Jurnal Teknik Informatika C.I.T Medicom, 16 (3) (2024) 143-157

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Topic modeling using LDA and performance evaluation of classification algorithm: k-NN, SVM, NBC, and DT

Yerik Afrianto Singgalen
Tourism Department, Faculty of Business Administration and Communication, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia

Article Info

Article history

Received : Apr 05, 2024
Revised : May 21, 2024
Accepted : Jun 25, 2024

Keywords:

Classification Algorithms;
Data Analysis Framework;
Latent Dirichlet Allocation (LDA);
Machine Learning;
Topic Modeling.

Abstract

This research investigates the integration of Latent Dirichlet Allocation (LDA) for topic modeling with the performance evaluation of various classification algorithms—specifically, k-nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), and Decision Trees (DT)—within the Digital Content Reviews and Analysis Framework. The framework systematically processes and analyzes digital content, including data cleaning, extraction, evaluation, and visualization techniques, to enhance machine learning models' interpretability and predictive accuracy. The study demonstrates that combining LDA with these classification algorithms significantly improves data interpretation and model performance, particularly in handling large-scale textual datasets. Notably, the Decision Tree algorithm achieved a 98.86% accuracy post-SMOTE. At the same time, the Support Vector Machine reached a near-perfect AUC of 1.000, highlighting the efficacy of these methods in managing imbalanced datasets. The findings provide valuable insights for optimizing model selection and developing more robust and adaptive machine-learning models across various applications. This research contributes to advancing the field of artificial intelligence by proposing a comprehensive framework that effectively addresses complex data-driven challenges, encouraging further exploration of more flexible and scalable models to accommodate evolving data environments.

Corresponding Author:

Yerik Afrianto Singgalen,
Tourism Department, Faculty of Business Administration and Communication
Atma Jaya Catholic University of Indonesia
Jl. Jend. Sudirman No.51 5, RT.004/RW.4, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12930
yerik.afrianto@atmajaya.ac.id

1. Introduction

The urgency of research on topic modeling using Latent Dirichlet Allocation (LDA) and performance evaluation of various classification algorithms, such as k-nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), and Decision Trees (DT), lies in the critical need to enhance the accuracy and efficiency of data-driven decision-making processes. With the exponential growth of textual data across diverse domains, LDA emerges as a robust method for uncovering hidden thematic structures within vast datasets, facilitating more refined data categorization and information retrieval [1]–[6]. Concurrently, evaluating the performance of different classification algorithms is essential to determine their effectiveness in handling complex and high-dimensional data [7]–[12]. It is argued that understanding the strengths and weaknesses of each algorithm provides valuable insights

into optimizing model selection and improving predictive accuracy in various applications, such as sentiment analysis, spam detection, and medical diagnosis [13]–[15]. Furthermore, a comprehensive analysis of these algorithms, considering factors like computational efficiency, scalability, and interpretability, underscores their potential to address specific challenges in real-world scenarios. This research is crucial as it contributes to developing more sophisticated and adaptive machine-learning models, ultimately advancing the field of artificial intelligence and its practical applications.

The primary aim of this research on topic modeling using Latent Dirichlet Allocation (LDA) and the performance evaluation of classification algorithms, specifically, k-nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), and Decision Trees (DT) is to advance the understanding and application of machine learning techniques in the processing and analysis of large textual datasets. The study employs LDA to identify and extract latent themes from extensive textual data systematically, enhancing data interpretation and thematic categorization [16]–[20]. Concurrently, the performance evaluation of different classification algorithms serves to discern the most effective models for accurately classifying and predicting outcomes within diverse datasets [20], [21]. It is posited that such a dual approach not only refines the model selection process but also bolsters the development of more robust and adaptable algorithms tailored to specific application needs. This comprehensive examination of topic modeling and classification performance is pivotal in optimizing machine learning workflows, contributing to more effective data-driven decision-making across various domains.

The novelty of this research lies in its innovative approach to integrating topic modeling via Latent Dirichlet Allocation (LDA) with a rigorous performance evaluation of classification algorithms, namely k-nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), and Decision Trees (DT). This study distinguishes itself by not only uncovering latent thematic structures within large-scale textual data using LDA but also systematically assessing the efficacy of various classification models in interpreting these structures. The research introduces a unique methodology that bridges the gap between data interpretation and predictive accuracy by juxtaposing topic modeling with algorithmic performance. This approach is particularly significant in scenarios where understanding the underlying topics within data is as critical as ensuring that the classification models employed are accurate and efficient. Integrating these two dimensions—topic extraction and classification performance—offers a novel perspective that enhances both the interpretability and applicability of machine learning techniques across diverse fields. Through this dual focus, the research contributes to advancing the state of knowledge in machine learning, offering valuable insights that push the boundaries of current methodologies.

