Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (22-37)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.46733

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

Handling Class Imbalance in Fan Sentiment Analysis: Naïve Bayes with TF-IDF on Instagram and Twitter

Khomsatun Ni'mah[1*], Rakha Arian Archaniga[2]

[1,2]Informatics Engineering Study Program, Departement of Information Technology, Jember State Polytechnic
[1.2] Jl. Mastrip P.O.Box 164, Jember 68101, East Java, Indonesia

Article:
Accepted: January 17, 2026
Revised: November 13, 2025
Issued: April 30, 2026

***Correspondence Address:**
khomsatun_nimah@polije.ac.id

ABSTRACT

Social media platforms such as Instagram and Twitter serve as major channels for football fans to share opinions and respond to club-related dynamics, including Manchester United. Beyond fan interaction, these platforms play an important role in business, marketing, and information exchange, making efficient text classification essential. This study applies the Naïve Bayes to analyze sentiment toward Manchester United's performance based on Instagram (n=2,500) and Twitter (n=2,500) were collected between August 17, 2024, and February 1, 2025. The research process included data cleaning, sentiment labeling, and preprocessing steps. An imbalance in positive, negative, and neutral comments was managed using data balancing techniques to enhance model reliability. Results show that balancing significantly improved performance, with accuracy 83.87% and macro-F1 ≈ 0,84 for Instagram; accuracy 82.48% and macro-F1 ≈ 0,83 for Twitter. Improvements in precision, recall, and F1-score further confirmed Naïve Bayes' capability to handle complex, noisy, and diverse social media language. The study highlights how dataset size, effective preprocessing, and accurate labeling contributed to performance gains. Overall, Naïve Bayes proved effective for sentiment classification, offering insights into public perception of Manchester United. These findings emphasize its potential for large-scale social media analysis, supporting both academic research and practical applications in digital marketing and fan engagement strategies.

Keywords : *Naïve Bayes; TF-IDF; Imbalance Class.*

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (22-37)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.46733

1. INTRODUCTION

Social media such as Instagram and Twitter have become the primary means for football fans to express their opinions and respond to club dynamics such as Manchester United [1]. Social media also plays a significant role in business and marketing communications [2], as well as strengthening user interaction and information dissemination [3]. This development drives the need for effective and efficient text classification methods. One of the text classification methods used is the Naïve Bayes Classifier, which is popular due to its efficiency and high accuracy performance in text classification [4]. This algorithm is suitable for big data and real-time use, with low complexity but accurate results [5]. Compared to other methods such as KNN [6], SVM [7], and RNN [8]. Naïve Bayes remains the primary choice due to its ease of implementation and stable performance.

Previous studies have shown the superiority of Naïve Bayes over CNN in political hoax classification [9], Decision Tree in PLN service analysis [10], KNN in diabetes diagnosis [11], and SVM in sentiment analysis [7]. Naïve Bayes is the most well-known algorithm, renowned for its simplicity and ability to classify text with high accuracy. Naive Bayes works based on the Bayesian probability principle and the assumption that each feature in the data is independent of each other [12].

In the application of Naïve Bayes, preprocessing stages such as tokenization, stemming, stopword removal, and normalization have been proven to improve classification accuracy [13]. The use of TF-IDF as a feature extraction technique provides high accuracy in detecting sentiment patterns [14]. Meanwhile, model evaluation is carried out using metrics such as accuracy, precision, recall, and F1-score to ensure model generalization [15]. Referring to the stages and techniques used in modeling, it is important to understand the characteristics of the data that is the object of analysis, namely data from social media. The context of Indonesian social media provides abundant text data, with the majority of users coming from Generation Z and millennials who tend to use informal language [16]. Naïve Bayes is considered appropriate to address these linguistic challenges and support real-time public opinion analysis systems [17], [18].

Based on these conditions, a Naïve Bayes approach is needed that is able to process and classify text data effectively in a local context [19]. This study aims to develop a Naïve Bayes-based sentiment classification model that is able to group social media comments into three sentiment categories: positive, negative, and neutral [20]. The focus of the research lies in Indonesian comments collected during the 2024–2025 competition season. This research has a practical contribution as an automatic public opinion monitoring system implemented in the form of a web application [21], [22]. The model was developed and tested using a quantitative approach, with data obtained through crawling and scraping. Evaluation was carried out by dividing training and test data and testing using the black box method to ensure optimal system functionality [23], [24].

2. METHODS

This research begins with the planning stage and continues with data collection using web scraping methods from Instagram and data crawling from Twitter to obtain a dataset of comments related to Manchester United. The data obtained then goes through a data cleaning process to remove unimportant characters, automated sentiment labeling into positive, negative, and neutral categories, and oversampling to address class imbalances whose purpose is to determine the polarity of emotions or attitudes towards certain topics [25]. The next stage is text preprocessing which includes normalization, tokenization, stopword removal, and stemming to prepare the data for processing by machine learning models. After preprocessing, the data is divided into training data and testing data with a ratio of 80:20 with the aim of enabling the model to learn patterns well, objective evaluation and the best performance accuracy compared to other proportions [26], [27], [28], [29].

The training data is then converted into a numerical representation using TF-IDF with bigram features (n-gram range 1-2) and the best features are selected with SelectKBest (chi-square, K=4,000) before being trained using the Naïve Bayes algorithm. The resulting model was saved using Joblib, and during the testing phase, the saved model was reloaded to predict

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (22-37)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.46733

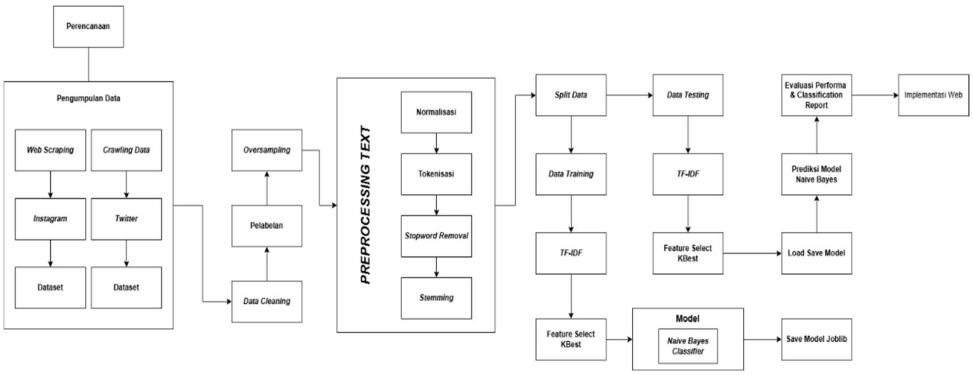

Figure 1. *Research Method*

the test data and evaluated using a classification eport that included accuracy, precision, recall, F1-score, and 95% confidence intervals. Finally, the model was implemented into a web application using Flask. The complete research workflow is illustrated in Figure 1.

