Science in Information Technology Letters
Vol. 3., No. 2, November 2022, pp. 59-72
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

Text classification of traditional and national songs using naïve bayes algorithm

Triyanti Simbolon [a,1], Aji Prasetya Wibawa [a,2,*], Ilham Ari Elbaith Zaeni [a,3], Amelia Ritahani Ismail [b,4]

[a] Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia
[b] Department of Computer Science, International Islamic University Malaysia, Malaysia
[1] tryntsmbln@gmail.com; [2] aji.prasetya.ft@um.ac.id; [3] ilham.ari.f t@um.ac.id; [4] amelia@iium.edu.my

* Corresponding Author

ARTICLE INFO

Article history
Received 26 September, 2022
Revised 20 October, 2022
Accepted 01 November, 2022

Keywords
Traditional songs
National songs
Multinomial naïve bayes
SMOTE
Text classification

ABSTRACT

In this research, we investigate the effectiveness of the multinomial Naïve Bayes algorithm in the context of text classification, with a particular focus on distinguishing between folk songs and national songs. The rationale for choosing the Naïve Bayes method lies in its unique ability to evaluate word frequencies not only within individual documents but across the entire dataset, leading to significant improvements in accuracy and stability. Our dataset includes 480 folk songs and 90 national songs, categorized into six distinct scenarios, encompassing two, four, and 31 labels, with and without the application of Synthetic Minority Over-sampling Technique (SMOTE). The research journey involves several essential stages, beginning with pre-processing tasks such as case folding, punctuation removal, tokenization, and TF-IDF transformation. Subsequently, the text classification is executed using the multinomial Naïve Bayes algorithm, followed by rigorous testing through k-fold cross-validation and SMOTE resampling techniques. Notably, our findings reveal that the most favorable scenario unfolds when SMOTE is applied to two labels, resulting in a remarkable accuracy rate of 93.75%. These findings underscore the prowess of the multinomial Naïve Bayes algorithm in effectively classifying small data label categories.

1. Introduction

Indonesia, a nation characterized by its intricate tapestry of ethnicities and cultural diversity, is a testament to the harmonious coexistence of various traditions [1]. Within this cultural mosaic, folk songs occupy a pivotal role, representing the cultural essence of Indonesia's diverse regions [2]. These songs encapsulate the linguistic and thematic nuances that are unique to each geographical locale, serving as sonic time capsules preserving the multifaceted cultural heritage of the archipelago [3], [4]. In concert with folk songs, Indonesia proudly features national songs, often synonymous with national anthems, which bear the solemn responsibility of vocalizing the nation's collective identity, patriotism, and unity.

Notwithstanding their historical significance and cultural importance, these traditional and national songs have encountered a troubling decline in their visibility and cultural resonance in contemporary Indonesian society [5], [6]. Their displacement from the cultural mainstream has inevitably culminated in a regrettable waning of public knowledge and interest [7]. It is within this cultural context that our research emerges as a clarion call to resuscitate these fading cultural jewels, employing advanced technology as a catalyst for reviving the traditions of the past and reigniting their significance in the present.

To facilitate the renewed appreciation and accessibility of these songs, a robust methodology is indispensable. Text classification, a method that leverages natural language processing and machine learning techniques [8] to categorize objects based on their intrinsic characteristics, emerges as the keystone of our endeavor [9]. The realm of text classification offers an array of approaches, each bearing unique merits and drawbacks. Among these, the Naïve Bayes algorithm stands out, and we have opted for the Multinomial Naïve Bayes (MNB) classification method due to its versatility and proven track record of effectiveness [10], [11]. MNB, a derivative of the Naïve Bayes classification family, operates on the foundation of probability theory [12], wielding the capacity to categorize documents based on the occurrence of words within single documents and their relative frequency across the entire corpus [13]–[15]. Its selection is underpinned by the pursuit of enhanced accuracy and efficiency, attributes deemed vital for our pursuit.

However, the application of MNB classification introduces a formidable challenge—imbalanced data class distribution [16], [17]. In the context of our research, this manifests as an unequal representation of folk songs and national songs within the dataset. To address this challenge, two principal approaches may be employed: the data approach and the algorithm approach [18]. In our research, we have chosen to adopt the data approach, and, more specifically, the Synthetic Minority Over-Sampling Technique (SMOTE) as our solution.

SMOTE represents an innovative methodology that rectifies imbalanced datasets by generating synthetic data instances [19], thereby equalizing the representation of minority classes and ensuring a more equitable classification outcome [20]. The application of SMOTE holds the promise of rectifying data class imbalances and fortifying the accuracy and comprehensiveness of our research. The ensuing sections will meticulously explore the application of the MNB method in conjunction with the SMOTE technique for classifying the lyrical content of folk songs and national songs. The overarching aim is to streamline the categorization process, thereby enhancing efficiency and fostering a renewed cultural understanding and appreciation of Indonesia's rich musical and lyrical heritage. This research seeks to bridge the past and present, serving as a testament to the enduring importance of Indonesia's cultural diversity and linguistic richness, as encapsulated within the verses of its folk and national songs.