This research offers substantial theoretical and practical contributions by enhancing the understanding of machine learning techniques and their applications in real-world scenarios. Theoretically, it provides a robust framework for combining topic modeling through Latent Dirichlet Allocation (LDA) with the performance evaluation of various classification algorithms, such as k-nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), and Decision Trees (DT). This dual approach advances the existing theoretical models for text analysis and classification and introduces new insights into the relationship between data representation and classification accuracy [22]–[24]. The study's findings have significant implications for fields that rely heavily on data analysis, such as natural language processing, market analysis, and healthcare analytics [25], [26]. By identifying the most effective algorithms for specific types of data and contexts, the research aids in optimizing machine learning workflows, thereby enhancing predictive capabilities and decision-making processes. This blend of theoretical innovation and practical application positions the research as a pivotal contribution to academic study and the practical deployment of advanced machine-learning methodologies.

Research exploring similar domains has focused extensively on integrating topic modeling techniques with machine learning algorithms to enhance data analysis and predictive accuracy. Studies have often employed Latent Dirichlet Allocation (LDA) for topic modeling, coupled with various classification algorithms like k-nearest Neighbors (k-NN) [27], [28], Support Vector Machines (SVM)

[29], [30], Naive Bayes Classifier (NBC) [31]–[33], and Decision Trees (DT) [34], [35], to categorize and predict outcomes based on textual data. It is argued that these efforts primarily concentrate on optimizing the individual components, either refining topic modeling techniques to achieve more nuanced data interpretation or enhancing the classification algorithms to improve accuracy and efficiency. However, a noticeable gap exists in the literature regarding the combined evaluation of these methods within a single comprehensive framework. While previous research has provided valuable insights into specific aspects of machine learning applications, a more holistic approach that addresses topic modeling and classification performance remains underexplored. This observation highlights the need for further studies integrating these methodologies to develop more sophisticated tools capable of tackling complex data-driven challenges across various domains.

The limitations of this research primarily stem from the inherent constraints of the methodologies and algorithms employed. The integration of Latent Dirichlet Allocation (LDA) with various classification algorithms, such as k-nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), and Decision Trees (DT), offers valuable insights into data interpretation and classification, it also presents specific challenges. One significant limitation is the reliance on predefined parameters and assumptions within these algorithms, which may not fully capture the complexities of diverse datasets or adapt well to dynamic data environments. Additionally, the performance of these algorithms can be significantly influenced by the quality and size of the input data, potentially limiting the generalizability of the findings to other contexts or domains. Furthermore, the study's focus on specific machine learning models may exclude alternative approaches that could offer different perspectives or more robust solutions to the problem. These limitations suggest a need for future research to explore more flexible and adaptive models that can better accommodate the evolving nature of data in various fields.

Future research should prioritize exploring more adaptive and scalable machine learning models to better accommodate data's evolving complexity and diversity across various domains. Expanding the focus beyond the current methodologies, it is advisable to investigate the integration of deep learning techniques with traditional algorithms such as k-nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), and Decision Trees (DT) to enhance predictive accuracy and model robustness. Additionally, incorporating real-time data processing and dynamic parameter adjustment could address the limitations of static model configurations, offering a more responsive approach to data analysis. This direction is particularly relevant as it aligns with the increasing demand for models that can adapt to changes in data patterns and maintain performance across different contexts. Emphasizing interdisciplinary approaches by combining insights from data science, domain expertise, and advanced computational techniques may yield innovative solutions that push the boundaries of current knowledge and practice. Such efforts will contribute to developing more sophisticated and reliable models, ultimately advancing the application of machine learning in both academic and practical settings.