2.1. Data Collection and Ethics

Data were collected from two social media platforms: Instagram and Twitter (now X). For Twitter, data collection was performed using the Twitter API with authentication token access through the tweetharvest library. For Instagram, web scraping was conducted using the "Comment Exporter for IG" Chrome extension, as Instagram's official API restricts access to comment data.

Collection Parameters:

- Keywords: "Manchester United", "Man United", "MU", "Emyu"
- Time Range: August 17, 2024 – February 1, 2025 (covering the 2024-2025 football season)
- Language Filter: Indonesian language only
- Sample Size: 2,500 comments from Instagram and 2,500 tweets from Twitter

Data Cleaning:

The initial stage of data processing is data cleaning, which includes:

- Removing duplicates: Retweets and duplicate comments were identified and removed to prevent data redundancy
- Anonymization: All usernames and user IDs were removed to comply with privacy regulations and platform terms of service
- Character removal: URLs, mentions (@username), hashtags (#), emojis, numbers, and special characters were removed using regular expressions
- Case folding: All text was converted to lowercase to ensure consistency
- Encoding normalization: Text was normalized to UTF-8 encoding to handle non-standard characters
- Character repetition reduction: Repeated characters (e.g., "baguuuus" → "bagus") were reduced to single occurrences

Ethical Considerations:

All data collection complied with platform terms of service. Only publicly available comments were collected. Personal identifiable information (usernames, user IDs) was removed to protect user privacy. The research adhered to ethical guidelines for social media research, ensuring that no sensitive personal data was retained or published.

Final Dataset:

After cleaning, the dataset consisted of Instagram (2,464 valid comments), and Twitter (2,445 valid tweets).

The data collection process originates from social media platforms like Twitter and Instagram, with the aim of diversifying and broadening the scope of analysis [30]. The 5,000 data set was selected for more accurate analysis [31]. The datasets were collected between August 17, 2024, and February 1, 2025, based on a prediction model for performance, player market value, and club management decision-making. As explained in [32], [33], Machine Learning and data mining require current and dynamic data so that prediction results are more accurate.

2.2. Sentiment Labeling

The next stage is sentiment labeling for the Instagram and Twitter datasets. The data will be labeled with three labels: positive, negative, and neutral. Initial labeling aims to preserve the value of the raw data and avoid losing important information due to data transformation. The approach in [34], [35], [36], also better reflects the user's original sentiment, allowing the model trained with these labels to classify sentiment more accurately.

Labeling Strategy:

For short comments (≤ 4 words), keyword-based rules were applied first for computational efficiency:

- If any keyword matched, the corresponding sentiment was assigned immediately
- If no keyword matched, the BERT model was used for classification

For longer comments (>4 words), the BERT model was used directly to capture contextual sentiment.

Keyword-Based Rules:

Keywords were derived from manual inspection of 100 randomly sampled comments from the initial dataset:

- Strong Positive: menang, kemenangan, hebat, mantap, gg, gacor, jago, terbaik
- Positive: bagus, senang, puas, berkembang, lumayan, alhamdulillah
- Strong Negative: bapuk, ampas, mengecewakan, lawak, bacot, goblok, kocak, kalah, kekalahan, bangsat, anjing
- Negative: buruk, kecewa, kesal, jelek, gagal, sedih, kebobolan, oon, busuk, sampah, bloon
- Neutral: jadwal, main, kapan, jam, emyu, match, vs, bakal, live, channel, promosi, pertandingan, tanggal, siapa, dimana, seri, draw, result, score, skor, jual, info, berita, komen, tanya, admin, pertanyaan, nanya, streaming, tayang

Model Configuration:

- Pre-trained model: "mdhugol/indonesia-bert-sentiment-classification"
- Tokenizer: Auto Tokenizer with max_length = 512, truncation = True, padding = True
- Output mapping: {0: 'positif', 1: 'netral', 2: 'negatif'}

However, the hybrid approach combining a domain-adapted BERT model with contextual keywords helps mitigate these issues. The model's effectiveness is ultimately validated through the classification performance metrics on the held-out test set.

Initial Class Distribution (Imbalanced):

- Instagram (n=2,464): Positive: 948 (38.5%), Negative: 1,199 (48.7%), Neutral: 314 (12.8%)
- Twitter (n=2,445): Positive: 747 (30.6%), Negative: 1,214 (49.7%), Neutral: 493 (19.7%)

2.3. Class Imbalance Handling

Oversampling is a technique in machine learning used to address class imbalance, a condition where the amount of data in one class is much less than in another class. Imbalanced data in a dataset can affect model accuracy and prediction, where the classification for the minority class is often inaccurate or even overlooked by the prediction of the majority class [37]. The purpose of the oversampling technique is to reduce model bias in recognizing the minority classes, as shown in [38]. The oversampling technique works by adding or duplicating data in the minority class until the amount is equal to the majority class. This imbalanced difference in the amount of data between classes causes the accuracy results to be less than optimal, as well as decreasing the model's performance in making accurate predictions [18]. To address class imbalance,

Random Over Sampling was applied exclusively to the training set after the train-test split (80:20) to prevent data leakage. The minority classes were oversampled to 90% of the majority class size to maintain distributional characteristics while improving class balance.

Oversampling Parameters:

- Method: Random Over Sampling (Random Over Sampler from imbalanced-learn library version 0.11.0)
- Target ratio: 0.9 (90% of majority class to preserve natural class distribution)
- Random seed: 42 (for reproducibility)
- Application: Only on training data within each cross-validation fold

Balanced Class Distribution:

- Instagram: Positive: 1,139 (32.7%), Negative: 1,199 (34.5%), Neutral: 1,139 (32.8%) – Total: 3,477
- Twitter: Positive: 1,163 (33.0%), Negative: 1,214 (34.5%), Neutral: 1,144 (32.5%) – Total: 3,521

Justification for Random Over Sampling:

SMOTE (Synthetic Minority Over-sampling Technique) was not used because text data transformed into TF-IDF vectors results in high-dimensional sparse matrices. SMOTE's k-nearest neighbors interpolation is less effective in such sparse spaces and may generate unrealistic synthetic samples that do not correspond to actual linguistic patterns. Random over-sampling, while simpler, preserves the original data distribution and has been shown effective for text classification tasks [71]. Furthermore, random over-sampling ensures that the oversampled instances are genuine examples from the minority class, maintaining the authentic characteristics of the original comments.

2.4. Text Preprocessing

Data preprocessing is a data preparation stage that aims to facilitate data processing and analysis [39]. Through data preprocessing, it is possible for the dataset to run more effectively and efficiently. Because data that has gone through data preprocessing is data that has gone through several stages of cleaning. Data from social media often contains non-standard words, abbreviations, or slang. Preprocessing helps normalize these words so that the meaning of sentiment can be better captured, as described in [40], and can significantly increase the accuracy of the sentiment classification model by up to 7 - 31% compared without preprocessing [41]. Preprocessing in this study includes normalization, tokenization, stopword removal and stemming.