2. Method

This study comprises a meticulously structured methodology, as illustrated in Fig. 1, which involves several critical stages, each contributing to the comprehensive text classification process.

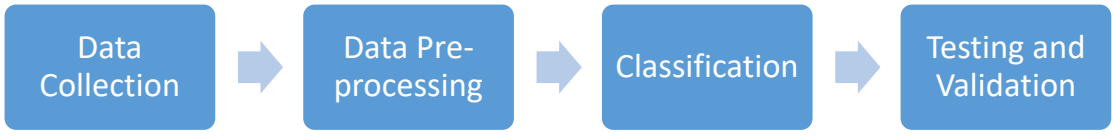

Fig. 1. Training Set

2.1. Data Collection

The initial phase of this research encompassed the meticulous collection of a diverse dataset comprising the lyrics of Indonesian folk songs and national songs [21]. The dataset's composition was informed by multiple sources, including folk songbooks, national songbooks, and online articles. Through this rigorous data collection process, a comprehensive corpus of 480 regional songs representing all 30 Indonesian provinces and 90 national songs was meticulously assembled.

This all-encompassing dataset forms the foundational resource upon which the entire research is conducted. It embodies a rich and varied collection of lyrical content from across the Indonesian archipelago, enabling in-depth analysis and classification of folk and national songs [22]. The dataset curation process reflects an extensive and methodical approach to ensure diversity and inclusivity, as it aims to encapsulate the linguistic and thematic nuances that emanate from the country's numerous provinces. The content extraction from various sources and the meticulous assembly of lyrics from both folk and national songs underscore the depth and comprehensiveness of this dataset, which serves as the cornerstone for the subsequent stages of the research.

2.2. Data Pre-Processing

The data pre-processing stage is a critical facet of this research, encompassing a series of meticulous procedures to prepare the textual content of the dataset for the subsequent analysis [23]. This stage is integral to ensuring the dataset's optimal quality and relevance in the text classification process [24], [25]. Below, we provide a detailed and in-depth analysis of the data pre-processing steps undertaken.

2.2.1. Removal of Punctuation

The removal of punctuation marks from the dataset is a foundational step in data pre-processing [26]. This process involves the elimination of non-alphabetic characters, such as question marks, commas, ellipsis marks, and more. Its purpose is to reduce the presence of noisy, non-alphabetic characters that could potentially interfere with the efficiency and accuracy of the subsequent classification process. Table 1 displays the results of this punctuation removal process.

Table 1. Results of Remove Punctuation

INPUT	OUTPUT
Bungong jeumpa, Bungong jeumpa, Meugah di Aceh... Bungong teuleubeh, teuleubeh, Indah lagoina...	Bungong jeumpa Bungong jeumpa Meugah di Aceh Bungong teuleubeh teuleubeh Indah lagoina

The removal of punctuation marks from the lyrical content serves to enhance the text's clarity and ensures that only the essential linguistic elements are retained for analysis. This step significantly contributes to reducing noise in the dataset.

2.2.2. Case Folding

The case folding process involves converting all capital letters within the song lyrics to lowercase. This step is integral to reducing data redundancy by standardizing the text's case format [27]. The uniformity achieved through case folding enhances the efficiency of the classification process, as it ensures that the same word in different cases is treated as a single entity. Table 2 illustrates the results of this case folding process.

Table 2. Results of Case Folding

Input	Output
Bungong jeumpa Bungong jeumpa Meugah di Aceh Bungong teuleubeh teuleubeh Indah lagoina	bungong jeumpa bungong jeumpa meugah di aceh bungong teuleubeh teuleubeh indah lagoina

Case folding plays a pivotal role in streamlining the dataset for effective text classification. By rendering all text in lowercase, the research minimizes variations in word case, resulting in a more efficient and accurate classification process.

2.2.3. Tokenization

Tokenization, a crucial step in data pre-processing, involves breaking down the text into individual words by identifying word boundaries through punctuation or spaces [28]. This process results in the isolation of individual words, creating a series of single-word tokens that form the basis for subsequent analysis. The tokenization process is demonstrated in Table 3.

Table 3. Results of Toknenization

Input	Output
bungong jeumpa bungong jeumpa meugah di aceh bungong teuleubeh teuleubeh indah lagoina	['bungong','jeumpa','bungong', 'jeumpa','meugah','di','aceh', 'bungong','teuleubeh','teuleubeh', 'indah','lagoina']

Tokenization serves to dissect the lyrical content into its constituent components, allowing for the analysis of individual words and phrases. These tokens become the input for various text classification processes, enabling in-depth analysis of the song lyrics.

