Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (303-315)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.46452

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Sentiment Classification in Imbalanced Data: Trade-Offs Between Metrics and Real-World Relevance

Indra Swanto Ritonga[1*], Wanayumini[2], Dedy Hartama[3]

[1]Master of Computer Science, Faculty of Computer Science, Potensi Utama University
[2]Informatics Engineering, Faculty of Engineering, Asahan University
[3]Information System, STIKOM Tunas Bangsa
[1]Jl. K.L Yos Sudarso, Gg. Famili No. 247, Medan, Indonesia
[2]Jl. Ahmad Yani No. 1, Kisaran Kota, Asahan, Sumatera Utara, Indonesia
[3]Jl. Kartini, Proklamasi, Pematang Siantar, Sumatera Utara, Indonesia

Article:
Accepted: June 20, 2025
Revised: April 30, 2025
Issued: October 30, 2025

***Correspondence Address:**
Indraswanto84@gmail.com

ABSTRACT

Sentiment analysis plays a crucial role in assessing public perception, particularly in healthcare services like BPJS Kesehatan, Indonesia's national health insurance program. However, sentiment classification faces a challenge due to class imbalance, where negative feedback dominates positive responses. This study investigates whether sentiment classification should prioritize traditional evaluation or maintain real-world data representation by preserving the original sentiment distribution. Two feature extraction methods, Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW), were evaluated using Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression with varying maximum feature counts (100–300) to examine the impact of feature dimensionality. Model performance was evaluated using traditional metrics, while sentiment distribution fidelity was assessed by comparing predicted proportions with the dataset. Results show TF-IDF achieves higher precision and recall but fails to capture positive sentiments, leading to a skewed representation of real-world trends, while BoW offers a more balanced distribution with slightly lower accuracy. Paired t-tests and Wilcoxon signed-rank tests confirmed differences in accuracy and recall are significant, but not in precision and sentiment distribution. These findings highlight a trade-off between performance and sentiment diversity, vital in healthcare services and other fields with imbalanced datasets, emphasizing the need to align evaluation metrics with real-world objectives. Future research should investigate advanced models, such as deep learning and transformer-based approaches, to enhance both accuracy and fairness when analyzing imbalanced data.

Keywords : *bag of words; bpjs Kesehatan; class imbalance; naïve bayes; sentiment analysis; text feature extraction; TF-IDF.*

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (303-315)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.46452

1. INTRODUCTION

Sentiment analysis has become an essential tool in evaluating public perception, particularly in service sectors such as healthcare [1]. One such critical area is the evaluation of public sentiment toward BPJS Kesehatan, Indonesia's national health insurance program [2]. BPJS Kesehatan represents public sentiment toward public services and reflects a broad social reality. Public sentiment toward these services is diverse, encompassing a range of aspects, from economic to emotional to existential. All of this results in strong and often unbalanced sentiments. The predominance of negative feedback leads to the neglect of positive feedback, which is crucial for supporting policy improvements and enhancing service quality. This makes BPJS an ideal case study for sentiment analysis because it is highly relevant: it reflects broad public perception, addresses basic community needs, and triggers strong opinions. Therefore, this research can be generalized to various domains due to the large-scale, unbalanced nature of the data and its origins in the context of public services. With a vast amount of textual data available from user feedback, extracting meaningful insights requires effective text feature extraction methods and a well-structured classification model.

Traditional approaches like Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) have been widely used to transform text into numerical representations for machine learning models [3], such as the Naïve Bayes classifier [4], SVM [18], and Logistic Regression [19]. However, when dealing with sentiment classification, a significant challenge arises-class imbalance [5]. Class imbalance occurs when the amount of data in each class in a dataset is unbalanced, resulting in one class (the majority class) having significantly more data than the other (the minority class). Class imbalance can cause machine learning models to be biased or favor the dominant class and perform poorly in predicting the minority class [6]. This occurs because the model tends to learn from a large amount of frequently occurring data, even though the minority class is often very important.

In real-world sentiment data, negative feedback is often more dominant than positive or neutral responses, leading to skewed distributions that affect model performance [7]. This imbalance raises an important question: should model development prioritize evaluation metrics such as accuracy, precision, and recall, or should it focus on ensuring that the model aligns with the actual data distribution and sentiment trends [8]? Many classification models perform well in controlled conditions but struggle when applied to real-world datasets, where misclassification of minority classes can lead to misleading conclusions [9], [10], [11]. Research related to sentiment classification on imbalanced data was conducted by [12], The study conducted sentiment analysis on restaurant review data using an ensemble approach (EBSVM) and the SMOTE resampling technique. This approach has been proven effective in improving accuracy and F1-Score, but there are limitations such as model generalization that is only carried out on one type of data without comparing it with a more diverse public dataset, the absence of statistical tests to determine the significance of model performance and not evaluating the trade-off between increased accuracy and changes in the original sentiment distribution after balancing. For instance, in healthcare, underestimating negative feedback could result in inadequate policy adjustments, ultimately affecting service quality [13].

To address this issue, this research investigates the trade-off between optimizing evaluation metrics and preserving real-world data representation in sentiment classification [14]. Specifically, the study compares the performance of TF-IDF and BoW as feature extraction methods [15], each tested with different maximum feature variants (100, 150, 200, 250, and 300). The machine learning algorithms used in this study to perform classification are Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression.

