Indonesian Journal of Electrical Engineering and Computer Science
Vol. 40, No. 2, November 2025, pp. 772~779
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v40.i2.pp772-779

Optimizing social issue sentiment analysis with hybrid Chi-square and bayesian-optimized binary coordinate ascent

Guilbert Nicanor Abiera Atillo[1,2], Ralph Alanunay Cardeno[2]
[1]Computer Science and Information Technology Department, College of Arts and Sciences (CAS), Negros Oriental State University (NORSU), Dumaguete City, Philippines
[2]Graduate School, Negros Oriental State University, Dumaguete City, Philippines

Article Info

Article history:

Received Oct 31, 2024
Revised Jul 10, 2025
Accepted Oct 14, 2025

Keywords:

Bayesian optimization
Binary coordinate ascent
Chi-square (Chi2)
Feature selection
Social Sentiment analysis

ABSTRACT

Feature selection aims to reduce the dimensionality of the feature space and prevent overfitting. However, when striving to produce accurate models for sentiment classification, feature selection introduces several challenges, particularly concerning textual content. Consequently, many researchers are exploring hybrid feature selection methods to customize the selection process and develop more advanced automated techniques, recognizing that the performance of these methods depends on hyperparameters. Integrating Bayesian Optimization into binary coordinate ascent (BCA) enhances the search for optimal solutions and improves classification performance in sentiment analysis, explicitly focusing on classifying abortion sentiment using Naïve Bayes. The effectiveness of combining Chi2 feature selection with the hybridized BCA and Bayesian Optimization approach is tested across multiple n-gram configurations. Results demonstrate significant improvements in accuracy and recall compared to Chi2 and BCA hybrid methods. For instance, the Bayesian Optimization-enhanced approach achieved up to 93.80% accuracy (1-gram) and 100% recall (4-gram), outperforming the baseline method. The study highlights trade-offs between computational efficiency and performance, noting that while the Chi2 and BCA hybrid method has lower training time complexity, the Bayesian Optimization-enhanced method excels in accuracy and recall during testing. The findings suggest that integrating Bayesian Optimization into feature selection improves sentiment classification performance and recommend further exploration of this approach with other classification algorithms, especially for social issues like abortion sentiment analysis.

Corresponding Author:

Guilbert Nicanor Abiera Atillo
Computer Science and Information Technology Department, College of Arts and Sciences (CAS)
Negros Oriental State University (NORSU)
Dumaguete City, Philippines
Email: alihas@upm.edu.my

1. INTRODUCTION

Social sentiment analysis has become an increasingly vital research area, offering valuable insights into public opinions on various social, economic, political, and personal issues. Social media platforms, such as Reddit, provide a wealth of user-generated content that can be analyzed to understand collective sentiments [1]. However, the sheer volume of data generated on these platforms presents a significant challenge in effectively extracting meaningful patterns. One of the primary techniques for improving the accuracy and performance of sentiment analysis models is feature selection, which reduces the dimensionality of the data. Yet, when applied to textual data, feature selection introduces considerable challenges, such as overfitting, high computational complexity, and difficulty identifying the optimal feature

subset [2], [3]. Similarly, overfitting occurs when a model adapts excessively to the training data and fails to generalize well to new or unseen data, resulting in poor performance and unreliable predictions [4]. Moreover, the curse of dimensionality complicates identifying the most relevant features from vast datasets, leading to suboptimal classification accuracy [5]. The Naïve Bayes algorithm, although widely used for text classification, struggles with large feature sets and complex feature interactions, leading to lower performance than more advanced algorithms [4], [5].

Recent advancements have explored various methods to improve feature selection and classification performance. Filter methods, such as Chi-square (Chi2), have been widely used to identify statistically significant features in text classification tasks [6], [7]. However, these methods ignore feature interactions, leading to suboptimal feature sets that may result in overfitting [8]. On the other hand, wrapper methods, which consider feature relationships, are more effective but computationally expensive [5]. Hybrid approaches, which combine filter, wrapper, and embedded methods, have gained attention due to their ability to balance computational efficiency with accuracy [2]. These methods address redundancy and irrelevant features, improving the overall performance of sentiment analysis models. Despite these advances, integrating feature selection methods with optimization techniques such as Bayesian Optimization has not been sufficiently explored in social sentiment analysis [9], [10]. Bayesian Optimization has shown promise in improving model accuracy and recall by automating the process of hyperparameter tuning [9], [11]. However, its integration with feature selection, particularly for classifying sentiment on complex social issues like abortion, remains an underexplored area.