The data pre-processing stage is fundamental to the research's success, as it ensures that the dataset is primed and optimized for the subsequent classification processes. By systematically removing punctuation, applying case folding, and tokenizing the text, this stage enhances the dataset's quality and cohesiveness, paving the way for accurate and meaningful text classification.

2.3. Word Weighting with Term Frequency (TF)

The word weighting stage, utilizing Term Frequency (TF) analysis, is a fundamental component of text classification, and it plays a pivotal role in determining the importance and relevance of words within the lyrical content. This section delves into a more detailed and in-depth analysis of this stage.

2.3.1. Term Frequency (TF) Measurement

Term Frequency (TF) represents a quantitative measure of the frequency with which a word or phrase appears in a document [29]. In the context of this research, TF is employed to ascertain the significance of individual words within the song lyrics. A high TF value denotes the prevalence of a word in a document, indicating its importance in the classification process.

TF is determined by calculating the frequency of each word within a given document, where a higher frequency suggests the word's significance in the research process. Furthermore, the TF value can also be leveraged to identify the class location of the same word in multiple classes by considering the TF values across different classes.

2.3.2. Significance of TF Value

The TF value carries substantial importance in text classification. By examining the TF values of words within a document, it is possible to identify those words that have the most substantial impact on class identification. Words with high TF values are likely to have a significant effect on the classification process, influencing the assignment of documents to their respective classes.

2.3.3. TF Application

In this research, the application of TF involves the measurement of word frequency within the dataset [30]. It helps determine the importance of specific words in the research process and their relevance to the classification of folk songs and national songs. The utilization of TF is particularly valuable in understanding the impact of individual words on the classification outcome. It aids in identifying words that are highly relevant to specific classes, thereby enhancing the accuracy of the text classification process.

2.3.4. In – Depth Analysis

The Word Weighting with Term Frequency (TF) stage is crucial in identifying the weight or significance of individual words [31] within the lyrical content of folk and national songs. By quantifying the frequency of each word, this process enables the research to pinpoint words that play a substantial role in the classification of songs. The high TF values of certain words indicate their importance in distinguishing between folk songs and national songs. Words that have a high TF value within specific categories are more likely to be influential in the classification process. This approach contributes to the robustness and precision of the text classification, ensuring that the most relevant linguistic elements are considered in determining the category of each document. The application of TF is a cornerstone of this research, as it provides a quantitative foundation for word importance, thereby bolstering the classification process. It assists in identifying and highlighting the linguistic elements that bear the most significant influence on categorizing the lyrical content of songs, ultimately contributing to the research's comprehensive and accurate results.

2.4. Text Classification using Multinomial Naïve Bayes (MNB)

The text classification process, hinging on the Multinomial Naïve Bayes (MNB) method [32], is the core of this research. Here, we delve into a detailed and comprehensive analysis of the MNB text classification, its principles, and its role in categorizing folk and national songs.

2.4.1. MNB: A Probability-Based Classification Method

Multinomial Naïve Bayes (MNB) is a classification method grounded in probability theory. Its fundamental principle is the calculation of document category based on the occurrence of words both within an individual document and across the entire dataset [33]. This probabilistic approach distinguishes MNB from other classification methods, making it a powerful tool for text categorization. MNB operates by determining the likelihood of a document belonging to a specific class, given the words that appear in the document. It combines prior probabilities (probability of a document's class of origin) with posterior probabilities (probability of a word being associated with the class of origin).

2.4.2. Preceding Steps in MNB

The MNB process commences with the assignment of data labels, extracting values from these labels, and calculating key statistics [34]. These include the number of documents, the number of classes, and the frequency of words across the entire dataset.

- **Prior Probability Calculation:** Prior probability is calculated to establish the likelihood of a document belonging to a specific class. It sets the stage for the subsequent classification process, providing a foundation for identifying the document's origin.
- **Posterior Probability Calculation:** Posterior probability serves to calculate the probability of a word or term being associated with the class of origin. This step is integral in determining the contribution of individual words to the classification outcome.

2.4.3. Advantages of MNB

Multinomial Naïve Bayes offers several advantages that render it a powerful choice for text classification [35] in this research:

- **Efficiency:** MNB's efficiency stems from its probabilistic approach, which streamlines the classification process based on word occurrences. This approach is notably efficient for large datasets.
- **Dominance:** MNB demonstrates dominance over other Naïve Bayes models, making it a robust choice for text classification in this research.
- **Probability-Based:** MNB leverages the theory of chance and probability, which aligns well with the nature of text classification where word occurrence probabilities are a crucial factor.