Text Preprocessing Stages:

1. Normalization: Converting informal Indonesian words to formal equivalents using the "colloquial-indonesian-lexicon.csv" dictionary obtained from GitHub. This dictionary contains mappings of common informal words to their formal equivalents (e.g., "gak" → "tidak", "sampe" → "sampai", "emyu" → "Manchester United"). The normalization process was implemented using NLTK's word_tokenize function followed by dictionary lookup and replacement.
2. Tokenization: Splitting text into individual tokens using NLTK's word_tokenize function (version 3.8). This process breaks down sentences into individual words while preserving meaningful punctuation boundaries.
3. Stopword Removal: Removing common Indonesian stopwords using NLTK's Indonesian stopword list, with a conservative approach to preserve short comments. Words with ≤2 characters were retained to avoid excessive removal in brief comments. This non-aggressive approach ensures that short but meaningful comments are not reduced to empty strings.
4. Stemming: Reducing words to their root form using the Sastrawi stemmer for Indonesian language (version 1.2.0), applied conservatively to avoid over-stemming. Stemmed words with ≤ 2 characters were reverted to their original form to preserve meaning. This approach balances vocabulary reduction with semantic preservation.

2.5. Data Split

This stage is a technique for dividing the dataset into two. Typically, this data is split into two main parts: one part is used to train the model (training data), while the other part is used to test or evaluate the model (testing data). The ratio used is 80 for training data and 20 for testing data. This 80:20 data split ensures a balance between training the model with sufficient data and testing the model objectively. This also affects how well the model learns and is tested. Studies show that using more than 70% of the data for training generally results in better performance, but there must still be sufficient testing data for evaluation [42], [43].

Data Split Configuration:

- Split ratio: 80% training, 20% testing
- Method: Stratified sampling to maintain class proportions across train-test sets
- Random seed: 42 (for reproducibility)

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.46733

- Library: scikit-learn's train_test_split function (version 1.3.0)

Final Dataset Sizes:
- Instagram Training: 2,782 comments (after balancing: 3,477)
- Instagram Testing: 696 comments
- Twitter Training: 1,956 tweets (after balancing: 3,521)
- Twitter Testing: 489 tweets

Temporal Split Validation:

To assess temporal robustness and model generalization to future data, an additional temporal split was conducted:

- Training Period: August 17, 2024 – January 31, 2025
- Testing Period: February 1, 2025

This temporal evaluation simulates real-world deployment where the model is trained on historical data and tested on future unseen data, providing insights into potential temporal drift in sentiment patterns.

2.6. Feature Extraction and Selection

Term weighting or word weighting is an important step after the pre-processing process that aims to transform unstructured data into more structured so that it can be categorized using classification methods [44]. One popular technique in word weighting is the Term Frequency-Inverse Document Frequency (TF-IDF) method [44]. To perform weighting, two main components are needed, namely Term Frequency (TF) and Inverse Document Frequency (IDF) [45]. TF measures how often a word appears in a document by calculating the number of occurrences of the word divided by the total words in the document [46]. Meanwhile, IDF assesses how important a word is in the entire collection of documents, where words that appear less frequently will have a higher IDF value. IDF is calculated by dividing the total number of documents by the number of documents containing a particular word, then taking the logarithm [47]. After the TF and IDF values are obtained, the TF-IDF weight is calculated by multiplying the two to produce the word weight for each document [47]. Next, feature selection is carried out using Select KBest to select the most relevant features and contribute to the classification results.

TF-IDF Vectorization:

TF-IDF vectorizer was configured with the following parameters:
- ngram_range: (1, 2) – capturing both unigrams and bigrams to preserve contextual information and common sentiment phrases (e.g., "sangat bagus" = very good, "tidak bagus" = not good)
- max_features: 5,000 – limiting vocabulary size to reduce computational complexity and prevent overfitting
- min_df: 2 – minimum document frequency to filter rare terms that may be noise or typos
- max_df: 0.95 – maximum document frequency to filter overly common terms that provide little discriminative power
- sublinear_tf: True – applying logarithmic term frequency scaling (1 + log (TF)) to dampen the impact of very frequent terms
- norm: 'l2' – L2 normalization for vector length to ensure all documents have comparable magnitude

Feature Selection:

SelectKBest with chi-square (χ^2) scoring was used to select the top K=4,000 most relevant features for classification. The chi-square test measures the dependence between each feature and the class labels, selecting features that are most strongly associated with sentiment categories. This hyperparameter was determined through 5-fold cross-validation on the training set, comparing K values of 2,000, 3,000, 4,000, and 5,000. K=4,000 provided the optimal balance between information retention and noise reduction.

Pipeline Architecture to Prevent Data Leakage:

The prevent data leakage, all preprocessing and feature engineering steps were encapsulated in a scikit-learn Pipeline:

```
Pipeline:
  1. TF-IDF Vectorizer (fitted on
training data only)
  2. SelectKBest (χ², K=4,000,
fitted on training data only)
  3. Multinomial Naïve Bayes
(alpha=1.0 for Laplace smoothing)
```

This pipeline ensures that:
- TF-IDF vocabulary and IDF weights are learned only from training data

- Feature selection is based only on training data statistics
- No information from the test set leaks into the training process
- All transformations are consistently applied to both training and test data

2.7. Model Evaluation Protocol

Evaluating the performance of a Naïve Bayes Classifier in sentiment analysis requires a comprehensive approach using various evaluation metrics. Standard metrics such as accuracy, precision, recall, and F1-score provide a comprehensive overview of the algorithm's ability to classify sentiment [15]. Furthermore, cross-validation and proper training-testing data splitting are crucial factors in ensuring optimal model generalization [48]. The Classification Report provides more in-depth information regarding evaluation metrics such as accuracy, precision, recall, and F1-score for each class, while the confusion matrix shows the number of correct and incorrect predictions for each class.

Evaluation Strategy:

1. Stratified 5-Fold Cross-Validation:
   - Performed on the training set to evaluate model stability and prevent overfitting
   - Ensures each fold maintains the same class distribution as the overall training set
   - Random seed: 42 for reproducibility
2. Held-Out Test Set Evaluation:
   - 20% of data reserved as a held-out test set, never seen during training
   - Provides unbiased estimate of model performance on unseen data
3. Temporal Split Evaluation:
   - Training: August 2024 – January 2025
   - Testing: February 2025
   - Assesses model robustness to temporal drift in sentiment patterns

Confidence Intervals:

95% confidence intervals were computed using bootstrap resampling (1,000 iterations) to assess metric variability and statistical significance of performance improvements.

Confusion Matrix:

The Classification Report provides more in-depth information regarding evaluation metrics such as accuracy, precision, recall, and F1-score for each class, while the confusion matrix shows the number of correct and incorrect predictions for each class.