The Naïve Bayes classifier, chosen for its simplicity and effectiveness in text classification [16], is used to categorize sentiments based on these extracted features and the corresponding labels in the dataset [17]. SVM are used because of their ability to handle high-dimensional data, they are utilized for text categorization in situations when there are more features than samples [18], and the logistic regression algorithm is used because it is an interpretable model, effective in high-dimensional spaces, and can be combined well with TF-IDF and BoW [19].

Model performance is evaluated using accuracy, precision, and recall as benchmark metrics, while the classification results are assessed against real-world data representation by comparing the proportion of predicted positive and negative sentiments to the original data distribution.

The findings of this study aim to determine whether prioritizing evaluation metrics or maintaining real-world data representation leads to a more optimal sentiment classification model. By addressing this trade-off, the research provides valuable insights for handling class imbalance in sentiment analysis, with implications not only for healthcare but also for other domains where imbalanced datasets are prevalent.

2. METHODS

This study investigates the trade-off between evaluation metrics and real-world data representation in sentiment classification using TF-IDF and Bag of Words (BoW) as feature extraction methods. The methodology consists of five key stages: data collection and preprocessing, feature extraction, sentiment classification, evaluation, and analysis.

2.1. Data Collection and Preprocessing

The sentiment dataset consists of user feedback related to BPJS Kesehatan, Indonesia's national health insurance program. The dataset, collected from Kaggle.com (https://www.kaggle.com/datasets/aeworld/sentimen-id-bpjs) [20], includes 15300 labeled sentiment classes (Positive and Negative), with 4.84% positive and 95.16% negative samples. However, the absence of information about the data collection period in the metadata or documentation of the dataset, with the last data recorded in 2022.

Before analysis, the following preprocessing steps were applied:

a. Text Cleaning

Removal of punctuation, stopwords, and special characters.

b. Tokenization

Splitting text into individual words or tokens.

c. Normalization

Converting words into lowercase and applying stemming to unify word forms.

Given the high imbalance in sentiment distribution (4.84% Positive vs. 95.16% Negative), no resampling or rebalancing techniques were applied, as the objective was to evaluate whether models should prioritize evaluation metrics or maintain real-world data representation. However, the impact of class imbalance on classification performance was analyzed to assess potential biases.

2.2. Feature Extraction Methods

Two common text feature extraction techniques were used to convert the processed text into numerical representations:

a. Term Frequency-Inverse Document Frequency (TF-IDF)

This method was chosen for its ability to reflect the importance of words in a document relative to a corpus [21].

b. Bag of Words (BoW)

This method was selected for its simplicity and effectiveness in capturing word frequencies [22].

To assess the impact of feature dimensionality, both methods were tested with five different maximum feature variants: 100, 150, 200, 250, and 300 features. These values were chosen to explore the trade-off between model complexity and performance.

2.3. Sentiment Classification Using Naïve Bayes, SVM, and Logistic Regression

The extracted features were used to train a Naïve Bayes classifier, Support Vector Machine, and Logistic Regression. Naïve Bayes classifier is a probabilistic mode l widely used in text classification due to its efficiency and robustness [23]. Naïve Bayes was selected because it performs well with high-dimensional text data and is computationally efficient [24]. Support Vector Machine (SVM) is a machine learning algorithm used to analyze data and recognize patterns, where data is grouped into two or more classes, aiming to determine the most optimal dividing boundary (hyperplane) between these data groups [25]. Meanwhile, logistic regression is an algorithm used in classification and predictive analysis. This algorithm utilizes a linear regression equation to produce a discrete binary output. The sigmoid function is used in this algorithm to transform the predicted results into the range 0 and 1 [19]. The logistic regression algorithm is used

because it is an interpretable model and can be combined well with TF-IDF and BoW.

To ensure robust evaluation of the model's performance, 10-fold cross-validation was employed. Stratified sampling was used to ensure that each fold maintained the same proportion of positive and negative sentiments as the original dataset [26].

This technique involves partitioning the dataset into 10 subsets, training the model on 9 subsets, and validating it on the remaining subset. This process is repeated 10 times, with each subset used exactly once as the validation set. By averaging the results across all folds, 10-fold cross-validation provides a more reliable estimate of the model's generalization performance compared to a simple train-test split. The model was trained and evaluated using this approach, ensuring that the results are robust and less prone to overfitting [27].

2.4. Evaluation Metrics and Benchmarking

To assess model performance, two evaluation perspectives were used:

a. Traditional Evaluation Metrics [28]
   Accuracy: Measures the overall correctness of predictions.
   Precision: Evaluates the proportion of correctly predicted positive sentiments.
   Recall: Measures the model's ability to identify all actual positive sentiments.

b. Real-World Data Representation

The predicted sentiment distributions were compared with the actual distribution of Positive and Negative sentiments in the dataset. This comparison was quantified using percentage difference, ensuring the model preserves real-world sentiment proportions. o quantify the alignment between model predictions and real-world sentiment distribution, percentage difference was calculated as:

$$Positive\ Difference = \frac{P_{pred} - P_{act}}{P_{act}} \quad (1)$$

$$Negative\ Difference = \frac{N_{pred} - N_{act}}{N_{act}} \quad (2)$$

Information :

P_{pred}	=	Predicted Positive Rate
P_{act}	=	Actual Positive Rate
N_{pred}	=	Predicted Negative Rate
N_{act}	=	Actual Negative Rate

A lower percentage difference indicates a better alignment with real-world sentiment distribution. A model with a difference below 5% is considered to have high representational fidelity, while models exceeding 10% indicate a significant deviation from real-world sentiment trends. Regarding the balancing technique used to assess how well the model works, this study uses undersampling and oversampling techniques as additional analysis, not used to perform sentiment distribution or statistical tests.