2.4.4. Sensitivity to Feature Selection

While MNB boasts many strengths, it is essential to acknowledge its sensitivity to feature selection. Specifically, when dealing with a considerable number of features in the classification process, MNB may encounter challenges [36]. A large feature set can lead to increased computation time, potentially affecting the overall accuracy of the results. Careful feature selection and dimensionality reduction strategies are pivotal to mitigating this challenge.

The text classification process using Multinomial Naïve Bayes (MNB) is the crux of this research. By calculating probabilities based on word occurrences, MNB efficiently determines the class of origin for each document in the dataset. MNB's probabilistic approach sets it apart, enabling it to handle text classification effectively. The combination of prior and posterior probabilities ensures that the classification process is well-informed and guided by the likelihood of word occurrences. MNB's efficiency and dominance make it an ideal choice for this research, where accuracy and efficiency are paramount.

However, it is vital to recognize that MNB's sensitivity to feature selection requires careful consideration when dealing with extensive feature sets. The MNB method stands as a robust and efficient tool for the classification of folk songs and national songs in this research. Its probabilistic underpinnings, combined with careful feature selection, ensure accurate and meaningful categorization of the lyrical content, further contributing to the research's overall success.

2.5. Dataset Test and SMOTE

The Dataset Test and Synthetic Minority Over-Sampling Technique (SMOTE) stage represents a pivotal phase in the research process, focusing on the validation and augmentation of the dataset to ensure robust and accurate text classification [37]. This section provides a detailed and comprehensive analysis of this crucial stage.

2.5.1. Six Test Scenarios

This research conducts dataset tests in six distinct scenarios, each designed to evaluate and validate the text classification process comprehensively. These scenarios are defined by variations in the number of labels and the application of SMOTE, a technique aimed at mitigating class imbalances :

- **Scenario 1 - No SMOTE (480 labels on folk songs, 90 labels on national songs):** This scenario operates without SMOTE and involves a sizeable number of labels for folk songs and national songs. It serves as a benchmark for assessing the baseline performance of the classification process.
- **Scenario 2 - No SMOTE (Varying k values):** Similar to Scenario 1, this scenario does not employ SMOTE. However, it introduces varying values of k, determining the number of nearest neighbors used in the SMOTE process. It tests the influence of k on the classification process, ranging from 9 to 89.
- **Scenario 3 - No SMOTE (Balanced dataset):** In this scenario, SMOTE is not applied. The dataset is balanced, with an equal number of labels for folk songs and national songs. This configuration assesses the classification process in a balanced context.
- **Scenario 4 - SMOTE (480 labels on folk songs, 90 labels on national songs):** This scenario introduces SMOTE to address class imbalances. Despite the same label distribution as Scenario 1, the addition of SMOTE aims to enhance classification accuracy and mitigate class imbalance challenges.
- **Scenario 5 - SMOTE (Varying k values):** SMOTE is applied, and k values vary as in Scenario 2. This scenario assesses the combined impact of SMOTE and k values on the classification process.
- **Scenario 6 - SMOTE (Balanced dataset):** Similar to Scenario 3, this scenario employs SMOTE to augment the dataset. It evaluates the effect of SMOTE on a balanced dataset.

2.5.2. Synthetic Minority Over-Sampling Technique (SMOTE):

SMOTE is a data augmentation technique that addresses class imbalances by generating synthetic samples for the minority class [38], [39]. It works by identifying the nearest neighbors of instances in the minority class and creating new synthetic instances between them. This technique has proven effective in handling class imbalances and reducing overfitting.

- **K-Value in SMOTE:** The choice of the k-value, representing the number of nearest neighbors considered in the SMOTE process, plays a crucial role in determining the effectiveness of the technique. Varying k-values are explored to understand their impact on classification performance.

2.5.3. Addressing Class Imbalances

The presence of class imbalances can skew the classification process, potentially leading to misclassifications and reduced accuracy. SMOTE offers an effective solution by augmenting the minority class, thereby equalizing class proportions and improving the reliability of the classification process.

2.5.4. Evaluating Classification Performance

Each of the six scenarios is designed to assess the performance of the text classification process under different conditions. This evaluation is crucial for understanding the impact of SMOTE, k-values, and class balance on the classification results. Accuracy, precision, recall, and F1-score are likely used metrics to evaluate the classification performance.

2.5.5. Significance of SMOTE

The application of SMOTE is of significant importance in this research. It addresses class imbalances, ensuring that the classification process is not unduly influenced by the majority class. SMOTE enhances the dataset's diversity by introducing synthetic samples, thereby contributing to more accurate and reliable text classification. The Dataset Test and SMOTE stage represents a meticulous evaluation of the text classification process. The six distinct scenarios, each with specific configurations, serve to comprehensively assess the performance of the Multinomial Naïve Bayes algorithm under different conditions. By addressing class imbalances and employing SMOTE, this research seeks to ensure that the classification process is robust and reliable, delivering accurate results. The significance of SMOTE lies in its ability to mitigate class imbalances, a common challenge in text classification.