2. Research Methodology

The framework employed in this research, the Digital Content Reviews and Analysis Framework, is designed to systematically process and analyze digital content systematically, mainly focusing on content reviews and travel vlogs. This framework encompasses several interconnected stages, beginning with raw data acquisition and then data cleaning and selection to ensure quality and relevance. Data extraction then plays a crucial role in isolating pertinent information for further analysis. The framework integrates data processing steps, which streamline the transformation of raw data into a more structured form, enabling more effective evaluation and visualization. Integrating model performance evaluation and data visualization techniques within this framework enhances the interpretability of the analyzed content and provides valuable insights into patterns and trends within digital media. This comprehensive approach facilitates a deeper understanding of content dynamics, ensuring that the analysis is rigorous and meaningful, thereby advancing the field of digital content analysis.

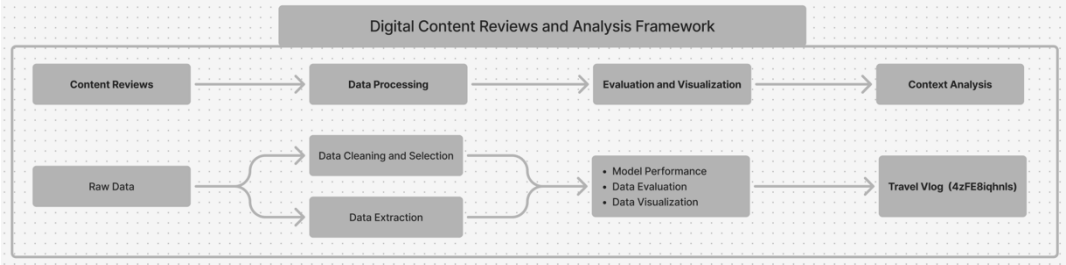

Figure. 1. Implementation of Digital Content Reviews and Analysis Framework

The data to be processed in this study is sourced from a YouTube video titled "Kelana Saka - Travel Vlog | Episode 1," which chronicles a journey through Eastern Indonesia, precisely in Ambon, Maluku. With 54,122 views and 172 comments since its premiere on December 23, 2023, this travel vlog provides a rich repository of user-generated content that reflects diverse viewer engagement and perspectives. The video's popularity and the substantial number of comments suggest a high level of viewer interaction, making it an ideal dataset for analyzing audience responses and sentiments. The study aims to uncover patterns in viewer behavior and sentiment towards the destination featured in the vlog by examining quantitative metrics like view counts and qualitative data like comment content. This approach is expected to yield valuable insights into digital content consumption patterns and viewer engagement, contributing to a more nuanced understanding of audience dynamics in travel-related media.

Data cleaning is critical in preparing datasets for analysis, ensuring accuracy and reliability in the outcomes. Initially, raw text data undergoes preprocessing, which involves tokenization, removal of stop words, and stemming or lemmatization to normalize the text. Next, relevant attributes are selected based on their importance to the research objectives, effectively reducing dimensionality and enhancing the dataset's focus. The next step involves removing duplicates to eliminate redundant entries, which helps minimize biases and errors in subsequent analyses. This procedure not only refines the dataset by maintaining unique and relevant data points but also ensures the integrity and quality of the data. A well-structured data-cleaning process, therefore, contributes significantly to the validity of any analytical results derived, facilitating more accurate, consistent, and trustworthy outcomes.

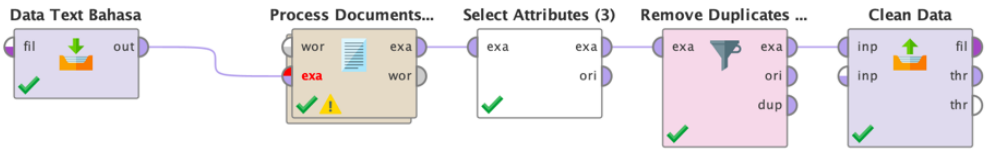

Figure 2. Cleaning Process

The outcome of data cleaning plays a pivotal role in shaping the reliability and validity of any research findings. The refined dataset emerges with enhanced clarity by systematically removing noise and inconsistencies, allowing for more precise and insightful analysis. This meticulous process involves eliminating irrelevant entries, correcting inaccuracies, and standardizing formats, collectively contributing to a dataset that reflects the underlying phenomena without distortions. Such refined data enables a more robust interpretation of patterns and relationships, reinforcing the strength and credibility of the analytical conclusions drawn. A thoroughly cleaned dataset is a solid foundation upon which sound, evidence-based decisions, and theoretical advancements can be confidently built.