2.5. Comparative Analysis

The final step involved comparing models based on two perspectives:

a. Which model performs better according to accuracy, precision, and recall?
b. Which model preserves the original sentiment distribution more accurately?

The results of this analysis were used to determine whether optimizing evaluation metrics or maintaining real-world data balance leads to a more reliable sentiment classification model. Paired t-tests were used to compare accuracy, precision, and recall scores across different feature extraction techniques because they are suitable for comparing paired samples. Wilcoxon signed-rank tests were applied to compare real-world sentiment representation deviations, as they are non-parametric and robust to non-normal distributions.

3. RESULTS AND DISCUSSION

This section presents the findings of the research, focusing on the performance of TF-IDF and Bag of Words (BoW) as feature extraction methods for sentiment classification using a Naïve Bayes classifier. The results are organized into four subsections: (1) the impact of preprocessing on the dataset [29], (2) model performance on traditional evaluation metrics (accuracy, precision, and recall) [30], (3) the alignment of predicted sentiment distributions with real-world data [31], and (4) statistical comparisons to assess the significance of observed differences [32]. Key findings include the trade-offs between evaluation metrics and real-world sentiment representation, the optimal feature count for each method, and the statistical significance of performance differences. These insights provide valuable guidance for selecting feature extraction methods in imbalanced

sentiment analysis tasks, particularly in applications like healthcare.

3.1. Preprocessing Results

The preprocessing steps significantly improved the quality of the dataset by removing noise and standardizing word forms. On average, the text length was reduced from 16.05 words before preprocessing to 10.13 words after preprocessing. This reduction highlights the elimination of stopwords, punctuation, and special characters, leading to a cleaner and more efficient dataset for feature extraction. By reducing unnecessary elements, the model can focus on meaningful words, improving classification accuracy. Table 1 presents an example of the preprocessing transformation:

Table 1. *Sample of Preprocessing Result*

	Text
Before	Sekalian kalau tidak punya BPJS tidak Bisa bayar pajak oke
After	bpjs bayar pajak oke

Note: The example text demonstrates the removal of stopwords ('sekalian,' 'kalau,' 'tidak'), punctuation, and special characters, as well as the conversion to lowercase and stemming.

3.2. Model Performance on Traditional Evaluation Metrics

To evaluate model performance, accuracy, precision, and recall were measured for TF-IDF + Naïve Bayes, BoW + Naïve Bayes, TF-IDF + SVM, BoW + SVM, TF-IDF + Logistic Regression, and BoW + Logistic Regression across different maximum feature counts (100, 150, 200, 250, and 300). The results are summarized in Tables 2 to 7 and visualized in Figures 1 to 3.

Table 2. *Evaluation Metric for TF-IDF + Naïve Bayes*

Max Feature	Accuracy	Precision	Recall
100	0.951634	0.905607	0.951634
150	0.951961	0.954269	0.951961
200	0.951634	0.905607	0.951634
250	0.951634	0.905607	0.951634
300	0.951634	0.905607	0.951634

Table 3. *Evaluation Metric for BoW + Naïve Bayes*

Max Feature	Accuracy	Precision	Recall
100	0.946078	0.931043	0.946078
150	0.944771	0.9314	0.944771
200	0.94281	0.931005	0.94281
250	0.942157	0.93107	0.942157
300	0.939216	0.929489	0.939216

The results provide valuable insights into the impact of feature extraction methods and feature count on model performance:

a. Best-Performing Feature Set

For TF-IDF, precision peaked at 150 features (0.9543), showing a notable improvement over other feature sets, while accuracy and recall remained stable (~0.9516).

For BoW, precision peaked at 150 features (0.9314), but accuracy showed a slight downward trend as feature count increased, suggesting diminishing returns.

b. Impact of Feature Count

TF-IDF: Increasing the feature count beyond 150 did not yield further performance gains, with accuracy and recall stabilizing around 0.9516 and precision fluctuating.

BoW: Accuracy showed a slight decline as feature count increased, possibly due to added noise in higher-dimensional representations.

c. Variation in Metrics

TF-IDF: Precision fluctuated the most, varying between 0.9056 and 0.9543, whereas recall remained stable at 0.9516 across feature sets.

BoW: Precision exhibited minor fluctuations, while recall stayed relatively consistent across all feature sets.

Table 4. *Evaluation Metric for TF-IDF + SVM*

Max Feature	Accuracy	Precision	Recall
100	0.951634	0.905607	0.951634
150	0.951307	0.921995	0.951307
200	0.950980	0.905577	0.950980
250	0.951307	0.905592	0.951307
300	0.952941	0.943632	0.952941

Table 5. *Evaluation Metric for BoW + SVM*

Max Feature	Accuracy	Precision	Recall
100	0.951634	0.905607	0.951634
150	0.951307	0.931452	0.951307
200	0.950980	0.931718	0.950980
250	0.949673	0.930910	0.949673
300	0.949673	0.934261	0.949673

The results provide valuable insights into the impact of feature extraction methods and feature count on model performance:

a. Best-Performing Feature Set

TF-IDF: Achieved the highest accuracy value of 0.9529 and precision of 0.9436. This indicates that adding up to 300 features can consistently improve model performance.