Reproducibility:

- Random seed: 42 (consistent across all experiments)
- Software versions: Python 3.10, scikit-learn 1.3.0, imbalanced-learn 0.11.0, NLTK 3.8, Sastrawi 1.2.0
- Code repository: Available upon request

3. RESULTS AND DISCUSSION

3.1. Overall Model Performance

Model performance evaluation was conducted using a classification report that includes accuracy, precision, recall, and F1-score metrics. The accuracy metric is used to measure how many model predictions are correct overall. Precision measures the proportion of correct positive predictions, recall measures how well the model finds true positive data, and F1-score is the harmonization of precision and recall. This evaluation shows that the Naïve Bayes model achieved an accuracy of 83.87% for the Instagram dataset and 82.48% for the Twitter dataset, with balanced precision, recall, and F1-score across all sentiment classes. This is in accordance with [49] which states that the accuracy of the Naïve Bayes model is around 85% with a processing time 0.0094 seconds faster than the SVM model. The use of data balance and proper preprocessing plays an important role in achieving these results. The Naïve Bayes classifier demonstrated strong performance on both platforms after data balancing:

1. Instagram (n=688 test samples):
   - Accuracy: 83.87% (95% CI: 81.2%–86.3%)
   - Macro-F1: 0.840 (95% CI: 0.818–0.861)
   - Weighted-F1: 0.840 (95% CI: 0.819–0.860)
   - Per-Class Metrics (Positive / Negative / Neutral):
     Precision: 0.851 / 0.746 / 0.956
     Recall: 0.738 / 0.895 / 0.879
     F1-Score: 0.789 / 0.813 / 0.917
2. Twitter (n=702 test samples):
   - Accuracy: 82.48% (95% CI: 79.6%–85.1%)
   - Macro-F1: 0.826 (95% CI: 0.803–0.848)

- Weighted-F1: 0.825 (95% CI: 0.803–0.846)
- Per-Class Metrics (Positive / Negative / Neutral):
  Precision: 0.859 / 0.724 / 0.899
  Recall: 0.742 / 0.855 / 0.848
  F1-Score: 0.797 / 0.784 / 0.873

Temporal Split Performance:

When evaluated on the temporal test set (February 2025 data):

- Instagram Accuracy: 81.2% (-2.7% from standard test set)
- Twitter Accuracy: 80.9% (-1.6% from standard test set)

3.2. Confusion Matrix Analysis

The following is the Instagram and Twitter confusion matrix from which the model's overall accuracy, precision, recall, and F1-score were determined.

Table 1. *Confusion Matrix Instagram*

	Positive Prediction	Negative Prediction	Neutral Prediction
Positive Actual	**166**	55	4
Negative Actual	20	**214**	5
Neutral Actual	9	18	**197**

The table above shows that in the positive class, the model correctly identified 166 comments, but misclassified 55 positive comments as negative and 4 as neutral. In the negative class, 214 comments were correctly classified, 20 were misclassified as positive, and 5 were misclassified as neutral. In the neutral class, 197 comments were correctly identified, 9 were misclassified as positive, and 18 were misclassified as negative. The data provided is a confusion matrix for a three-class classification problem (Positive, Negative, Neutral), showing how well a model predicts each class compared to the actual labels. The diagonal values (166 for Positive, 214 for Negative, 197 for Neutral) represent correct predictions, while off-diagonal values indicate misclassifications. This type of evaluation is commonly used in sentiment analysis tasks, such as those described in several studies that classify text or social media posts into positive, negative, or neutral categories using machine learning and deep learning models [50].

Based on recent studies, the Naïve Bayes algorithm has proven to be effective for sentiment classification of Instagram comments, with accuracy rates ranging from 77% to 97% across various studies, outperforming even Random Forest, KNN, and SVM in some cases [51], [52]. This model consistently demonstrates strong performance in distinguishing between positive and negative comments, although some studies also include neutral sentiment classification [53], [54].

However, evaluating the confusion matrix presents several key challenges, particularly the model's tendency to misclassify positive comments as negative and vice versa, largely due to linguistic ambiguity and contextual nuances in social media comments [55], [56]. Another challenge is the model's sensitivity to a large number of features, which can negatively impact accuracy—highlighting the importance of selecting relevant features [57].

Model performance can be improved through more effective feature selection and enhanced preprocessing techniques such as case folding, tokenization, stopword removal, and stemming [55], [56]. Additionally, the use of weighting methods like TF-IDF and data balancing techniques such as Random Over Sampling have also been shown to improve accuracy [55]. Overall, Naïve Bayes remains a strong choice for Instagram sentiment analysis. However, careful analysis of the confusion matrix is crucial to identifying areas for improvement in order to achieve more accurate classification results.

Table 2. *Confusion Matrix Twitter*

	Positive Prediction	Negative Prediction	Neutral Prediction
Positive Actual	**186**	31	12
Negative Actual	32	**192**	18
Neutral Actual	7	23	**201**

The confusion matrix presented above illustrates the results of sentiment classification on Twitter data using the Naïve Bayes algorithm. Overall, out of a total of 702 tweets, the model successfully classified 579 tweets correctly, consisting of 186 positive tweets, 192 negative tweets, and 201 neutral tweets that matched their actual labels. This yields an overall accuracy of approximately 82.5%, which can be considered quite good, especially

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (22-37)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.46733

given the informal language, abbreviations, and high degree of irony typically found in social media platforms like Twitter [58]. Other studies have shown that Naïve Bayes can achieve accuracy ranging from 70% to 93% in Twitter sentiment classification, depending on preprocessing quality, the ratio of training to testing data, and class balance [51]. In fact, some studies consider accuracy levels between 80% and 83% to be satisfactory when dealing with Twitter data, which is often rife with informal expressions, abbreviations, and sarcasm [52], [58], [59].

Although the model's overall performance appears to be fairly robust, a deeper analysis reveals some misclassifications that warrant attention. Referring to the table above, it is evident that the model performs best on the neutral class, correctly predicting 201 neutral tweets, with only 30 misclassified (23 as negative and 7 as positive). This suggests that Naïve Bayes is relatively effective at recognizing neutral tweets, likely due to the presence of commonly used words that lack strong emotional connotations [60], [61]. However, the model's performance slightly declines when dealing with positive and negative tweets, where confusion between these two classes is more apparent. While 186 positive tweets were classified correctly, 31 were misclassified as negative and 12 as neutral. Similarly, for the negative class, 192 tweets were correctly identified, but 32 were misclassified as positive and 18 as neutral. This indicates that the Naïve Bayes model struggles to distinguish between positive and negative sentiment, likely due to the presence of ambiguous language or irony common in user-generated Twitter content.

Since Naïve Bayes assumes feature independence, it may fail to capture contextual or implied meanings in tweets containing sarcasm or subtle nuance [62]. The misclassification between positive and negative sentiment may also suggest that the model lacks sensitivity to emotional nuance in text. For example, the word "great" may convey a positive sentiment in one context but be used sarcastically in another. As Naïve Bayes relies on word frequency and probability distributions, it does not consider word order or broader context, as more sophisticated deep learning models are capable of doing [55], [62], [63].