The main contribution of this study is to integrate Bayesian Optimization with Binary Coordinate Ascent (BCA) and Chi2 feature selection to optimize feature selection and enhance sentiment classification performance. This hybrid approach reduces overfitting by selecting the most relevant features, thereby improving the model's ability to generalize to new data. Additionally, combining Bayesian Optimization with BCA allows for efficient feature space exploration, significantly reducing computational load while maintaining high accuracy and recall, even with higher-order n-grams. This research provides a novel approach to sentiment analysis, demonstrating that integrating Bayesian Optimization with feature selection techniques can significantly improve the performance of sentiment classification models, mainly when applied to socially sensitive topics like abortion.

The paper is structured as follows: the Materials and Methods section describes the dataset, preprocessing steps, and hybrid feature selection methodology. The Results and Discussion section presents the proposed method's comparative performance against traditional Chi2 and BCA techniques, highlighting improvements in both accuracy and recall. The Conclusion summarizes the findings and suggests directions for future research in sentiment analysis using hybrid feature selection techniques.

2. MATERIALS AND METHODS

This section describes the detailed experimental procedure used to conduct sentiment analysis on abortion-related sentiment data. The methodology integrates established techniques, such as data preprocessing, feature extraction, and classification, and novel optimizations, like Bayesian Optimization and binary coordinate ascent (BCA), ensuring the study is valid, reproducible, and efficient. The methods outlined here address the key questions and knowledge gaps identified in the Introduction, particularly regarding handling missing data, selecting meaningful features from the text, optimizing model parameters, and improving sentiment classification performance.

2.1. Data collection and preparation

This research utilizes a dataset of 3,000 ProChoice_ProLife sentiment reviews scraped from Reddit threads on abortion between 2016 and 2018. The dataset contains two variables: 1) "text" for comment content and 2) "target" for binary labels (pro-choice or pro-life). Missing values were removed to maintain data integrity. The experimental setup involves a Lenovo IdeaPad 300 with an Intel Core i5-6200U processor (2.30GHz), 16GB RAM, and a 64-bit Windows 10 operating system. Python 3.11.3 was developed and deployed the hybrid feature selection model, along with libraries like Numpy, Keras, Matplotlib, Pandas, and Sci-Learn. Keras also allowed parameter configuration for the model. The stratified split was applied to divide the ProChoice_ProLife sentiment review dataset using an 80/20 split ratio, acknowledging the importance of data distribution on experiment success rates. If success rates decline, adjustments to the training ratio may be necessary [12].

2.2. Data preprocessing

The first and most crucial step in this methodology is data preprocessing, ensuring the dataset is clean, consistent, and ready for machine learning analysis. Median imputation was chosen to handle missing values, as it preserves the overall distribution of the data and is less sensitive to outliers than mean

imputation, making it ideal for text data that often contain missing values and irregularities [13]. Using median imputation, missing values in a feature are replaced with the median of the observed values for that feature, ensuring that the dataset remains intact without introducing bias.

Following imputation, TF-IDF (Term Frequency-Inverse Document Frequency) was applied for feature extraction and outlier detection. TF-IDF is a well-established technique for identifying significant terms in text data, assigning higher weights to words frequent within a document but rare across the entire corpus [14]. This ensures that important words contribute meaningfully to the sentiment analysis process while minimizing the influence of common words, such as stopwords, that do not carry relevant sentiment. The TF-IDF formula calculates each term's term frequency (TF) and inverse document frequency (IDF), with the resulting TF-IDF score reflecting the term's importance within the corpus [15]. This method ensures that the model focuses on meaningful terms and reduces the impact of irrelevant common terms.

The formula used for TF-IDF is:

$$tf(w,d) = log\ log\ (1 + fw,d) \quad (1)$$

$$idf(w,d) = \left(\frac{N}{f(w,d)}\right) \quad (2)$$

$$tfidf(w,d,D) = tfw,d * idf(w,D) \quad (3)$$

Where:

- TF is the term frequency of a word in a document.
- IDF is the inverse document frequency.
- N is the total number of documents in the corpus.
- d is the given document.
- D is the total document used.
- w is a word in document d.