2.6. K-Fold Cross-Validation

The K-Fold Cross-Validation stage is the final and crucial step in the research process. It serves as a method of rigorously testing and validating the performance of the Multinomial Naïve Bayes (MNB) text classification under various conditions [40]. This section provides a detailed and comprehensive analysis of the K-Fold Cross-Validation process.

2.6.1. Validation through K-Fold Cross-Validation

K-Fold Cross-Validation is employed to assess the generalization and robustness of the MNB classification model. It systematically divides the dataset into k equally-sized subsets (folds), where k is set at 10 for this research [41]. The process operates through a sequence of iterations, ensuring that each fold serves as both a training and a testing set.

2.6.2. Key Aspects of K-Fold Cross-Validation

- **Dividing the Dataset:** The dataset is divided into k subsets, ensuring that each fold is statistically representative of the overall dataset. This division guarantees comprehensive testing and validation.
- **Training and Testing:** In each iteration, one fold is designated as the testing set, while the remaining k-1 folds are used for training the MNB classification model. This approach ensures that every data point is tested and validated.
- **Multiple Iterations:** The K-Fold Cross-Validation process is conducted through a series of iterations, with each fold taking turns as the testing set. This iteration reinforces the model's reliability and ability to generalize.

2.6.3. Assessing Classification Performance

The performance of the MNB classification model is assessed using various evaluation metrics in each iteration. Metrics such as accuracy, precision, recall, and F1-score are pivotal in gauging the model's performance in accurately categorizing folk songs and national songs.

2.6.4. Significance of K-Fold Cross-Validation

K-Fold Cross-Validation holds significant importance in this research for several reasons :

- **Robustness Testing:** By systematically testing the classification model under various conditions, K-Fold Cross-Validation evaluates its robustness and ability to deliver consistent results.
- **Generalization Assessment:** This method assesses how well the MNB classification model generalizes to new, unseen data, a critical consideration in text classification.

- **Model Evaluation:** The evaluation metrics used in K-Fold Cross-Validation provide insights into the model's performance, accuracy, and reliability.

K-Fold Cross-Validation is the ultimate validation step in the research process, ensuring that the MNB classification model performs consistently and reliably. By testing the model through multiple iterations and using rigorous evaluation metrics, this stage seeks to validate the effectiveness of the text classification process. The significance of K-Fold Cross-Validation lies in its ability to test the model's performance under various conditions, ensuring that it can generalize and deliver accurate results. The iterative nature of this process adds depth to the research's validation, bolstering the confidence in the accuracy and robustness of the classification model.

3. Results and Discussion

In this section, we delve into a deeper analysis of the research findings, focusing on the scenarios, and the impact of the Synthetic Minority Over-Sampling Technique (SMOTE) on the classification performance using the Multinomial Naïve Bayes (MNB) method. Each scenario's implications are thoroughly examined, considering the nuances of label distribution, class imbalance, and the effect of different k-values in SMOTE.

3.1. Scenario-Based Testing

The scenarios designed for this research encompass a range of complexities in label distribution, class representation, and the utilization of SMOTE. These scenarios offer a rich dataset for evaluation, as outlined in Table 4. The following explore the key takeaways from each scenario.

Table 4. Scenarios Results

Scenario	Label		Smote
	National Songs	Folk Songs	
1	90	480	-
2	90	480	K = 9, 19, 29, 39, 49, 59, 69, 79, 89.
3	90	West: 312 Center: 121 East: 42	-
4	90	West: 312 Center: 121 East: 42	K = 11, 21, 31, 41.
5	90	30 Provinces Maximum: 42 (West Java) Minimum: 4 (North Kalimantan)	-
6	90	30 Provinces Maximum: 42 (West Java) Minimum: 4 (North Kalimantan)	K = 2, 3.

Scenario 1 and 2: In scenarios 1 and 2, the dataset was divided into two labels - national songs and folk songs. The critical distinction was the application of SMOTE in scenario 2 with varying k-values. The results, as displayed in Table 5, demonstrate that employing SMOTE with k-values between 9 and 89 significantly enhances accuracy, precision, and recall compared to scenario 1 without SMOTE. This underscores the importance of addressing class imbalances for improved classification performance.