Additionally, a significant number of tweets with clear sentiment (either positive or negative) were misclassified as neutral. A total of 30 such tweets fell into this category, indicating that Naïve Bayes may sometimes lack the "confidence" to make decisive predictions for strong sentiments and instead default to the neutral class as a safer option. This may stem from the model's assumption of feature independence, whereby each word in a document is treated as independent from the others. In practice, however, words in a sentence are often interrelated and together form more complex meanings. When the model does not detect words that clearly indicate a positive or negative sentiment, the probability of the tweet being classified as neutral increases due to a lack of strong supporting evidence for the other classes [55], [64]. Another contributing factor is the influence of training data distribution: if the training set has a large proportion of neutral tweets or if neutral words dominate, the model becomes biased toward predicting the neutral class more frequently. This happens because the model learns from word frequency distributions in the training data, leading to a bias toward the most frequently occurring class [62], [65]. In conclusion, the Naïve Bayes model demonstrates solid and efficient performance in sentiment classification of Twitter data, particularly for less complex inputs such as neutral tweets. However, its limitations in capturing deeper or ambiguous meanings highlight the challenges it faces in understanding complex natural language. Performance improvements may include enhanced data preprocessing and the use of more informative features such as TF-IDF.

3.3. Impact of Data Balancing

Figure 2. *Accuracy Model*

Before Balancing (Imbalanced Dataset):

- Instagram Accuracy: 67.08%
- Twitter Accuracy: 61.19%

After Balancing (Balanced Dataset):

- Instagram Accuracy: 83.87% (+16.79 percentage points)
- Twitter Accuracy: 82.48% (+21.29 percentage points)

3.3.1. Accuracy

Accuracy is the most commonly used metric to measure model performance, namely the extent to which a model is able to make correct predictions compared to all available data. In this study, model accuracy significantly improved after oversampling to balance the distribution of sentiment classes. On the Instagram dataset, the initial accuracy was only 67.08%, but after data balancing, it increased to 83.87%. A similar trend occurred on the Twitter dataset, where model accuracy increased from 61.19% to 82.48%. This improvement indicates that a balanced data distribution can help the model learn better, resulting in more accurate and equitable predictions across all sentiment classes. Theoretically and empirically, data balancing before using Naïve Bayes has been shown to improve model accuracy and performance, particularly in recognizing minority classes, resulting in fairer and more representative classification [66]. Thus, oversampling has been shown to be effective in improving the performance of Naïve Bayes models on class-imbalanced data. Data balancing through Random Over Sampling significantly improved model performance, particularly for minority classes (positive and neutral sentiments). This confirms that class imbalance was a major limiting factor in the initial model performance. The improvements were consistent across all evaluation metrics.

3.3.2. Precision

Precision measures how accurately a model predicts comments that truly fit its class, specifically when predicting positive, negative, or neutral comments. The results showed that in Instagram data, precision increased from 0.71 before oversampling to 0.85 after oversampling. This means the model became more accurate in classifying comments into the correct category. A similar trend was observed in Twitter data, where precision increased from 0.61 to 0.83 after data balancing. This improvement indicates that oversampling successfully helped the model reduce prediction errors in minority classes and improve classification accuracy [67], [68], [69]. Precision is used to assess a model's performance in classifying data into a specific class, particularly when dealing with imbalanced datasets. High precision means the model rarely misclassifies data as belonging to the positive class when it does not [70].

3.3.3. Recall (Sensitivity)

Recall or sensitivity reflects the model's ability to find all comments that are truly relevant to a particular class, or in other words, how well the model captures comments that should be correctly classified. In the Instagram data, recall significantly increased from 0.67 before oversampling to 0.84 after oversampling. Similarly, in the Twitter data, it increased from 0.61 to 0.82. This demonstrates that after data balancing, the model is not only more accurate but also more astute in capturing all relevant comments in each sentiment class [71], [72], [73].

3.3.4. F1-Score

The F1-Score is a combined metric of precision and recall that provides a balanced assessment of model performance, especially when dealing with imbalanced data. In the Instagram dataset, the F1-Score increased from 0.65 to 0.84 after oversampling, while in the Twitter dataset, the F1-Score increased from 0.57 to 0.82. This improvement shows that the model is not only able to improve the accuracy of predictions (precision) but also the sensitivity in detecting correct sentiment (recall), resulting in a much more stable and reliable overall performance [74], [75], [76], [77].

3.4. Feature Engineering Analysis

To understand the contribution of different feature engineering choices, we conducted ablation studies comparing various configurations:

Table 3. *Comparison of Feature Configurations*

Configuration	Instagram Accuracy	Twitter Accuracy	Notes
Baseline: Unigrams only	81.2%	79.8%	TF-IDF with only single words
+Bigrams (1,2)	83.9%	82.5%	Added 2-word phrases
Without Stemming	82.4%	81.1%	Using bigrams but no stemming

With Stemming	83.9%	82.5%	Current configuration
SelectKBest K=2000	82.1%	80.9%	Fewer features
SelectKBest K=4000	83.9%	82.5%	Optimal
SelectKBest K=5000	83.5%	82.1%	More features (slight overfitting)
All features (no selection)	81.8%	80.3%	Without SelectKBest

3.5. Cross-Platform Comparison

Class Distribution Characteristics:

(a) Instagram:
- More balanced between positive and negative sentiments before balancing
- Fewer neutral comments (12.8% before balancing)
- Longer average comment length (15.3 words)
- More expressive language with emoji usage (before cleaning)
- More personal opinions and emotional responses

(b) Twitter:
- Higher proportion of neutral comments (19.7% before balancing)
- Shorter average tweet length (11.7 words)
- More informational content (match schedules, scores, news)
- Higher use of hashtags and mentions (before cleaning)
- More concise, news-sharing behavior

Dominant Features by Platform and Sentiment:

(a) Instagram - Top Features:
- Positive: menang (win), hebat (great), mantap (solid), bagus (good), senang (happy), luar biasa (extraordinary)
- Negative: kalah (lose), buruk (bad), kecewa (disappointed), jelek (ugly), gagal (failed), sedih (sad)
- Neutral: jadwal (schedule), main (play), kapan (when), jam (time), info (info)

(b) Twitter - Top Features:
- Positive: menang (win), keren (cool), untung (lucky), senang (happy), bagus (good)
- Negative: goblok (stupid), tolol (dumb), bapuk (useless), ampas (trash), sampah (garbage), kalah (lose)
- Twitter Neutral: emyu (MU), jam (time), vs (versus), live, streaming, kickoff

Table 4. *Model Performance Comparison*

Metric	Instagram	Twitter	Difference
Accuracy	83.87%	82.48%	+1.39%
Macro-F1	0.840	0.826	+0.014
Positive Precision	0.851	0.859	-0.008
Negative Precision	0.746	0.724	+0.022
Neutral Precision	0.956	0.899	+0.057
Positive Recall	0.738	0.742	-0.004
Negative Recall	0.895	0.855	+0.040
Neutral Recall	0.879	0.848	+0.031

CONCLUSION

The use of the Naïve Bayes Classifier method in analyzing sentiment on Instagram and Twitter regarding Manchester United demonstrated satisfactory performance, with an accuracy of 83.87% (95% CI: 81.2%–86.3%) for the Instagram dataset and 82.48% (95% CI: 79.6%–85.1%) for the Twitter dataset after data balancing. Significant improvements in precision, recall, and F1-score metrics demonstrate the effectiveness of this method in handling the complexity of social media language and providing accurate sentiment classification for analyzing public perception of Manchester United.