Once the TF-IDF computation was completed, data transformations were performed by applying min-max normalization to the sentiment scores. This transformation scales the sentiment scores to a standard range of 0 to 1, ensuring that each feature contributes equally to the model's output. Normalization helps prevent features with larger numeric ranges from disproportionately affecting the analysis, thus ensuring that all features contribute equally to the final predictions and that outliers have minimal influence [16]. The formula for min-max normalization is:

$$Normalized\ Value = \frac{X - Xmin}{NXmax - Xmin} \quad (4)$$

X is the feature's original value, and Xmin and Xmax are its minimum and maximum values, respectively.

This transformation ensures consistency across sentiment scores derived from different abortion-related reviews.

2.3. Feature engineering

For feature engineering, both sentiment lexicons and n-grams were employed. VADER (Valence Aware Dictionary and Sentiment Reasoner) evaluated the polarity of abortion-related opinions, categorizing sentiment as positive, negative, or neutral [17]. VADER is especially effective for analyzing informal text data, such as social media content, because it accounts for emoticons, slang, and abbreviations commonly used on platforms like Reddit. Additionally, n-grams (unigrams, bigrams, trigrams, and fourgrams) were extracted to capture the text's local word patterns and contextual meanings [18], [19]. N-grams help the model understand the relationships between consecutive words, which is essential for capturing sentiment in context-dependent phrases like "not good" or "very bad" [18]. Using multiple n-gram types ensures a broad representation of word patterns, enhancing the model's ability to interpret complex sentiment expressions [20].

2.4. Feature selection and statistical analysis

Feature selection was carried out using the Chi-square (Chi2) test, a statistical method for evaluating the relationship between features (n-grams) and sentiment labels (positive, negative, and neutral) [21]. The Chi2 test compares the observed frequency of a feature with the expected frequency under the assumption of no association between the feature and the sentiment label [6]. Features with higher Chi2 scores were selected for their strong association with sentiment labels, improving the model's focus on relevant features

and reducing dimensionality. This approach ensures that only the most significant features are retained, which enhances model efficiency.

2.5. Binary coordinate ascent and bayesian optimization

Binary Coordinate Ascent (BCA) was applied to optimize the feature set further. BCA is an optimization technique that iteratively refines binary features selected by Chi2 to maximize the model's performance. It is especially effective for optimizing binary features in text-based datasets, where each feature represents the presence or absence of a term [22]. After BCA, Bayesian Optimization was used to fine-tune the model's hyperparameters. Bayesian Optimization efficiently explores the hyperparameter space by focusing on the most promising regions, leading to faster convergence and better results than traditional grid or random search methods [9]. The combination of BCA and Bayesian Optimization enables effective optimization, helping to mitigate challenges like overfitting, class imbalance, and handling zero occurrences in the dataset. In the initialization phase, a set of solutions is represented as binary vectors, either randomly generated or based on domain knowledge. During the BCA phase, the binary vector is iteratively refined by flipping individual bits to improve the objective function's value, incorporating a regularization term to prevent overfitting. In the Bayesian Optimization phase, a surrogate model, typically a Gaussian Process, predicts the objective function values for unexplored regions, guiding the search process to balance exploration and exploitation.

The iterative process alternates between the BCA and Bayesian Optimization phases until convergence criteria are met, such as reaching a maximum number of iterations, exceeding a time limit, or achieving a satisfactory objective value. This hybrid approach effectively combines the global exploration capabilities of Bayesian Optimization with the local optimization strengths of BCA, leading to enhanced model performance [23]. Figure 1 is the Bayesian-optimized Binary Coordinate Ascent (BCA) block diagram.