BoW: Achieved the highest precision value of 0.9343 at 300 features. However, accuracy and recall decreased, indicating that

adding features did not significantly improve overall model performance.

b. Impact of Feature Count

TF-IDF: There was no significant increase in accuracy or recall when increasing the number of features from 100 to 300, remaining stable at around 0.9516. However, precision fluctuated across feature sets, indicating that increasing the number of features does not always guarantee improved model consistency.

BoW: Accuracy decreased with increasing feature number despite a slight increase in precision, indicating that additional features can introduce noise into high-dimensional data representations.

c. Variation in Metrics

TF-IDF: Precision was the most volatile metric, ranging from 0.9056 to 0.9436. Recall, on the other hand, remained stable at 0.9516, indicating that the model consistently detected all relevant instances.

BoW: Precision showed significant variation, ranging from 0.9076 to 0.9314, while recall was more stable, ranging from 0.9372 to 0.9444.

Table 6. *Evaluation Metric for TF-IDF + Logistic Regression*

Max Feature	Accuracy	Precision	Recall
100	0.951961	0.954269	0.951961
150	0.952288	0.942756	0.952288
200	0.952288	0.942756	0.952288
250	0.952288	0.942756	0.952288
300	0.952614	0.954862	0.952614

Table 7. *Evaluation Metric for BoW + Logistic Regression*

Max Feature	Accuracy	Precision	Recall
100	0.952941	0.937842	0.952941
150	0.951634	0.934046	0.951634
200	0.951634	0.934332	0.951634
250	0.949020	0.928844	0.949020
300	0.949020	0.929386	0.949020

The results provide valuable insights into the impact of feature extraction methods and feature count on model performance:

a. Best-Performing Feature Set

TF-IDF: Achieved high precision at 300 features, with a value of 0.9549, followed by 0.9543 at 100 features. This indicates that both small and large feature sets can yield high precision.

BoW: At 100 features, the result was 0.9378, which then decreased with increasing features. The highest accuracy was achieved at 100 features, with a value of 0.9529, indicating the best performance achieved with a smaller feature set.

b. Impact of Feature Count

TF-IDF: Adding features from 100 to 300 only provides a small increase in precision, with no significant impact on accuracy or recall.

BoW: Accuracy decreased from 0.9529 at 100 features to 0.9490 at 300 features. Precision also showed a downward trend, indicating that adding features increases noise and decreases performance.

c. Variation in Metrics

TF-IDF: Precision fluctuated from 0.9428 to 0.9549, while accuracy and recall remained stable. This indicates that the model tends to be more selective towards positive predictions.

BoW: All metrics decreased with the addition of features, indicating a general performance degradation at higher feature dimensions.

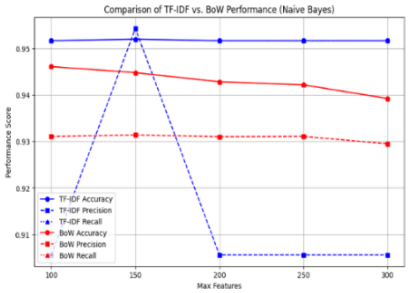

Figure 1. *Evaluation metric comparison naïve bayes*

The graph in Figure 1 shows that TF-IDF + Naïve Bayes achieves peak precision at 150 features but declines afterward, while accuracy and recall remain stable. However, the precision and recall graphs appear overlapping because they have very close or identical values for most of the feature counts. This indicates that the model can maintain a balance between precision and recall and reflects the performance of the TF-IDF model in handling text data across a wide range of feature counts. In contrast, BoW + Naïve Bayes maintains more consistent precision but experiences a slight decline in accuracy and recall as feature count increases. This suggests that TF-IDF benefits from careful feature selection, whereas BoW is more stable but less sensitive to feature count adjustments.

Overall, 150 features appear optimal for TF-IDF, while BoW exhibits gradual performance degradation with more features. These findings indicate that while TF-IDF achieved higher peak precision, BoW demonstrated greater consistency across feature variations.

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (303-315)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.46452

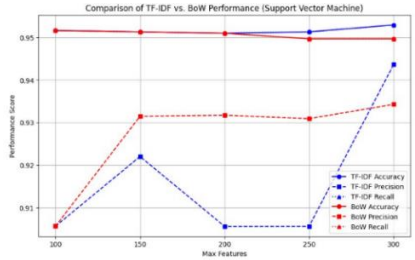

Figure 2. *Evaluation metric comparison support vector machine*

Figure 2 displays a graph of TF-IDF + SVM, which initially had a low accuracy of around 0.91-0.92 but showed improvements in accuracy, precision, and recall as the number of features increased, especially at 300 features. In contrast, the graph of BoW + SVM displays relatively stable performance, with accuracy remaining above 0.95 and not changing significantly even as the number of features increases. The precision and recall for BoW rose initially at 15 features, then slightly decreased again. This indicates that enriching text representation with TF-IDF benefits from increasing the number of features. Meanwhile, BoW tends to be more consistent and stable, although its performance tends to plateau at a certain level due to being less responsive to increases in features.