Table 5. Comparison of Classification Performance in Scenarios 1 and 2

Value k	Accuracy (%)	Precision (%)	Recall (%)
9	92,92	93,86	92,93
19	92,81	93,75	92,82
29	93,12	94,07	93,13
39	93,54	94,35	93,53
49	93,23	94,12	93,24
59	93,33	94,12	93,34
69	93,44	94,20	93,45
79	93,65	94,41	93,66
89	93,75	94,48	93,76
-	89,12	94,31	65,56

Scenario 3 and 4: In these scenarios, the dataset was divided into four labels based on the region (West, Center, and East). Scenario 3 omitted the use of SMOTE, while scenario 4 introduced SMOTE with k-values ranging from 11 to 41. Table 6 highlights that scenario 4, with the highest k-value of 41 in SMOTE, yielded the best accuracy, precision, and recall values. This emphasizes the need to address class imbalance and underscores the impact of the choice of k-value.

Table 6. Comparison of Classification Performance in Scenarios 3 and 4

K Value	Accuracy (%)	Precision (%)	Recall (%)
11	91,38	92,25	91,25
21	91,46	92,15	91,41
31	91,61	92,47	91,31
41	91,77	92,57	91,79
-	65,26	50,38	40,07

Scenario 5 and 6: The most complex scenarios involved the dataset being divided into 31 labels, corresponding to the provinces in Indonesia. The results in Table 7 indicate that the application of SMOTE with k-values of 2 and 3 improved classification performance significantly in scenario 6. The choice of k-value in SMOTE played a pivotal role in enhancing classification accuracy. The stark contrast in performance with and without SMOTE highlights the challenges of handling class imbalance.

Table 7. Comparison of Classification Performance in Scenarios 5 and 6

K Value	Accuracy (%)	Precision (%)	Recall (%)
2	90,65	91,41	90,66
3	90,75	91,48	90,76
-	21,58	8,45	6,03

3.2. Discussion of Results

The findings from these scenarios give rise to profound insights and implications.

- **SMOTE as an Effective Class Imbalance Mitigator:** The most discernible observation pertains to the substantial enhancement in classification performance through the application of SMOTE. SMOTE operates as a robust solution for ameliorating class imbalance by synthesizing additional

data for the minority class. This, in turn, leads to more balanced data distributions and, subsequently, elevated accuracy, precision, and recall.

- **The Significance of the K-Value:** The choice of the k-value in SMOTE stands out as a pivotal factor influencing the quality of the synthetic data. Across scenarios, higher k-values consistently corresponded to improved accuracy, precision, and recall. This underscores the capacity of SMOTE to generate diverse and representative synthetic data, thus diminishing the repercussions of class imbalance.
- **Scenario-Specific Findings:** Each scenario presented distinct challenges and opportunities. The scenario featuring two labels underscored the advantages of dealing with larger class datasets, rendering the classification process relatively straightforward. Nevertheless, even within this scenario, the incorporation of SMOTE further elevated performance. In scenarios with multiple labels, the complexities of managing class imbalance were evident, with SMOTE emerging as a critical tool in addressing these challenges.

The research findings underpin the imperative understanding of dataset characteristics, class distribution, and the role of SMOTE in bolstering text classification. The implications extend beyond this specific study, offering insights into the development of resilient text classification models that can categorize diverse textual content effectively.

4. Conclusion

In this study, we delved deep into the Multinomial Naïve Bayes (MNB) algorithm's effectiveness in classifying Indonesia's folk songs and national songs, shedding light on their regional origins. The research findings have unveiled pivotal insights that ripple through the realms of text classification and cultural heritage preservation. Notably, MNB showcased its proficiency in distinguishing between these musical genres using textual attributes. Additionally, the Synthetic Minority Over-Sampling Technique (SMOTE) emerged as a formidable asset in enhancing classification performance by rectifying class imbalances. These discoveries form the bedrock for a compelling fusion of technology and cultural conservation.

Future research trajectories beckon, advocating for a more comprehensive approach that integrates audio analysis in conjunction with lyrical content. The inclusion of audio data promises a deeper understanding of songs' regional origins, offering the potential for even more precise classification results. Refinements in data pre-processing, encompassing techniques like stemming and stopword removal, stand as promising avenues for fine-tuning datasets and augmenting classification outcomes. The expansive landscape of text classification beckons further exploration, with room for diverse Naïve Bayes algorithm models, alternative datasets, and innovative data pre-processing strategies. This research encapsulates the confluence of technology and cultural legacy, underscoring the vital importance of preserving tradition in an ever-evolving digital era.

Acknowledgment

We express our sincere gratitude to the Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, for their invaluable support and resources that were instrumental in the successful execution of this research. Their commitment to fostering academic excellence and research excellence has been a cornerstone of our pursuit of knowledge and innovation.

References

[1] M. G. C. Njoku, L. A. Jason, and R. B. Johnson, "Global Perspectives on Personal Peace, Children and Adolescents, and Social Justice," in *The Psychology of Peace Promotion*, M. G. C. Njoku, L. A. Jason, and R. B. Johnson, Eds. Cham: Springer International Publishing, p. 251, 2019, doi: 10.1007/978-3-030-14943-7 .