Data extraction is an essential operation in data analysis, designed to retrieve relevant information from raw datasets systematically. Initially, this process involves preprocessing the data to

ensure uniformity and consistency, which is crucial for practical analysis. Then, specific attributes are selected based on their relevance to the analytical goals, narrowing the focus to only the most pertinent data. Subsequently, sentiment analysis might be applied, particularly in text data, to classify the data, such as positive, negative, or neutral sentiments, thereby enriching the dataset with valuable contextual insights. Removing duplicates at this stage is critical, as it prevents data redundancy and enhances the dataset's quality by maintaining unique entries. A well-executed data extraction process not only streamlines the dataset for subsequent analyses but also enhances the overall reliability and validity of the findings derived from the data.

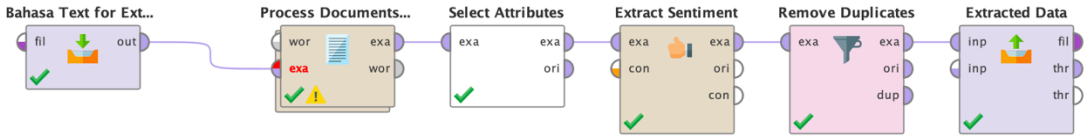

Figure 3. Extraction Process

The results of the data extraction process significantly determine the quality and depth of subsequent analytical insights. After carefully extracting relevant attributes and sentiment information, the refined dataset becomes a robust foundation for deeper exploration and interpretation. This curated data is devoid of redundancies and enriched with meaningful patterns, allowing for a nuanced understanding of underlying trends and relationships within the data. By focusing on the most pertinent information and eliminating irrelevant or duplicate entries, the dataset's coherence and relevance are markedly enhanced, leading to more accurate and insightful outcomes. A well-prepared extracted dataset thus catalyzes valuable conclusions, enabling more precise forecasting, pattern recognition, and decision-making processes.

Topic modeling is an advanced computational method designed to identify patterns within textual data by categorizing words into clusters representing various topics. The procedure begins with document preprocessing, which includes tokenization and normalization, to prepare the text for more intricate analysis. Next, Latent Dirichlet Allocation (LDA) models detect word associations and generate topic distributions across the corpus. This stage is critical as it uncovers latent structures in the data, allowing for extracting meaningful patterns that reflect the core themes discussed within the documents. The rationale behind topic modeling lies in its ability to distill complex textual information into discernible themes, significantly aiding in interpreting large datasets. By transforming unstructured text into structured insights, topic modeling is vital for enhancing understanding and supporting decision-making processes in various fields, ranging from social sciences to market analysis.

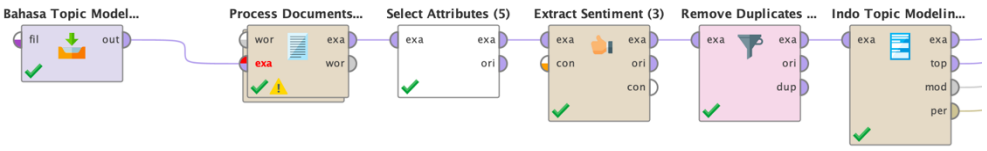

Figure 4. Topic Modeling LDA

The provided dataset performance metrics indicate that the topic modeling results reveal a nuanced understanding of the underlying textual data. The model's log-likelihood of -9118.442 and perplexity of 469.673 suggest a moderate fit to the data, implying that while the model captures some patterns in the text, there is room for optimization. The average number of tokens per document (111.300) and average document entropy (3.599) indicate a diverse set of topics, which can enrich the interpretative power of the analysis. Additionally, the average word length (5.160) and coherence score (-17.452) reflect a need for further refinement in topic coherence to ensure a more meaningful grouping of

terms. The metrics on distribution uniformity (Avg(uniform_dist): 2.349, Avg(corpus_dist): 2.201) and adequate number of words (56.123) suggest a balanced representation of terms across topics. The low average token-document difference (0.004) and moderate exclusivity score (0.387) indicate a reasonable distinction between topics, though further enhancement could improve exclusivity. AlphaSum (12.936) and Beta (0.418) values reflect the model's assumptions about topic distributions and word distributions within topics, which are crucial for understanding topic coherence and diversity. Collectively, these performance indicators underscore the effectiveness of the topic model in generating insightful thematic structures while highlighting potential areas for improvement to achieve greater precision and interpretability in future analyses.