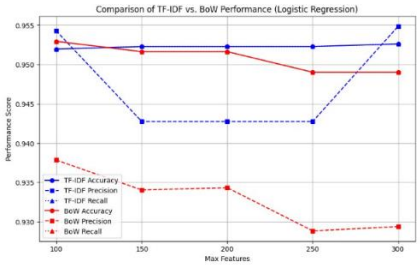

Figure 3. *Evaluation metric comparison logistic regression*

Figure 3 presents a graph of TF-IDF combined with Logistic Regression, which maintains stable accuracy but shows a drop in precision and recall at around 150 features, before increasing again at 300 features. This suggests that TF-IDF benefits from a larger number of features to achieve optimal performance. Meanwhile, BoW combined with Logistic Regression demonstrates fairly consistent results across different feature counts, but remains below TF-IDF for all metrics. This indicates that BoW is relatively stable but does not significantly improve as the number of features grows, showing less sensitivity to feature variations. Overall, TF-IDF performs better with increasing features, while BoW remains stable but less effective overall.

Overall, the features appear optimal for TF-IDF, while BoW experiences a gradual decline in performance as features increase. Thus, TF-IDF excels with high precision, while BoW performs stably.

To better illustrate the classification results, the confusion matrix of one of the models used, namely Naïve Bayes with TF-IDF, is shown in Figure 4.

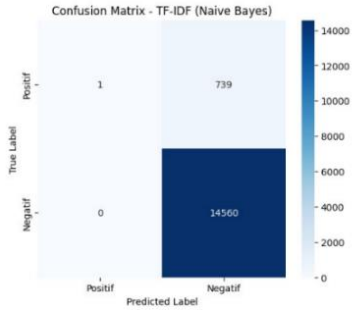

Figure 4. *Confusion matrix naïve bayes with TF-IDF*

This model shows a significant imbalance in model predictions, where it tends to classify data into the negative class predominantly, thus indicating an imbalance in the classification of positive and negative labels, so that a data balancing approach is needed.

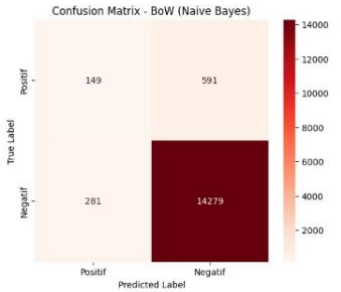

Figure 5. *Confusion matrix naïve bayes with BoW*

In Figure 5, when compared with Figure 4, it can be seen that BoW is better able to recognize the class with the label "positive". Although TF-IDF shows higher accuracy, BoW is more suitable for real-world use, especially when positive sentiment is less than negative sentiment.

To address this class imbalance, a balancing technique using random undersampling and random oversampling was implemented. This technique was used to assess the model's performance and served only as a supplementary analysis; it was not used to perform sentiment distribution or statistical

tests in this study. The use of the random undersampling technique is shown in Table 8 as follows.

Table 8. *Random Undersampling*

Max Feature	Accuracy	Precision	Recall
100 TF-IDF	0.672297	0.672305	0.672297
150 TF-IDF	0.702703	0.702851	0.702703
200 TF-IDF	0.699324	0.699771	0.699324
250 TF-IDF	0.716216	0.717208	0.716216
300 TF-IDF	0.722973	0.723626	0.722973
100 BoW	0.679054	0.679719	0.679054
150 BoW	0.695946	0.696520	0.695946
200 BoW	0.699324	0.699552	0.699324
250 BoW	0.712838	0.712848	0.712838
300 BoW	0.706081	0.706090	0.706081

Table 8, using the random undersampling method, shows consistency across feature dimensions. This is understandable because the undersampling technique reduces the amount of data from the majority class, thereby risking the loss of important information and causing the model to lose generalization capacity. The use of the random oversampling technique is shown in Table 9 as follows.

Table 9. *Random Oversampling*

Max Feature	Accuracy	Precision	Recall
100 TF-IDF	0.734203	0.734262	0.734203
150 TF-IDF	0.774382	0.774383	0.774382
200 TF-IDF	0.786058	0.786060	0.786058
250 TF-IDF	0.818510	0.819061	0.818510
300 TF-IDF	0.825034	0.825888	0.825034
100 BoW	0.716346	0.717927	0.716346
150 BoW	0.737294	0.738944	0.737294
200 BoW	0.755151	0.757107	0.755151
250 BoW	0.780220	0.781090	0.780220
300 BoW	0.790179	0.790740	0.790179

Table 9, using the random oversampling technique, yields higher performance across all evaluation metrics. These results indicate that balancing classes by adding copies of data from the minority class is more effective in improving model performance than reducing data from the majority class.

The results in Tables 8 and 9 demonstrate that the application of the balancing technique significantly impacts the performance of the classification model. The random oversampling technique proved superior on the imbalanced data used in this study, especially when used in conjunction with the high-dimensional TF-IDF feature representation.

3.3. Real-World Sentiment Representation Analysis

To assess whether the models accurately reflect real-world sentiment proportions, we examine the predicted sentiment distributions at different feature counts. Tables 10 and 11 summarize the proportion of positive and negative responses predicted by the models.

Table 10. *Predicted Sentiment Distribution (TF-IDF + Naïve Bayes)*

Max Feature	Positive Response	Negative Response
100	0%	100%
150	0.03268%	99.96732%
200	0%	100%
250	0%	100%
300	0%	100%

Table 11. Predicted Sentiment Distribution (BoW + Naïve Bayes)

Max Feature	Positive Response	Negative Response
100	2.254902%	97.745098%
150	2.581699%	97.418301%
200	2.908497%	97.091503%
250	3.039216%	96.960784%
300	3.333333%	96.666667%

The results reveal significant differences between TF-IDF and BoW in sentiment representation:

a. Which feature extraction method better preserves sentiment distribution?