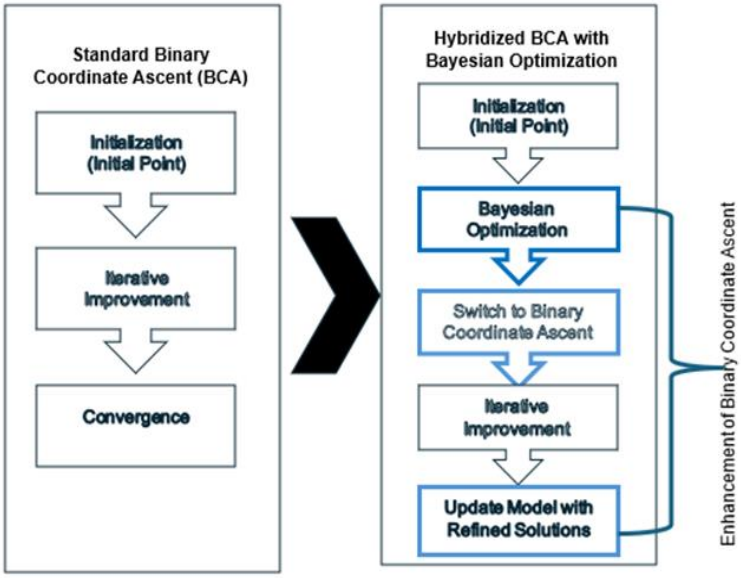

Figure 1. Bayesian-optimized Binary Coordinate Ascent (BCA)

2.6. Classification with Naïve Bayes

Naïve Bayes was selected for classification due to its simplicity, interpretability, and strong performance in text classification tasks [24], mainly when working with smaller datasets [25]. Naïve Bayes assumes feature independence, a reasonable assumption for text data, where each word is treated as an independent feature. The model was trained using features selected by Chi2 and optimized using BCA and Bayesian Optimization. The performance of the model was evaluated using accuracy and recall. Accuracy measures the proportion of correct predictions, while recall focuses on the model's ability to identify positive instances. This is especially important in sentiment analysis, where detecting positive sentiment is a priority [4]. The formulas for these metrics are:

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Where TP refers to True Positives, TN refers to True Negatives, FP refers to False Positives, and FN refers to False Negatives.

This methodology integrates well-established techniques and novel optimizations to improve the accuracy and efficiency of sentiment analysis. Ensuring the analysis is valid and reproducible by selecting feature extraction, feature selection, and optimization methods. The methodology provides a robust framework for replicating the study and confirming the results.

3. RESULTS AND DISCUSSION

The following is the model evaluation results for applying the Chi-Square and Bayesian-Optimized Binary Coordinate Ascent. Tables 1 and 2 present the accuracy and recall outcomes for applying Hybrid Chi2 with BCA and Chi2 with BCA and Bayesian Optimization. Figures 2 and 3 show these models' comparative graphical accuracy and recall performance.

Table 1 presents the accuracy results for two feature selection methods-Hybrid Chi2 with BCA and Chi2 with BCA with Bayesian Optimization-across various n-gram features (1-gram to 4-gram) on a dataset split 80/20 (training/testing). The accuracy results show that the Hybrid Chi2 with BCA method performs best with 1-gram features, achieving an accuracy of 90.16%. However, its performance decreases as the n-gram size increases, with accuracy dropping to 83.00% for 2-grams, 71.33% for 3-grams, and 66.16% for 4-grams. This trend indicates that the Hybrid Chi2 and BCA method struggles to capture the complex dependencies in higher-order n-grams, resulting in decreased performance as the feature space expands. These findings suggest that Hybrid Chi2 and BCA are better suited for simpler models that can effectively capture the relationships between terms.

Table 1. Accuracy

Hybrid Chi^2 and BCA				Chi^2 and BCA with Bayesian Optimization			
(1-gram)	(2-gram)	(3-gram)	(4-gram)	(1-gram)	(2-gram)	(3-gram)	(4-gram)
90.16%	83.00%	71.33%	66.16%	93.80%	92.69%	92.22%	92.22%

On the other hand, the Chi2 and BCA with the Bayesian Optimization method show a more consistent performance across all n-gram sizes. This method achieves the highest accuracy of 93.80% for 1-grams, with only a slight decrease to 92.69% for 2-grams and a stable 92.22% for both 3-grams and 4-grams. The results highlight the effectiveness of Bayesian Optimization in fine-tuning model parameters, enabling the method to handle the increased complexity of higher-order n-grams without significant loss of accuracy. These findings support the hypothesis that Bayesian Optimization enhances the model's robustness and adaptability to more complex feature sets.