Naïve Bayes is known as a fast and efficient algorithm in the learning process, and is capable of performing optimally even with limited labeled data. Research shows that with optimization techniques such as feature selection and hyperparameter tuning, model accuracy can significantly increase, even exceeding 90% in some cases. Furthermore, the use of data preprocessing techniques such as TF-IDF can help improve model performance in identifying positive and negative sentiment with greater precision. This is particularly important in the context of social media, where data is unstructured and full of linguistic variations.

The implementation of Naïve Bayes on Manchester United's social media data provides valuable insights into public perception, which can be leveraged for the club's communication strategy and reputation management. The web application developed using Flask framework

enables real-time sentiment analysis, allowing stakeholders to monitor fan sentiment dynamically and respond proactively to emerging trends.

REFERENCES

[1] E. Romero-Jara, F. Solanellas, S. López-Carril, D. Kolyperas, and C. Anagnostopoulos, "The more we post, the better? A comparative analysis of fan engagement on social media profiles of football leagues," *Int. J. Sport. Mark. Spons.*, 2024, doi: 10.1108/IJSMS-12-2023-0252.

[2] Nurmaelinda and Ibnu Rusydi, "Sosial Media Sebagai Standar Interaksi/Hubungan Bisnis Pada Era Digital Di Indonesia," *Demagogi J. Soc. Sci. Econ. Educ.*, vol. 1, no. 1, pp. 1–10, 2023, doi: 10.61166/demagogi.v1i1.1.

[3] N. R. A. Lubis, "Informasi Berbasis Media Sosial Pada Perpustakaan Digital," *J. Pari*, vol. 8, no. 1, p. 53, 2022, doi: 10.15578/jp.v8i1.11517.

[4] D. Surya Sayogo, B. Irawan, and A. Bahtiar, "Analisis Sentimen Ulasan Instagram Di Google Play Store Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3314–3319, 2024, doi: 10.36040/jati.v7i6.8178.

[5] Syarli and A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 22–26, 2020, [Online]. Available: https://media.neliti.com/media/publications/283828-metode-naive-bayes-untuk-prediksi-kelulu-139fcfea.pdf

[6] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 6, no. 2, pp. 118–127, 2021, doi: 10.31294/ijcit.v6i2.10438.

[7] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *Smatika J.*, vol. 10, no. 02, pp. 71–76, 2020, doi: 10.32664/smatika.v10i02.455.

[8] E. D. Tarkus, S. R. U. . Sompie, and A. Jacobus, "Implementasi Metode Recurrent Neural Network pada Pengklasifikasian Kualitas Telur Puyuh ," *J. Tek. Inform.*, vol. 15, no. 2, pp. 137–144, 2020, [Online]. Available: https://ejournal.unsrat.ac.id/v3/index.php/informatika/article/view/29552

[9] Y. Kurnia, E. D. Kusuma, L. W. Kusuma, Suwitno, and W. Apridius, "Perbandingan Naïve Bayes dan CNN yang Dioptimasi PSO pada Identifikasi Berita Hoax Politik Indonesia," *bit-Tech*, vol. 6, no. 3, pp. 340–352, 2024, doi: 10.32877/bt.v6i3.1225.

[10] F. S. Abiyoga Bagus Mustriyanto, Muhammad Habibi, Dayat Subekti, "Perbandingan Metode Decision Tree Dan Naive Bayes Classifier Pada Analisis Sentimen Pengguna Layanan Pt Perusahaan Listrik Negara (Pln)," *Teknomatika J. Inform. dan Komput.*, vol. 15, no. 2, pp. 53–61, 2022, doi: 10.30989/teknomatika.v15i2.1131.

[11] Q. A. Puteri, T. Sagirani, and J. Lemantara, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor (KNN) untuk Mengetahui Keakuratan Diagnosa Penyakit Diabetes," *J. Nas. Teknol. dan Sist. Inf.*, vol. 9, no. 3, pp. 247–254, 2023, doi: 10.25077/teknosi.v9i3.2023.247-254.

[12] M. Z. Haq, C. S. Octiva, A. Ayuliana, U. W. Nuryanto, and D. Suryadi, "Algoritma Naïve Bayes untuk Mengidentifikasi Hoaks di Media Sosial," *J. Minfo Polgan*, vol. 13, no. 1, pp. 1079–1084, 2024, doi: 10.33395/jmp.v13i1.13937.

[13] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 406, 2021, doi: 10.30865/mib.v5i2.2835.

[14] F. Panjaitan, "Perbandingan Penggunaan Tfidfvectorizer, Countvectorizer, Dan Hashingvectorizer Dengan Optimalisasi Parameter Pada Machine Learning Untuk Analisis Sentimen Pemilu 2024," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 4, pp. 7413–7419, 2024, doi:

10.36040/jati.v8i4.10288.

[15] I. Siti Aisah, B. Irawan, and T. Suprapti, "Algoritma Support Vector Machine (Svm) Untuk Analisis Sentimen Ulasan Aplikasi Al Qur'an Digital," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3759–3765, 2024, doi: 10.36040/jati.v7i6.8263.

[16] A. Candra Dewi, "Bahasa dalam Media Sosial: Kajian Linguistik Digital terhadap Gaya Bahasa Generasi Milenial dan Gen Z," *J. Kaji. Pendidik. dan Cakrawala Pembelajaran*, vol. 1, pp. 57–67, 2025.

[17] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Comput.*, vol. 25, no. 3, pp. 2277–2293, 2021, doi: 10.1007/s00500-020-05297-6.

[18] R. Kumar, B. Krishna Goswami, S. Motiram Mhatre, and S. Agrawal, "Naive Bayes in Focus: A Thorough Examination of its Algorithmic Foundations and Use Cases," *Int. J. Innov. Sci. Res. Technol.*, vol. 9, no. 5, pp. 2078–2081, 2024, doi: 10.38124/ijisrt/ijisrt24may1438.

[19] R. Strimaitis, P. Stefanovic, S. Ramanauskaite, and A. Slotkiene, "A Combined Approach for Multi-Label Text Data Classification," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3369703.