BoW outperforms TF-IDF in maintaining real-world sentiment proportions, as it consistently predicts a higher percentage of positive responses. In contrast, TF-IDF nearly always predicts negative sentiment, leading to an imbalanced distribution.

b. Which feature dimension gives the most accurate representation?

For BoW, increasing feature count improves sentiment balance, with the most accurate representation at 300 features (3.33% positive).

TF-IDF fails to capture positive sentiment, regardless of feature count.

c. Is there a trade-off between accuracy and sentiment distribution?

While TF-IDF achieves high accuracy (see Table 2), its sentiment predictions are skewed, potentially making it unsuitable for real-world applications.

BoW shows a slight accuracy decline (Table 3), but better maintains sentiment distribution, suggesting a trade-off between accuracy and real-world representativeness.

These findings indicate that while TF-IDF achieves higher precision, it significantly underestimates positive sentiments, which could lead to misleading conclusions in real-world applications. A few examples of incorrect classifications show how TF-IDF has trouble

detecting positive emotion. One example of a misclassified negative was "Awalnya saya kecewa, tapi ternyata pelayanan sekarang sangat baik." This was probably caused by the influence of high-frequency negative phrases like "kecewa." Similarly, "Saya sangat terbantu dengan adanya BPJS, meski antre." was also incorrectly classified, suggesting that TF-IDF tends to underrepresent positive context and overweight isolated negative phrases. These examples show that TF-IDF is not semantically sensitive enough to handle nuanced or conflicting sentiments. BoW, on the other hand, provides a more balanced representation but with slightly lower precision. This highlights the trade-off between evaluation metrics and real-world data representation, which is critical for applications like sentiment analysis in healthcare.

3.4. Statistical Comparison (Paired t-Test & Wilcoxon Test Results)

To determine whether the performance differences between TF-IDF and BoW are statistically significant, a paired t-test was conducted for accuracy, precision, and recall, while a Wilcoxon signed-rank test was used for sentiment distribution. Table 12 presents the p-values for each comparison.

Table 12. *Statistical Test Results (Naïve Bayes)*

Metric	Paired t-Test (p-value)	Wilcoxon Test (p-value)
Accuracy	0.0017	-
Precision	0.1821	-
Recall	0.0017	-
Sentiment Distribution	-	0.0625

The results indicate the following:

a. Accuracy and Recall (p = 0.0017, statistically significant)

Since the paired t-test p-value is below 0.05, the accuracy and recall differences between TF-IDF and BoW are statistically significant. This means that one method consistently outperforms the other in these metrics, and the observed difference is unlikely due to chance.

b. Precision (p = 0.1821, not statistically significant)

Since the p-value is above 0.05, the precision difference between TF-IDF and BoW is not statistically significant. This suggests that both methods perform similarly in terms of precision, and any observed differences might be due to random variation rather than a meaningful performance gap.

c. Sentiment Distribution (p = 0.0625, not statistically significant)

The Wilcoxon signed-rank test p-value is slightly above 0.05, meaning the difference in sentiment representation between TF-IDF and BoW is not statistically significant at the conventional 5% level. However, since 0.0625 is relatively close to 0.05, there may still be a slight trend toward a difference, but it is not strong enough to be conclusive. This result, while inconclusive, could achieve significance with a larger dataset if there are small, significant differences.

These findings suggest that while TF-IDF may be more reliable for applications where accuracy and recall are critical, both methods are comparable in terms of precision and sentiment representation. This highlights the importance of considering both statistical significance and practical implications when choosing a feature extraction method for sentiment analysis.

In this discussion, we delve into the comparative analysis of Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) feature extraction methods within the realm of sentiment analysis. Our investigation reveals that while TF-IDF demonstrates higher precision and recall, it notably underrepresents positive sentiments, resulting in a skewed depiction of real-world data. Conversely, BoW offers a more balanced sentiment distribution, albeit with a slight compromise in accuracy. Statistical evaluations confirm that the disparities in accuracy and recall between these methodologies are significant, whereas differences in precision and sentiment distribution are not.

3.5. Summary of Key Findings

The study revealed that TF-IDF achieves higher precision and recall but significantly underestimates positive sentiments, leading to a skewed representation of real-world data. In contrast, BoW maintains a more balanced sentiment distribution but with slightly lower accuracy. Statistical tests confirmed that the differences in accuracy and recall between the two methods are significant, while precision and sentiment distribution differences are not.

3.6. Interpretation of Results

The superior precision of TF-IDF can be attributed to its ability to reflect the importance of words in a document relative to a corpus. However, its failure to capture positive sentiments suggests that it may not be suitable for applications where balanced representation is critical, such as healthcare sentiment analysis. BoW, while less precise, provides a more accurate reflection of real-world sentiment proportions, making it a better choice for applications requiring balanced representation.

3.7. Comparison with Previous Studies

These findings align with previous studies highlighting the challenges of class imbalance in sentiment analysis [33], [34], [35]. While TF-IDF is often favored for its precision, its limitations in handling imbalanced datasets have been noted in other domains [36]. BoW's consistency across feature variations is consistent with its reputation for simplicity and robustness, though its performance may vary depending on the dataset [37].

3.8. Addressing the Research Questions

The study answers the research questions as follows:

a. Should model development prioritize evaluation metrics or real-world data representation?