Table 2 presents the recall results, an important metric in sentiment analysis, as it evaluates the model's ability to identify positive instances correctly. The Hybrid Chi2 and BCA method shows a gradual improvement in recall as n-gram size increases, from 93.49% for 1-grams to 98.37% for 4-grams. However, the Chi2 and BCA with Bayesian Optimization method outperform the Hybrid Chi2 and BCA method across all n-gram sizes, achieving recall values ranging from 99.48% for 1-grams to 100.00% for 4 grams. These results underscore the critical role of Bayesian Optimization in improving recall, as it allows the model to better capture complex patterns in the data, particularly for higher-order n-grams [14]. This improvement in recall emphasizes the model's effectiveness in capturing positive sentiment, which is often the focus of sentiment analysis tasks.

Table 2. Recall

Hybrid Chi^2 and BCA				Chi^2 and BCA with Bayesian Optimization			
(1-gram)	(2-gram)	(3-gram)	(4-gram)	(1-gram)	(2-gram)	(3-gram)	(4-gram)
93.49	94.85	96.20	98.37	99.48	99.30	99.82	100.00

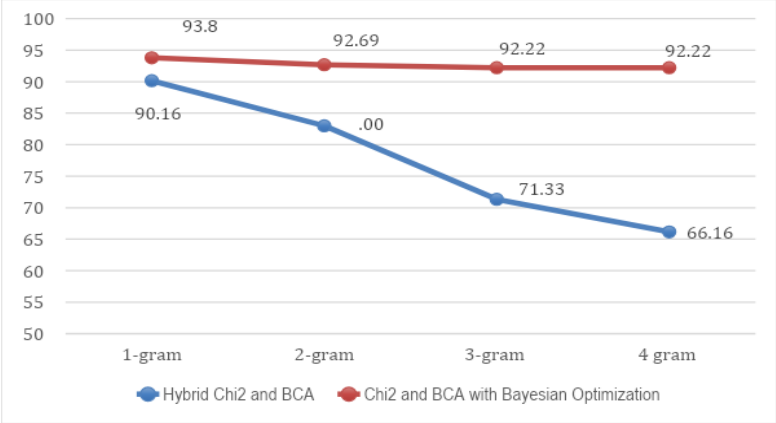

Figure 2. Graphic of Accuracy Performance

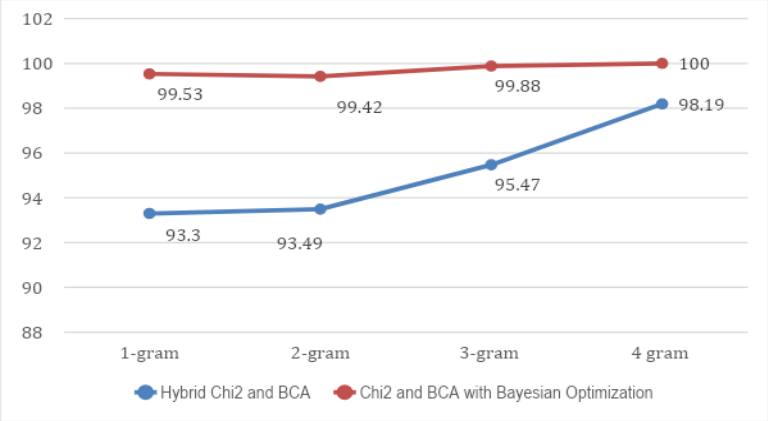

Figure 3. Graphic of Recall Performance

Furthermore, the findings of this study align with previous research suggesting that simpler models, such as unigrams, perform better in specific contexts, mainly when dealing with noisy or irregular text data. For instance, found that unigram models typically outperform more complex n-gram models in accuracy, especially in environments where feature selection and noise reduction are critical [14]. In this study, the Hybrid Chi2 and BCA methods performed best with 1-gram features, confirming [14] findings. Additionally, research on hotel review sentiment analysis, such as [17], also showed that unigrams are more effective than higher-order n-grams for sentiment classification, supporting the results observed in this study.

Conversely, applying Bayesian Optimization to enhance model performance, especially with higher-order n-grams, is consistent with findings from [9], [18]. These studies highlighted the significant improvements in model performance when Bayesian Optimization is applied to feature selection methods like Chi2, particularly in high-dimensional spaces. This study confirms these findings, demonstrating that Bayesian Optimization mitigates the performance drop associated with higher-order n-grams, maintaining high accuracy and recall.