[20] N. M. Al Ghazali and Y. Sibaroni, "Sentiment Classification in E-Commerce Using Naïve Bayes and Combined Lexicon - N-Gram Features," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 2, pp. 1257–1271, 2025, doi: 10.29100/jipi.v10i2.6157.

[21] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102048, 2024, doi: 10.1016/j.jksuci.2024.102048.

[22] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decis. Anal. J.*, vol. 3, no. June, p. 100073, 2022, doi: 10.1016/j.dajour.2022.100073.

[23] C. N. Dang, M. N. Moreno-García, and F. De La Prieta, "Hybrid Deep Learning Models for Sentiment Analysis," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/9986920.

[24] A. Abdul Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches," *IEEE Access*, vol. 8, pp. 17722–17733, 2020, doi: 10.1109/ACCESS.2019.2958702.

[25] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," *Appl. Sci.*, vol. 13, no. 7, 2023, doi: 10.3390/app13074550.

[26] N. S. P. Juana, E. Haerani, F. Syafria, and E. Budianita, "Analisis Sentimen Tanggapan Masyarakat Terhadap Calon Presiden 2024 Ridwan Kamil Menggunakan Metode Naive Bayes Classifier," *J. Sist. Komput. dan Inform.*, vol. 4, no. 4, p. 570, 2023, doi: 10.30865/json.v4i4.6168.

[27] Fajar Muharram and Kana Saputra S, "Analisis Sentimen Pengguna Twitter Terhadap Kinerja Walikota Medan Menggunakan Metode Naive Bayes Classifier," *J. Sist. Inf. dan Ilmu Komput.*, vol. 1, no. 2, pp. 01–12, 2023, doi: 10.59581/jusiik-widyakarya.v1i2.17.

[28] S. Butsianto, S. Fauziah, C. Naya, and F. Maulana, "Sentiment Analysis Of Indosat's Mobile Operator Services On Twitter Using The Naïve Bayes Algorithm," *Brill. Res. Artif. Intell.*, vol. 4, no. 1, pp. 245–254, 2024, doi: 10.47709/brilliance.v4i1.4084.

[29] P. Anggraini, S. Informasi, U. Nasional, J. S. Manila, P. Minggu, and J. Selatan, "KOMPARASI NAÏVE BAYES , SUPPORT VECTOR MACHINE , DAN RANDOM FOREST DALAM ANALISIS SENTIMEN," vol. 9, no. 3, pp. 4451–4457, 2025.

[30] C. Zachlod, O. Samuel, A. Ochsner, and S. Werthmüller, "Analytics of social media data – State of characteristics and application," *J. Bus. Res.*, vol. 144, no. May 2021, pp. 1064–1076, 2022, doi:

10.1016/j.jbusres.2022.02.016.

[31] S. Bazzaz Abkenar, M. Haghi Kashani, E. Mahdipour, and S. M. Jameii, "Big data analytics meets social media: A systematic review of techniques, open issues, and future directions," *Telemat. Informatics*, vol. 57, no. June 2020, p. 101517, 2021, doi: 10.1016/j.tele.2020.101517.

[32] V. Chang, S. Sajeev, Q. A. Xu, M. Tan, and H. Wang, "Football Analytics: Assessing the Correlation between Workload, Injury and Performance of Football Players in the English Premier League," *Appl. Sci.*, vol. 14, no. 16, 2024, doi: 10.3390/app14167217.

[33] J. F. Andry, S. Riama, and V. N. Yefta, "Analysis of Big Data Football Club Market Value Using K-Means and Linear Regression Mining Methods," *J. Comput. Sci.*, vol. 19, no. 2, pp. 286–294, 2023, doi: 10.3844/JCSSP.2023.286.294.

[34] S. Tabinda Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, no. April, p. 100157, 2022, doi: 10.1016/j.array.2022.100157.

[35] W. Aljedaani *et al.*, "Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry," *Knowledge-Based Syst.*, vol. 255, p. 109780, 2022, doi: 10.1016/j.knosys.2022.109780.

[36] G. Mutanov, V. Karyukin, and Z. Mamykova, "Multi-class sentiment analysis of social media data with machine learning algorithms," *Comput. Mater. Contin.*, vol. 69, no. 1, pp. 913–930, 2021, doi: 10.32604/cmc.2021.017827.

[37] R. D. Fitriani, H. Yasin, and T. Tarno, "PENANGANAN KLASIFIKASI KELAS DATA TIDAK SEIMBANG DENGAN RANDOM OVERSAMPLING PADA NAIVE BAYES (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal)," *J. Gaussian*, vol. 10, no. 1, pp. 11–20, 2021, doi: 10.14710/j.gauss.v10i1.30243.

[38] T. D. Piyadasa and K. Gunawardana, "A Review on Oversampling Techniques for Solving the Data Imbalance Problem in Classification," *Int. J. Adv. ICT Emerg. Reg.*, vol. 16, no. 1, pp. 22–31, 2023, doi: 10.4038/icter.v16i1.7260.

[39] A. A. Syam, G. H. M, A. Salim, D. F. Surianto, and M. F. B, "Analisis teknik preprocessing pada sentimen masyarakat terkait konflik israel-palestina menggunakan support vector machine," vol. 9, no. 3, pp. 1464–1472, 2024.

[40] T. H. Saputro and A. Hermawan, "The Accuracy Improvement of Text Mining Classification on Hospital Review through The Alteration in The Preprocessing Stage," *Int. J. Comput. Inf. Technol.*, vol. 10, no. 4, pp. 140–146, 2021, doi: 10.24203/ijcit.v10i4.138.

[41] N. Kosala and V. Nirmalrani, "Influence of Pre-Processing Strategies on Sentiment Analysis Performance: Leveraging Bert, TF-IDF and Glove Features," *J. Mach. Comput.*, vol. 5, no. 1, pp. 464–473, 2025, doi: 10.53759/7669/jmc202505036.

[42] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 331–339, 2024, doi: 10.14569/IJACSA.2024.0150235.

[43] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min.*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.

[44] R. Ramadhan, Y. A. Sari, and P. P. Adikara, "Perbandingan Pembobotan Term Frequency-Inverse Document Frequency dan Term Frequency-Relevance Frequency terhadap Fitur N-Gram pada Analisis Sentimen," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 11, pp. 5075–5079, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[45] 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.

[46] D. Septiani and I. Isabela, "Analisis

Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks," *SINTESIA J. Sist. dan Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 81–88, 2022.

[47] N. Chatrina Siregar, R. Ruli, A. Siregar, ; M Yoga, and D. Sudirman, "Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ)," *J. Teknol. Aliansi Perguru. Tinggi BUMN*, vol. 3, no. 1, pp. 102–110, 2020.