[2] L. Mueller *et al.*, "Agricultural Landscapes: History, Status and Challenges," in *In: Mueller, L., Sychev, V.G., Dronin, N.M., Eulenstein, F. (eds) Exploring and Optimizing Agricultural Landscapes. Innovations in Landscape Research*, 2021, pp. 3–54, doi: 10.1007/978-3-030-67448-9.

[3] R. Mountain, "Music: a versatile interface for explorations in art & science," *Interdiscip. Sci. Rev.*, vol. 47, no. 2, pp. 243–258, Apr. 2022, doi: 10.1080/03080188.2022.2035107.

[4] D. D. Wiebe, "Music and Religion: Trends in Recent English-Language Literature (2015–2021)," *Religions*, vol. 12, no. 10, p. 833, Oct. 2021, doi: 10.3390/rel12100833.

[5] Y. Zhu, "Conformity and Contestation in Cultural Production," in *Media Power and its Control in Contemporary China*, Singapore: Springer Nature Singapore, 2022, pp. 37–78, doi: 10.1007/978-981-19-6917-1_2 .

[6] J. D. Lomas and H. Xue, "Harmony in Design: A Synthesis of Literature from Classical Philosophy, the Sciences, Economics, and Design," *She Ji J. Des. Econ. Innov.*, vol. 8, no. 1, pp. 5–64, 2022, doi: 10.1016/j.sheji.2022.01.001.

[7] A. Silke and J. Morrison, "Gathering Storm: An Introduction to the Special Issue on Climate Change and Terrorism," *Terror. Polit. Violence*, vol. 34, no. 5, pp. 883–893, Jul. 2022, doi: 10.1080/09546553.2022.2069444.

[8] Y. Fauziah, S. Saifullah, and A. S. Aribowo, "Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID-19 pandemic," in *Proceeding of LPPM UPN "Veteran" Yogyakarta Conference Series 2020–Engineering and Science Series*, 2020, vol. 1, no. 1, pp. 253–261. [Online]. Available at: https://proceeding.researchsynergypress.com/index.php/ess/article/view/117.

[9] M. N. Asim, M. Wasim, M. U. Ghani Khan, N. Mahmood, and W. Mahmood, "The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval," *IEEE Access*, vol. 7, pp. 21662–21686, 2019, doi: 10.1109/ACCESS.2019.2897849.

[10] A. Ali and N. Nimat Saleem, "Classification of Software Systems attributes based on quality factors using linguistic knowledge and machine learning: A review.," *J. Educ. Sci.*, vol. 31, no. 3, pp. 66–90, Sep. 2022, doi: 10.33899/edusj.2022.134024.1245.

[11] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, and I. Pappas, "A Review on Text-Based Emotion Detection -- Techniques, Applications, Datasets, and Future Directions," p. 74, Apr. 2022. [Online]. Available at: https://arxiv.org/abs/2205.03235.

[12] J. Zhang, S. Wang, L. Chen, and P. Gallinari, "Multiple Bayesian discriminant functions for high-dimensional massive data classification," *Data Min. Knowl. Discov.*, vol. 31, no. 2, pp. 465–501, Mar. 2017, doi: 10.1007/s10618-016-0481-y.

[13] S. Saifullah, Y. Fauziah, and A. S. Aribowo, "Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety Based on Social Media Data," pp. 45-55, Jan. 2021. [Online]. Available at: 10.26555/jifo.v15i1.a20111.

[14] M. Abbas, S. Memon, K. A. Memon, A. A. Jamali, and A. Ahmed, "Multinomial Naive Bayes Classification Model for Sentiment Analysis," *Int. J. Comput. Sci. Netw. Secur.*, vol. 19, no. 3, pp. 62–67, 2019, doi: 10.13140/RG.2.2.30021.40169.

[15] A. A. Farisi, Y. Sibaroni, and S. Al Faraby, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier," *J. Phys. Conf. Ser.*, vol. 1192, p. 012024, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012024.

[16] A. P. Ardhana, D. E. Cahyani, and Winarno, "Classification of Javanese Language Level on Articles Using Multinomial Naive Bayes and N-Gram Methods," *J. Phys. Conf. Ser.*, vol. 1306, no. 1, p. 012049, Aug. 2019, doi: 10.1088/1742-6596/1306/1/012049.

[17] A. T. Akbar, R. Husaini, B. M. Akbar, and S. Saifullah, "A proposed method for handling an imbalance data in classification of blood type based on Myers-Briggs type indicator," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 4, pp. 276–283, Oct. 2020, doi: 10.14710/jtsiskom.2020.13625.