Figure 5. Topic Modeling LDA

The visualization generated through Topic Modeling using Latent Dirichlet Allocation (LDA) reveals key thematic patterns and prevalent terms from the comments section of the travel vlog. Words such as "Ambon," "pantai," "Sudah," and "bagus" emerge prominently, indicating frequent discussions around the location and its appealing features. This suggests a strong viewer interest in Ambon's natural and cultural attractions and positive sentiments towards the vlog content. The recurrence of terms like "tua," "nyanya," and "datang" may reflect narratives or personal stories shared by viewers, thereby highlighting the interactive and engaging nature of the comment section. Such findings indicate a broader viewer engagement trend characterized by appreciative expressions and conversational interactions. Analyzing these patterns provides valuable insights into the audience's preferences and the effectiveness of the vlog in capturing viewer interest. Ultimately, these insights contribute to a deeper understanding of digital audience behaviors and the potential influence of multimedia content in shaping perceptions of travel destinations.

The performance evaluation of the Naive Bayes Classifier (NBC), Decision Tree (DT), k-nearest Neighbors (k-NN), and Support Vector Machine (SVM) with and without the application of the Synthetic Minority Over-sampling Technique (SMOTE) reveals significant insights into their predictive capabilities on imbalanced datasets. The models without SMOTE generally exhibit lower accuracy and recall, particularly in handling minority classes, as the imbalanced data leads to a biased model that favors the majority class. Conversely, applying SMOTE improves the training dataset's balance, enhancing the classifiers' ability to correctly predict both minority and majority classes, thereby improving overall model performance. Among the models evaluated, the Decision Tree and SVM with SMOTE show marked improvements in classification metrics, suggesting that these

algorithms are particularly well-suited to benefit from the augmented sample diversity provided by SMOTE. The enhancement in the k-NN model's performance also underscores the importance of balanced class representation, as its distance-based classification is heavily reliant on the availability of sufficient minority class instances. Consequently, employing SMOTE as a preprocessing step in machine learning pipelines is a valuable strategy for boosting model robustness and ensuring more equitable predictive accuracy across all classes.

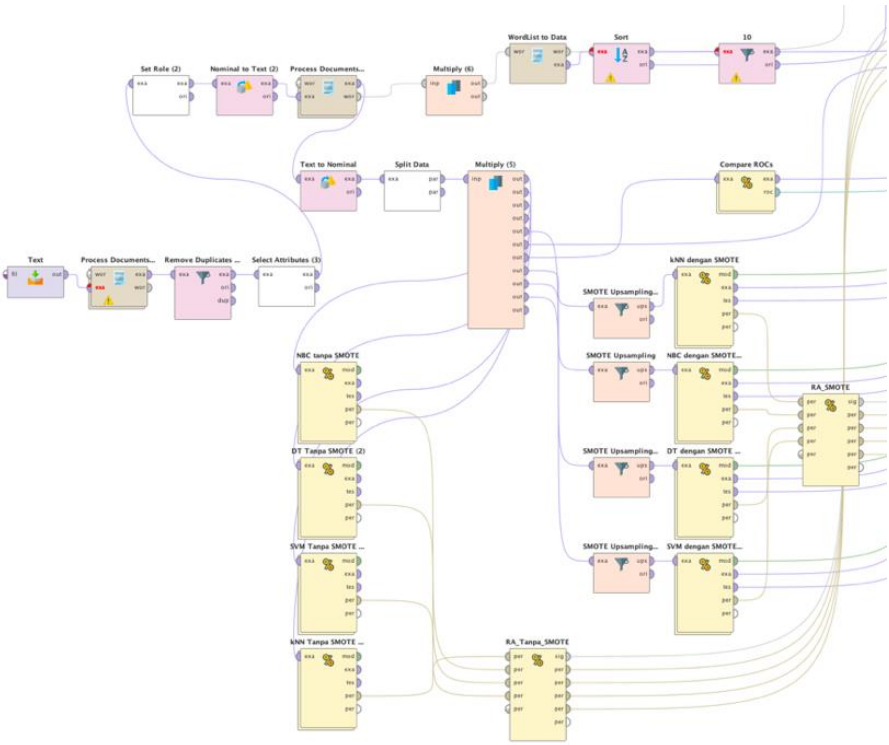

Figure 6. Performance Evaluation

Synthetic Minority Over-sampling Technique (SMOTE) significantly impacts classification by addressing the common issue of class imbalance in datasets, which can lead to biased and inaccurate model predictions. In an imbalanced dataset, models tend to favor the majority class because there are more instances of that class to learn from during training, resulting in poor performance when predicting the minority class. SMOTE mitigates this problem by generating synthetic samples for the minority class, thereby increasing its representation in the dataset. This is achieved by creating new, synthetic data points that are interpolations between existing minority class samples. As a result, SMOTE enhances the classifier's ability to learn the decision boundaries more accurately for both classes, leading to improved sensitivity and recall for the minority class without affecting the specificity or precision for the majority class. This balanced approach ensures the model performs more reliably across all classes, enhancing its overall predictive power and robustness.

