highest accuracy at 0.89. Logistic Regression also demonstrated a commendable balance between precision and recall, making it

particularly useful for cases where false positives are a concern. On the other hand, Naive Bayes, known for its efficiency, is advantageous for handling large datasets or when computational resources are limited, despite its potential for oversimplification due to the assumption of feature independence. XGBoost and Random Forest are better suited for datasets with imbalances, with XGBoost showing a propensity to over-classify reviews as positive—a concern that could be addressed with parameter tuning. Random Forest, while excellent in precision for negative sentiments, had a lower recall for these cases, which might not be ideal if detecting every negative sentiment is critical. SVM, though providing robust performance across metrics, can be computationally demanding, especially with large or complex datasets. Given these nuances, Logistic Regression emerges as a particularly strong candidate for many sentiment analysis applications due to its high accuracy, interpretability, and effective handling of binary classifications. This model is beneficial not only for its performance but also for its capacity to provide insights into how different features influence predictions—valuable in scenarios requiring transparency and explainability. However, if dealing with particularly complex data characteristics, experimenting with models like XGBoost or Random Forest and adjusting their parameters might yield improved results.

## 4. CONCLUSION

The comparative analysis of the five algorithm in this research underscores the varied capabilities of each model in handling sentiment analysis tasks. The study incorporated TF-IDF for feature extraction, which effectively captured the importance of terms within the documents, contributing to improved model performance. Logistic Regression emerged as the most suitable model due to its high accuracy, balanced precision and recall, and excellent interpretability, making it highly applicable in environments where understanding the influence of input features on predicted outcomes is crucial. Naive Bayes proved to be highly effective, particularly noted for its speed and efficiency, beneficial for processing large data volumes. Models like XGBoost and Random Forest showed robustness in managing data complexities and imbalances, although they require careful tuning to optimize performance, particularly in improving recall rates. SVM, despite its robust performance, might pose challenges in scalability and computational efficiency with larger datasets. The implementation of SMOTE further balanced model performance across positive and negative classes, as seen in the closely aligned precision, recall, and F1 scores. However, a limitation of this study is the dataset size, which may impact the generalizability of the findings. Future research could explore hybrid models, ensemble methods, and advanced optimization techniques to enhance sentiment analysis accuracy and reliability, especially in complex and large-scale scenarios.

## REFERENCES

[1] R. Wijayanti and A. Arisal, "Automatic Indonesian sentiment lexicon curation with sentiment valence tuning for social media sentiment analysis," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1–16, 2021.

[2] D. Fimoza, A. Amalia, and T. H. F. Harumy, "Sentiment analysis for movie review in Bahasa Indonesia using BERT," in *2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, IEEE, 2021, pp. 27–34.

[3] Y. Fauziah, B. Yuwono, and A. S. Aribowo, "Lexicon based sentiment analysis in Indonesia languages: A systematic literature review," in *RSF Conference Series: Engineering and Technology*, 2021, pp. 363–367.

[4] J. Ipmawati, S. Saifulloh, and K. Kusnawi, "Analisis Sentimen Tempat Wisata Berdasarkan Ulasan pada Google Maps Menggunakan Algoritma Support Vector Machine: Sentiment Analysis of Tourist Attractions Based on Reviews on Google Maps Using the Support Vector Machine Algorithm," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 247–256, 2024.

[5] A. Harun and D. P. Ananda, "Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve bayes dan Decission Tree: Analysis of Public Opinion Sentiment About Covid-19 Vaccination in Indonesia Using Naïve Bayes and Decission Tree," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 1, no. 1, pp. 58–64, 2021.

[6] E. Ditendra, S. Suryani, S. Romelah, M. H. A. Tanjung, and M. Sarah, "Perbandingan Algoritma Klasifikasi untuk Analisis Sentimen Islam Nusantara di Indonesia: Comparison of Classification Algorithms for Sentiment Analysis of Islam Nusantara in Indonesia," *Malcom: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 1, pp. 71–77, 2022.