The results suggest that the choice depends on the application. For tasks requiring high precision, TF-IDF may be preferable, but for applications where balanced representation is critical, BoW is more suitable.

b. Which feature extraction method is more suitable for imbalanced sentiment analysis?

BoW is more suitable for imbalanced sentiment analysis due to its ability to maintain real-world sentiment proportions, despite slightly lower accuracy.

3.9. Practical Implications

These findings have important implications for sentiment analysis in healthcare and other domains with imbalanced datasets. For example, healthcare policymakers relying on sentiment analysis to gauge public perception of services should prioritize methods like BoW that provide balanced representation, even if they sacrifice some accuracy.

3.10. Limitations of the Study

This study has several limitations. First, the dataset, while substantial, is highly imbalanced, which may affect the generalizability of the results. Second, the use of Naïve Bayes, SVM, and Logistic Regression as classifiers limits the exploration of how other models can handle class imbalance. Finally, the study focused only on TF-IDF and BoW, leaving room for future research on other feature extraction methods.

3.11. Future Work

Future research could explore the following directions:

a. Testing other classifiers, such as Random Forest or deep learning models, to see if they handle class imbalance better.
b. Investigating advanced feature extraction methods, such as word embeddings (e.g., Word2Vec, GloVe) or transformer-based models (e.g., BERT).
c. Applying the methods to other domains with imbalanced datasets to validate the findings.

CONCLUSION

This study compared TF-IDF and BoW for sentiment analysis in an imbalanced dataset, revealing a trade-off between evaluation metrics and real-world sentiment representation. While TF-IDF achieved higher accuracy and recall, it significantly underestimated positive sentiments, whereas BoW maintained a more balanced sentiment distribution despite slightly lower accuracy. Statistical analysis confirmed significant differences in accuracy and recall but found no substantial variation in precision or sentiment distribution. These findings suggest that model selection should consider both performance metrics and real-world applicability, as TF-IDF is preferable for tasks requiring high precision, while BoW is better suited for applications needing balanced sentiment representation.

This study offers important insights into the trade-off between performance and sentiment representation, despite the limitations which include the extremely unbalanced

dataset, the use of simple extraction techniques, and the lack of investigation of other machine learning models in addressing class imbalance. These findings have practical implications for healthcare policy since they can be utilized to improve public services, undertake data-driven policy analysis, and fairly evaluate public perception (not just negative sentiment). Future research should explore advanced feature extraction techniques, alternative classifiers, and applications in different domains to validate these insights.

REFERENCES

[1] L. Abualigah, H. E. Alfar, M. Shehab, and A. M. A. Hussein, "Sentiment Analysis in Healthcare: A Brief Review," in *Studies in Computational Intelligence*, vol. 874, no. December 2019, 2020, pp. 129–141. doi: 10.1007/978-3-030-34614-0_7.

[2] T. D. Dikiyanti, A. M. Rukmi, and M. I. Irawan, "Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm," *J. Phys. Conf. Ser.*, vol. 1821, no. 1, 2021, doi: 10.1088/1742-6596/1821/1/012054.

[3] H. D. Abubakar and M. Umar, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec," *SLU J. Sci. Technol.*, vol. 4, no. 1&2, pp. 27–33, Aug. 2022, doi: 10.56471/slujst.v4i.266.

[4] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bull. Electr. Eng. Informatics*, vol. 10, no. 5, pp. 2780–2788, 2021, doi: 10.11591/eei.v10i5.3157.

[5] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/ACCESS.2022.3149482.

[6] H. R. Sneha and B. Annappa, "Exploratory Analysis of Methods, Techniques, and Metrics to Handle Class Imbalance Problem," *Procedia Comput. Sci.*, vol. 235, pp. 863–877, 2024, doi: 10.1016/j.procs.2024.04.082.

[7] C. Suhaeni and H. S. Yong, "Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences," *Appl. Sci.*, vol. 13, no. 17, 2023, doi: 10.3390/app13179766.

[8] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, "Framework for Improved Sentiment Analysis via Random Minority Oversampling for User Tweet Review Classification," *Electron.*, vol. 11, no. 19, pp. 1–17, 2022, doi: 10.3390/electronics11193058.

[9] J. Qiu, C. Liu, Y. Li, and Z. Lin, "Leveraging sentiment analysis at the aspects level to predict ratings of reviews," *Inf. Sci. (Ny).*, vol. 451–452, pp. 295–309, Jul. 2018, doi: 10.1016/j.ins.2018.04.009.

[10] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci. (Ny).*, vol. 513, pp. 429–441, Mar. 2020, doi: 10.1016/j.ins.2019.11.004.

[11] J. Yun and J.-S. Lee, "Learning from class-imbalanced data using misclassification-focusing generative adversarial networks," *Expert Syst. Appl.*, vol. 240, p. 122288, Apr. 2024, doi: 10.1016/j.eswa.2023.122288.

[12] 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, doi: 10.17485/ijst/v15i17.2339.

[13] T. Mirzoev and S. Kane, "Key strategies to improve systems for managing patient complaints within health facilities–what can we learn from the existing literature?," *Glob. Health Action*, vol. 11, no. 1, 2018, doi: 10.1080/16549716.2018.1458938.

[14] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans. Assoc. Comput. Linguist.*, vol. 12, no. 2018, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675.