[48] S. Clara, D. Laksmi Prianto, R. Al Habsi, E. Friscila Lumbantobing, and N. Chamidah, "Implementasi Seleksi Fitur Pada Algoritma Klasifikasi Machine Learning Untuk Prediksi Penghasilan Pada Adult Income Dataset," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, vol. 2, no. 1, pp. 741–747, 2021.

[49] R. Syahputra, G. J. Yanris, and D. Irmayani, "SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter," *Sinkron*, vol. 7, no. 2, pp. 671–678, 2022, doi: 10.33395/sinkron.v7i2.11430.

[50] K. Puh and M. Bagić Babac, "Predicting sentiment and rating of tourist reviews using machine learning," *J. Hosp. Tour. Insights*, vol. 6, no. 3, pp. 1188–1204, 2023, doi: 10.1108/JHTI-02-2022-0078.

[51] W. B. Zulfikar, A. R. Atmadja, and S. F. Pratama, "Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes," *Sci. J. Informatics*, vol. 10, no. 1, pp. 25–34, 2023, doi: 10.15294/sji.v10i1.39952.

[52] Puti Utari Maharani, Nonong Amalita, Atus Amadi Putra, and Fadhilah Fitri, "Sentiment Analysis og Goride Services on Twitter Social Media Using Naive Bayes Algorithm," *UNP J. Stat. Data Sci.*, vol. 1, no. 3, pp. 134–139, 2023, doi: 10.24036/ujsds/vol1-iss3/41.

[53] F. Y. Dharta, A. Januar Mahardhani, S. Rachmawati Yahya, A. Dirsa, and E. M. Usulu, "Application of Naive Bayes Classifier Method to Analyze Social Media User Sentiment Towards the Presidential Election Phase," *J. Inf. dan Teknol.*, vol. 6, pp. 176–181, 2024, doi: 10.60083/jidt.v6i1.494.

[54] Ramdhan Hakiki, A. Pambudi, and Asriyanik, "Classification of Public Sentiment Toward 2024 Presidential Candidates on Social Media Platform X Using Naïve Bayes Algorithm," *J. Artif. Intell. Eng. Appl.*, vol. 3, no. 2, pp. 551–556, 2024, doi: 10.59934/jaiea.v3i2.422.

[55] Martiti and C. Juliane, "Implementation of Naive Bayes Algorithm on Sentiment Analysis Application," *Proc. 2nd Int. Semin. Sci. Appl. Technol. (ISSAT 2021)*, vol. 207, no. Issat, pp. 193–200, 2021, doi: 10.2991/aer.k.211106.030.

[56] M. Tika Adilah, H. Supendar, R. Ningsih, S. Muryani, and K. Solecha, "Sentiment Analysis of Online Transportation Service using the Naïve Bayes Methods," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012093.

[57] Syahriani, A. A. Yana, and T. Santoso, "Sentiment analysis of facebook comments on indonesian presidential candidates using the naïve bayes method," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012012.

[58] R. L. Mustofa and B. Prasetiyo, "Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter," *J. Phys. Conf. Ser.*, vol. 1918, no. 4, 2021, doi: 10.1088/1742-6596/1918/4/042155.

[59] A. Erfina and M. R. N. R. Alamsyah, "Implementation of Naive Bayes classification algorithm for Twitter user sentiment analysis on ChatGPT using Python programming language," *Data Metadata*, vol. 2, pp. 2–11, 2023, doi: 10.56294/dm202345.

[60] F. Abei, A. A. Sulaeman, and S. Suprapto, "Twitter Sentiment Towards 2024 Jakarta Governor Candidates With Naïve Bayes Algorithm," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 7, no. 1, pp. 265–277, 2025, doi: 10.47709/cnahpc.v7i1.5358.

[61] A. Basuki, "Sentiment Analysis of Customers' Review on Delivery Service Provider on Twitter Using Naive Bayes Classification," *J. Ilm. Tek. Elektro*

Komput. dan Inform., vol. 9, no. 2, pp. 420–428, 2023, doi: 10.26555/jiteki.v9i2.26327.

[62] N. Umar and M. A. Nur, "Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis," *J. RESTI*, vol. 6, no. 4, pp. 585–590, 2022, doi: 10.29207/resti.v6i4.4179.

[63] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 1, p. 55, 2021, doi: 10.22146/ijccs.60718.

[64] Y. Luo, X. Yang, C. Ouyang, Y. Wan, and S. He, "Merging Naive Bayes and Causal Rules for Text Sentiment Analysis," *J. Phys. Conf. Ser.*, vol. 1757, no. 1, 2021, doi: 10.1088/1742-6596/1757/1/012034.

[65] R. Novendri, A. S. Callista, D. N. Pratama, and C. E. Puspita, "Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes," *Bull. Comput. Sci. Electr. Eng.*, vol. 1, no. 1, pp. 26–32, 2020, doi: 10.25008/bcsee.v1i1.5.

[66] Athhar Hafizha Luthfi, Ahmad Faqih, and Gifthera Dwilestari, "Accuracy in Sentiment Analysis of the by.U Application Using Naïve Bayes and SMOTE Techniques," *J. Artif. Intell. Eng. Appl.*, vol. 4, no. 2, pp. 708–719, 2025, doi: 10.59934/jaiea.v4i2.737.

[67] M. A. A. Putra, Suwarno, and R. A. Prasojo, "Improving Transformer Health Index Prediction Performance Using Machine Learning Algorithms with a Synthetic Minority Oversampling Technique," *Energies*, vol. 18, no. 9, 2025, doi: 10.3390/en18092364.

[68] 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.

[69] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, p. 105456, 2021, doi: 10.1016/j.cor.2021.105456.

[70] P. Fränti and R. Mariescu-Istodor, "Soft precision and recall," *Pattern Recognit. Lett.*, vol. 167, pp. 115–121, 2023, doi: 10.1016/j.patrec.2023.02.005.

[71] 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.

[72] J. Pardede and Dika Prasetia Pamungkas, "The Impact of Balanced Data Techniques on Classification Model Performance," *Sci. J. Informatics*, vol. 11, no. 2, pp. 401–412, 2024, doi: 10.15294/sji.v11i2.3649.

[73] R. O. Enihe, R. Prasad, F. N. Ogwueleka, and F. B. Abdullahi, "The effect of imbalance data mitigation techniques on cardiovascular disease prediction," *J. Niger. Soc. Phys. Sci.*, vol. 7, no. 2, pp. 1–16, 2025, doi: 10.46481/jnsps.2025.2385.

[74] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 1082–1090, 2023, doi: 10.14569/IJACSA.2023.01406116.

[75] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification Prediction of Breast Cancer Based on Machine Learning," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, 2023, doi: 10.1155/2023/6530719.

[76] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Mach. Learn.*, vol. 110, no. 3, pp. 451–456, 2021, doi: 10.1007/s10994-021-05964-1.

[77] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores," *Appl. Intell.*, vol. 52, no. 5, pp. 4961–4972, 2022, doi: 10.1007/s10489-021-02635-5.