[7] I. G. B. A. Budaya, I. K. Dharmendra, D. P. Agustino, I. G. Harsemadi, I. M. P. P. Wijaya, and I. G. P. M. Yusadara, "Evaluation of Public Health Centers Performance through Sentiment Analysis using LSTM in Bali Province, Indonesia," in *2023 11th International Conference on Cyber and IT Service Management (CITSM)*, IEEE, 2023, pp. 1–6.

[8] O. A. El-Said, "Impact of online reviews on hotel booking intention: The moderating role of brand image, star category, and price," *Tour Manag Perspect*, vol. 33, p. 100604, 2020.
[9] V. Schoenmueller, O. Netzer, and F. Stahl, "The polarity of online reviews: Prevalence, drivers and implications," *Journal of Marketing Research*, vol. 57, no. 5, pp. 853–877, 2020.
[10] S. A. H. Bahtiar, C. K. Dewa, and A. Luthfi, "Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling," *Journal of Information Systems and Informatics*, vol. 5, no. 3, pp. 915–927, 2023.
[11] M. Z. Yumarlin, J. E. Bororing, and S. Rahayu, "Analisis Sentimen Terhadap Layanan Tokopedia Berdasarkan Twitter dengan Metode Klasifikasi Support Vector Machine," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 12, no. 1, pp. 153–163, 2023.
[12] J. Ipmawati, S. Saifulloh, and K. Kusnawi, "Analisis Sentimen Tempat Wisata Berdasarkan Ulasan pada Google Maps Menggunakan Algoritma Support Vector Machine: Sentiment Analysis of Tourist Attractions Based on Reviews on Google Maps Using the Support Vector Machine Algorithm," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 247–256, 2024.
[13] K. Afifah, I. N. Yulita, and I. Sarathan, "Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier," in *2021 international conference on artificial intelligence and Big data analytics*, IEEE, 2021, pp. 22–27.
[14] A. Samih, A. Ghadi, and A. Fennan, "Enhanced sentiment analysis based on improved word embeddings and XGboost.," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 13, no. 2, 2023.
[15] B. Warsito and A. Prahutama, "Sentiment analysis on tokopedia product online reviews using random forest method," in *E3S Web of Conferences*, EDP Sciences, 2020, p. 16006.
[16] S. Khomsah, "Sentiment analysis on youtube comments using word2vec and random forest," *Telematika: Jurnal Informatika dan Teknologi Informasi*, vol. 18, no. 1, pp. 61–72, 2021.
[17] I. G. B. A. Budaya, L. P. S. Pratiwi, and D. P. Agustino, "Klasifikasi Sentimen untuk Analisis Kepuasan Pelayanan Puskesmas Berbasis Arsitektur LSTM," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 12, no. 4, pp. 941–948, 2023.
[18] S. George and V. Srividhya, "Performance evaluation of sentiment analysis on balanced and imbalanced dataset using ensemble approach," *Indian J Sci Technol*, vol. 15, no. 17, pp. 790–797, 2022.
[19] S. Singh, K. Kumar, and B. Kumar, "Sentiment analysis of Twitter data using TF-IDF and machine learning techniques," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, IEEE, 2022, pp. 252–255.
[20] W. A. Prabowo and F. Azizah, "Sentiment analysis for detecting cyberbullying using tf-idf and svm," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 4, no. 6, pp. 1142–1148, 2020.
[21] S. Kedas, A. Kumar, and P. K. Jain, "Dealing with Class Imbalance in Sentiment Analysis Using Deep Learning and SMOTE," in *Advances in Data Computing, Communication and Security: Proceedings of I3CS2021*, Springer, 2022, pp. 407–416.
[22] D. P. Chatterjee, S. Mukhopadhyay, S. Goswami, and P. K. Panigrahi, "Efficacy of oversampling over machine learning algorithms in case of sentiment analysis," in *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2020, Volume 2*, Springer, 2021, pp. 247–260.
[23] M. H. Setiawan, I. G. A. Gunadi, and G. Indrawan, "Klasifikasi Pelayanan Kesehatan Berdasarkan Data Sentimen Pelayanan Kesehatan menggunakan Multiclass Support Vector Machine," *Jurnal Sistem dan Informatika (JSI)*, vol. 17, no. 1, pp. 47–54, 2022.