[18] B. Santoso, H. Wijayanto, K. A. Notodiputro, and B. Sartono, "Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 58, p. 012031, Mar. 2017, doi: 10.1088/1755-1315/58/1/012031.

[19] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.

[20] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70-75, Sep. 2020, doi: 10.52465/joscex.v1i1.5.

[21] P. Lestari and L. H. Sihombing, "The Portrait of Nationalism in The Superman Is Dead's Song, Jadilah Legenda," *Virtuoso J. Pengkaj. dan Pencipta. Musik*, vol. 5, no. 1, pp. 57–64, Jun. 2022, doi: 10.26740/vt.v5n1.p57-64.

[22] M. I. Munandar and J. Newton, "Indonesian EFL teachers' pedagogic beliefs and classroom practices regarding culture and interculturality," *Lang. Intercult. Commun.*, vol. 21, no. 2, pp. 158–173, Mar. 2021, doi: 10.1080/14708477.2020.1867155.

[23] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[24] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities," *Neural Comput. Appl.*, vol. 33, no. 22, pp. 15091–15118, Nov. 2021, doi: 10.1007/s00521-021-06406-8.

[25] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[26] M. A. Palomino and F. Aider, "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," *Appl. Sci.*, vol. 12, no. 17, p. 8765, Aug. 2022, doi: 10.3390/app12178765.

[27] U. Hasanah, T. Astuti, R. Wahyudi, Z. Rifai, and R. A. Pambudi, "An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian," in *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*, Nov. 2018, pp. 230–234, doi: 10.1109/ICITISEE.2018.8720957.

[28] R. Egger and E. Gokce, "Natural Language Processing (NLP): An Introduction," in *In: Egger, R. (eds) Applied Data Science in Tourism. Tourism on the Verge.*, 2022, pp. 307–334, doi: 10.1007/978-3-030-88389-8_15.

[29] N. H. Cahyana, S. Saifullah, Y. Fauziah, A. S. Aribowo, and R. Drezewski, "Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, pp. 147-151, 2022, doi: 10.14569/IJACSA.2022.0131020.

[30] T. Dogan and A. K. Uysal, "A novel term weighting scheme for text classification: TF-MONO," *J. Informetr.*, vol. 14, no. 4, p. 101076, Nov. 2020, doi: 10.1016/j.joi.2020.101076.

[31] Z. Jiang and H. N. Huynh, "Unveiling music genre structure through common-interest communities," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, p. 35, Dec. 2022, doi: 10.1007/s13278-022-00863-2.

[32] E. Dias Canedo and B. Cordeiro Mendes, "Software Requirements Classification Using Machine Learning Algorithms," *Entropy*, vol. 22, no. 9, p. 1057, Sep. 2020, doi: 10.3390/e22091057.

[33] N. J. Prottasha *et al.*, "Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning," *Sensors*, vol. 22, no. 11, p. 4157, May 2022, doi: 10.3390/s22114157.

[34] S. Kumar, A. Sharma, B. K. Reddy, S. Sachan, V. Jain, and J. Singh, "An intelligent model based on integrated inverse document frequency and multinomial Naive Bayes for current affairs news categorisation," *Int. J. Syst. Assur. Eng. Manag.*, vol. 13, no. 3, pp. 1341–1355, Jun. 2022, doi: 10.1007/s13198-021-01471-7.

[35] H. I. Abdalla and A. A. Amer, "On the integration of similarity measures with machine learning models to enhance text classification performance," *Inf. Sci. (Ny).*, vol. 614, pp. 263–288, Oct. 2022, doi: 10.1016/j.ins.2022.10.004.

[36] A. P. Rodrigues *et al.*, "Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–14, Apr. 2022, doi: 10.1155/2022/5211949.

[37] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: 10.1109/ACCESS.2021.3083638.

[38] M. Temraz and M. T. Keane, "Solving the class imbalance problem using a counterfactual method for data augmentation," *Mach. Learn. with Appl.*, vol. 9, p. 100375, Sep. 2022, doi: 10.1016/j.mlwa.2022.100375.

[39] P. Shamsolmoali, M. Zareapoor, L. Shen, A. H. Sadka, and J. Yang, "Imbalanced data learning by minority class augmentation using capsule adversarial networks," *Neurocomputing*, vol. 459, pp. 481–493, Oct. 2021, doi: 10.1016/j.neucom.2020.01.119.

[40] D. F. Oliveira, A. S. Nogueira, and M. A. Brito, "Performance Comparison of Machine Learning Algorithms in Classifying Information Technologies Incident Tickets," *AI*, vol. 3, no. 3, pp. 601–622, Jul. 2022, doi: 10.3390/ai3030035.

[41] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS One*, vol. 14, no. 11, p. e0224365, Nov. 2019, doi: 10.1371/journal.pone.0224365.