A key strength of this study is the integration of Bayesian Optimization with Chi2 and BCA for sentiment analysis. The ability to maintain high accuracy and recall, even as the complexity of n-grams increases, showcases the robustness of the Bayesian Optimization-enhanced model. This approach is particularly valuable for applications requiring consistent performance across diverse feature complexities, such as sentiment analysis of social media content where n-gram representations vary.

However, the study has limitations. While Hybrid Chi2 and BCA performed well with 1 gram, their accuracy and recall declined with more complex n-grams. This suggests that while accuracy decreases with higher-order features, the model can still effectively capture positive sentiment. Another limitation is the computational cost of Bayesian Optimization. While it improves model performance, the optimization process can be resource-intensive, especially with large datasets or complex models. Future studies should

explore more efficient optimization techniques or hybrid models that reduce computational overhead while maintaining performance.

An unexpected finding was the Bayesian Optimization method's consistent performance, even with 4-grams. Despite the higher complexity, the model maintained excellent recall, achieving 100.00% recall for 4-grams. This suggests that Bayesian Optimization improves model performance and helps fine-tune the model to extract meaningful features from complex n-gram representations. This is an encouraging result for future sentiment analysis applications, where capturing subtle language nuances is essential.

4. CONCLUSION

In conclusion, this study demonstrates that Chi2 and BCA with Bayesian Optimization outperform the Hybrid Chi2 and BCA methods across all n-gram sizes, particularly regarding accuracy and recall. Bayesian Optimization is critical in enhancing model performance, ensuring high accuracy even with more complex n-gram features. The findings confirm the importance of optimization techniques for improving the performance of feature selection models in sentiment analysis tasks and suggest that Bayesian Optimization offers a promising approach for handling complex text.

Additionally, the results of this study suggest several avenues for future research. First, exploring Bayesian Optimization with other feature selection methods beyond Chi2 could help determine whether this approach can be generalized to a broader range of text classification tasks. Second, further research is needed to explore the computational efficiency of Bayesian Optimization in large-scale datasets and its potential for real-time applications. Finally, examining the impact of additional model fine-tuning techniques, such as ensemble methods or deep learning architectures, could provide further insights into improving the accuracy and recall of sentiment analysis models, particularly when handling high-dimensional, unstructured text data.

FUNDING INFORMATION

Authors state no funding involved.

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P. M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Systems with Applications*, vol. 223, p. 119862, Aug. 2023, doi: 10.1016/j.eswa.2023.119862.

[2] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, no. 2, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.

[3] M. García-Torres, R. Ruiz, and F. Divina, "Evolutionary feature selection on high dimensional data using a search space reduction approach," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105556, Jan. 2023, doi: 10.1016/j.engappai.2022.105556.

[4] S. Dey Sarkar, S. Goswami, A. Agarwal, and J. Aktar, "A novel feature selection technique for text classification using Naïve Bayes," *International Scholarly Research Notices*, vol. 2014, pp. 1–10, Oct. 2014, doi: 10.1155/2014/717092.

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

[6] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, "Comprehensive review of text-mining applications in finance," *Financial Innovation*, vol. 6, no. 1, p. 39, Dec. 2020, doi: 10.1186/s40854-020-00205-1.

[7] Y. Cahyono, "Sentiment analysis pada Sosial Media Twitter using Naïve Bayes classifier with feature selection particle swarm optimization dan term frequency," *Jurnal Informatika Universitas Pamulang*, vol. 2, no. 1, p. 14, 2017, doi: 10.32493/informatika.v2i1.1500.

[8] M. Frąckiewicz, "Feature selection methods in AI: filter, wrapper, and embedded techniques," TS2 Space. Accessed: Jul. 16, 2025. [Online]. Available: https://ts2.space/en/feature-selection-methods-in-ai-filter-wrapper-and-embedded-techniques/

[9] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: a review of Bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016, doi: 10.1109/JPROC.2015.2494218.

[10] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, 2012, pp. 2951–2959.