3. Result and Disscusion

Decision Tree

The performance metrics for the Decision Tree (DT) model without using SMOTE reveal a high accuracy rate of 97.75%, with a precision similarly at 97.75%, indicating the model's effectiveness in predicting the majority class correctly. However, the confusion matrix shows that while the model successfully identifies 131 positive cases, it fails to classify any negative cases, highlighting a significant bias towards the positive class. This bias is further reflected in the recall rate, which is 100%, suggesting

that the model correctly identified all actual positive instances but failed to represent the negative class. The Area Under the Curve (AUC) values also tell the story: an optimistic AUC of 0.650, a pessimistic AUC of 0.350, and a standard AUC of 0.500 demonstrate a limited ability to distinguish between the classes effectively. These metrics suggest that although the DT model achieves high accuracy and recall, its performance is skewed due to the imbalance in the dataset, leading to an inflated assessment of its actual classification ability. Consequently, the model's overall effectiveness is compromised when applied to datasets with significant class imbalance, emphasizing the need for strategies like SMOTE to enhance the robustness and fairness of predictive modeling.

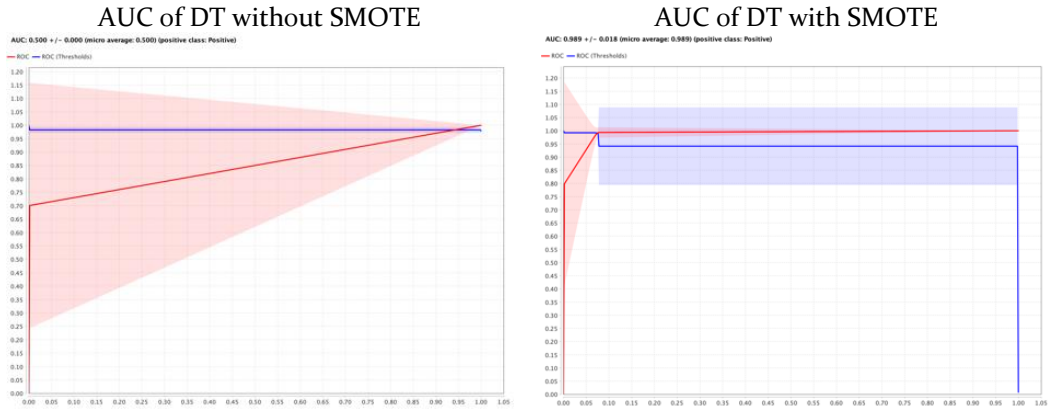

Figure 7. DT Performance

After applying SMOTE, the Decision Tree (DT) model significantly enhances classification metrics, particularly in handling an imbalanced dataset. The model achieves an impressive accuracy of 98.86% and a precision of 98.52%, reflecting its ability to classify positive and negative instances with minimal error correctly. The confusion matrix further underscores this balanced performance, with 129 true negatives, 130 true positives, and only minimal misclassifications (1 false positive and two false negatives). The high Area Under the Curve (AUC) values—optimistic at 1.000, standard at 0.989, and pessimistic at 0.978—indicate a robust capacity to distinguish between the classes, showcasing near-perfect sensitivity and specificity. The recall rate of 99.29% also highlights the model's exceptional ability to identify almost all positive cases, enhancing its reliability for predictive tasks. This overall improvement in performance metrics suggests that incorporating SMOTE effectively mitigates the challenges of class imbalance, leading to a more accurate, equitable, and generalizable decision-making model. Thus, SMOTE is a valuable strategy in optimizing the Decision Tree's capability to effectively handle diverse and unbalanced datasets.