[15] Z. Qiu *et al.*, "Assessing the impact of

bag-of-words versus word-to-vector embedding methods and dimension reduction on anomaly detection from log files," *Int. J. Netw. Manag.*, vol. 34, no. 1, pp. 1–20, 2024, doi: 10.1002/nem.2251.

[16] I. Verawati and B. S. Audit, "Algoritma Naïve Bayes Classifier Untuk Analisis Sentiment Pengguna Twitter Terhadap Provider By.u," *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1411, 2022, doi: 10.30865/mib.v6i3.4132.

[17] Y. P. Astuti, A. R. Wibowo, E. Kartikadarma, E. R. Subhiyakto, N. A. Sri Winarsih, and M. S. Rohman, "Penerapan Metode Naïve Bayes Classifier Untuk Klasifikasi Sentimen Pada Judul Berita," *LogicLink*, vol. 1, no. 1, pp. 1–12, 2024, doi: 10.28918/logiclink.v1i1.7684.

[18] Israt Jahan, Md Nakibul Islam, Md Mahadi Hasan, and Md Rafiuddin Siddiky, "Comparative analysis of machine learning algorithms for sentiment classification in social media text," *World J. Adv. Res. Rev.*, vol. 23, no. 3, pp. 2842–2852, 2024, doi: 10.30574/wjarr.2024.23.3.2983.

[19] Y. Jaswanth, R. Muni, S. Kumar, R. M. Sudhan, M. Vijaya Kumar, and M. Rajagopalam, "Sentiment analysis using logistic regression algorithm," *Eur. J. Mol. Clin. Med.*, vol. 7, no. 4, pp. 2081–2086, 2020, [Online]. Available: https://ejmcm.com/article_1947.html

[20] "Sentimen-ID-BPJS." Accessed: Jul. 17, 2025. [Online]. Available: https://www.kaggle.com/datasets/aeworld/sentimen-id-bpjs

[21] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[22] K. Juluru, H. H. Shih, K. N. K. Murthy, and P. Elnajjar, "Bag-of-words technique in natural language processing: A primer for radiologists," *Radiographics*, vol. 41, no. 5, pp. 1420–1426, 2021, doi: 10.1148/rg.2021210025.

[23] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, 2021, doi: 10.1186/s13634-021-00742-6.

[24] S. Dey Sarkar, S. Goswami, A. Agarwal, and J. Aktar, "A Novel Feature Selection Technique for Text Classification Using Naïve Bayes," *Int. Sch. Res. Not.*, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/717092.

[25] T. H. J. Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," *Procedia Comput. Sci.*, vol. 197, no. 2021, pp. 660–667, 2021, doi: 10.1016/j.procs.2021.12.187.

[26] M. T R, V. K. V, D. K. V, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthc. Anal.*, vol. 4, no. July, p. 100247, 2023, doi: 10.1016/j.health.2023.100247.

[27] D. Wilimitis and C. G. Walsh, "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial," *Jmir Ai*, vol. 2, no. 1, 2023, doi: 10.2196/49023.

[28] P. Alkhairi, E. R. Batubara, R. Rosnelly, W. Wanayaumini, and H. S. Tambunan, "Effect of Gradient Descent With Momentum Backpropagation Training Function in Detecting Alphabet Letters," *Sinkron*, vol. 8, no. 1, pp. 574–583, 2023, doi: 10.33395/sinkron.v8i1.12183.

[29] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, pp. 1–22, 2020, doi: 10.1371/journal.pone.0232525.

[30] S. Akuma, T. Lubem, and I. T. Adom, "Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets," *Int. J. Inf. Technol.*, vol. 14, no. 7, pp. 3629–3635, Dec. 2022, doi: 10.1007/s41870-022-01096-4.

[31] C. A. Nurhaliza Agustina, R. Novita, Mustakim, and N. E. Rozanda, "The Implementation of TF-IDF and

Word2Vec on Booster Vaccine Sentiment Analysis Using Support Vector Machine Algorithm," *Procedia Comput. Sci.*, vol. 234, pp. 156–163, 2024, doi: 10.1016/j.procs.2024.02.162.

[32] Dedy Sugiarto, Ema Utami, and Ainul Yaqin, "Perbandingan Kinerja Model TF-IDF dan BOW untuk Klasifikasi Opini Publik Tentang Kebijakan BLT Minyak Goreng," *J. Tek. Ind.*, vol. 12, no. 3, pp. 272–277, 2022, doi: 10.25105/jti.v12i3.15669.

[33] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00943-4.

[34] Z. Nassr, F. Benabbou, N. Sael, and T. Hamim, "Improving Sentiment Analysis Performance on Imbalanced Moroccan Dialect Datasets Using Resample and Feature Extraction Techniques," *Inf.*, vol. 16, no. 1, 2025, doi: 10.3390/info16010039.

[35] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Enhancing machine learning-based sentiment analysis through feature extraction techniques," *PLoS One*, vol. 19, no. 2 February, 2024, doi: 10.1371/journal.pone.0294968.

[36] S. J. Basha, S. R. Madala, K. Vivek, E. S. Kumar, and T. Ammannamma, "A Review on Imbalanced Data Classification Techniques," in *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, IEEE, Mar. 2022, pp. 1–6. doi: 10.1109/ICACTA54488.2022.9753392.

[37] Z. Shuai *et al.*, "Comparison of different feature extraction methods for applicable automated ICD coding," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–15, 2022, doi: 10.1186/s12911-022-01753-5.