[11] Y. Mate and N. Somai, "Hybrid feature selection and Bayesian optimization with machine learning for Breast Cancer prediction," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2021,

pp. 612–619. doi: 10.1109/ICACCS51430.2021.9441914.
[12] M. K. Uçar, M. Nour, H. Sindi, and K. Polat, "The effect of training and testing process on machine learning in biomedical datasets," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–17, May 2020, doi: 10.1155/2020/2836236.
[13] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, Mar. 2021, doi: 10.3389/fenrg.2021.652801.
[14] J Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003, pp. 29–48.
[15] A. Addiga and S. Bagui, "Sentiment analysis on twitter data using term frequency-inverse document frequency," *Journal of Computer and Communications*, vol. 10, no. 08, pp. 117–128, 2022, doi: 10.4236/jcc.2022.108008.
[16] L. Zheng, H. Wang, and S. Gao, "Sentimental feature selection for sentiment analysis of chinese online reviews," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 1, pp. 75–84, Jan. 2018, doi: 10.1007/s13042-015-0347-4.
[17] C. J. Hutto and E. Gilbert, "VADER: a parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/icwsm.v8i1.14550.
[18] S. Liu and T. Forss, "Combining N-gram based similarity analysis with sentiment analysis in web content classification," in *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, SCITEPRESS - Science and and Technology Publications, 2014, pp. 530–537. doi: 10.5220/0005170305300537.
[19] L. Zhu, W. Wang, M. Huang, M. Chen, Y. Wang, and Z. Cai, "A N-gram based approach to auto-extracting topics from research articles," *Journal of Intelligent and Fuzzy Systems*, vol. 43, no. 5, pp. 6137–6146, Sep. 2022, doi: 10.3233/JIFS-220115.
[20] V. H. Nguyen, H. T. Nguyen, H. N. Duong, and V. Snasel, "N-Gram-based text compression," *Computational Intelligence and Neuroscience*, vol. 2016, 2016, doi: 10.1155/2016/9483646.
[21] S. Vairavasundaram and L. R., "Applying semantic relations for automatic topic ontology construction," in *Developments and trends in intelligent technologies and smart systems. IGI Global*, 2017, pp. 48–77. doi: 10.4018/978-1-5225-3686-4.ch004.
[22] Z. Liu, J. Yang, L. Wang, and Y. Chang, "A novel relation aware wrapper method for feature selection," *Pattern Recognition*, vol. 140, p. 109566, Aug. 2023, doi: 10.1016/j.patcog.2023.109566.
[23] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of internet restaurant reviews written in Cantonese," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7674–7682, Jun. 2011, doi: 10.1016/j.eswa.2010.12.147.
[24] A. Solikhatun and E. Sugiharti, "Application of the Naïve Bayes classifier algorithm using N- Gram and information gain to improve the accuracy of restaurant review sentiment analysis," *Journal of Advances in Information Systems and Technology*, vol. 2, no. 2, pp. 1–12, 2020.
[25] G. Liu, Y. Luo, and J. Sheng, "Research on application of naive bayes algorithm based on attribute correlation to unmanned driving ethical dilemma," *Mathematical Problems in Engineering*, vol. 2022, pp. 1–9, Aug. 2022, doi: 10.1155/2022/4163419.

BIOGRAPHIES OF AUTHORS

Dr. Guilbert Nicanor Abiera Atillo holds dual doctorates in Information Technology and Public Administration. He is an accomplished researcher who has done extensive sentiment analysis and has published in socio-economic studies. His research contributions have been presented at various international conferences. Dr. Atillo currently serves as the Director of Community Extension Services and Manager of FABLAB-Negros. He is deeply involved in community development initiatives, consulting local cooperatives, and leading efforts aimed at community empowerment. Dr. Atillo is an active member of the Philippine Society of Information Technology Educators and the Philippine Association of Graduate Educators. His work in Information Technology and Public Administration has earned local and international recognition. He can be contacted at email: guilbertnicanor.atillo@norsu.edu.ph.

Dr. Ralph A. Cardeño is a Professor 2 at the Negros Oriental State University (NOrSU), Negros Oriental, Dumaguete City, the Philippines. He holds a Doctor of Philosophy in English (Language as Concentration) degree. Aside from teaching courses in the same university's undergraduate and graduate programs, he has traveled to present papers at various international conferences here and abroad. He has published research papers in refereed journals focusing on the lens of both quantitative and qualitative research explorations. His research focuses on reading literacy, discourse, critical discourse analyses, conversation analysis, sociolinguistics, and pragmatics. At present, he serves as the Assistant Dean of the College of Arts and Sciences and at the same time, the Director for Curriculum and Instruction at NOrSU. He can be contacted at email: alihas@upm.edu.my.