Naïve Bayes Classifier

The performance evaluation of the Naive Bayes Classifier (NBC) without applying SMOTE indicates suboptimal handling of class imbalance, as reflected by an accuracy of 63.35% and a recall rate of 64.69%. The confusion matrix reveals a significant limitation in predicting negative instances, with no true negatives identified and 46 false negatives, suggesting a pronounced bias towards the positive class. Despite a high precision of 96.35%, which indicates the model's effectiveness in correctly identifying optimistic predictions, the Area Under the Curve (AUC) values present a stark contrast, with an optimistic AUC of 0.178 and a pessimistic AUC of 0.000, highlighting a severe deficiency in distinguishing between the two classes. The F-measure of 76.93% further underscores the model's imbalanced performance, where high precision does not translate into effective overall classification due to poor recall. This analysis suggests that while NBC may maintain high precision, its utility in imbalanced datasets is limited without SMOTE. This is crucial for improving the model's ability to

recognize minority class instances and provide a more balanced and comprehensive classification outcome.

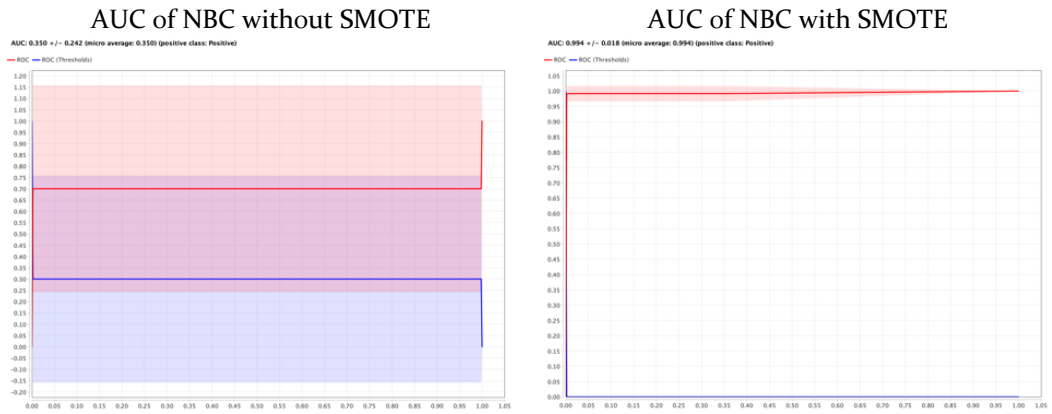

Figure 8. NBC Performance

The performance of the Naive Bayes Classifier (NBC) after applying SMOTE reveals a nuanced improvement in handling an imbalanced dataset, as evidenced by an overall accuracy of 81.00% and a perfect precision rate of 100%. The confusion matrix shows that the model correctly identifies 131 negative and 81 positive cases but misclassifies 50 positive instances as negative, reflecting a limitation in sensitivity. Despite these misclassifications, the high Area Under the Curve (AUC) values—optimistic at 0.997, standard at 0.994, and pessimistic at 0.992—indicate a strong ability of the model to distinguish between the classes effectively. However, the recall rate of 62.16% suggests that while the model is exact, it has difficulty capturing all true positive cases, leading to a moderate F-measure of 75.67%. This performance indicates that while SMOTE enhances the model's capability to identify positive instances with high precision, it does not fully address the recall deficiencies inherent in NBC, mainly when dealing with complex or overlapping class distributions. Consequently, while SMOTE contributes to balancing the dataset, further refinement or combination with other techniques may be necessary to optimize recall and precision in predictive modeling with Naive Bayes.

K-Nearest Neighbour

The performance of the k-Nearest Neighbors (k-NN) model without applying SMOTE demonstrates a solid tendency to correctly identify the majority class, achieving a high accuracy of 97.75% and a recall rate of 100%. However, the confusion matrix indicates a clear bias, as the model fails to identify any negative instances, resulting in no true negatives and only classifying 131 positive cases correctly, with three misclassifications. This imbalance is further reflected in the Area Under the Curve (AUC) values, where the optimistic AUC is 0.600, and the pessimistic AUC drops to 0.300, highlighting a significant limitation in distinguishing between classes. The low standard AUC of 0.050 suggests the model's poor ability to differentiate between positive and negative instances despite its high precision of 97.75%. The F-measure of 98.83% indicates excellent performance in predicting the positive class, yet this metric is inflated due to the class imbalance. These findings suggest that while k-NN can maintain high accuracy and recall for the majority class, its effectiveness diminishes in datasets with imbalanced class distributions. This underscores the need for techniques like SMOTE to improve the model's discriminatory power and overall balance.

AUC of k-NN without SMOTE AUC of k-NN with SMOTE

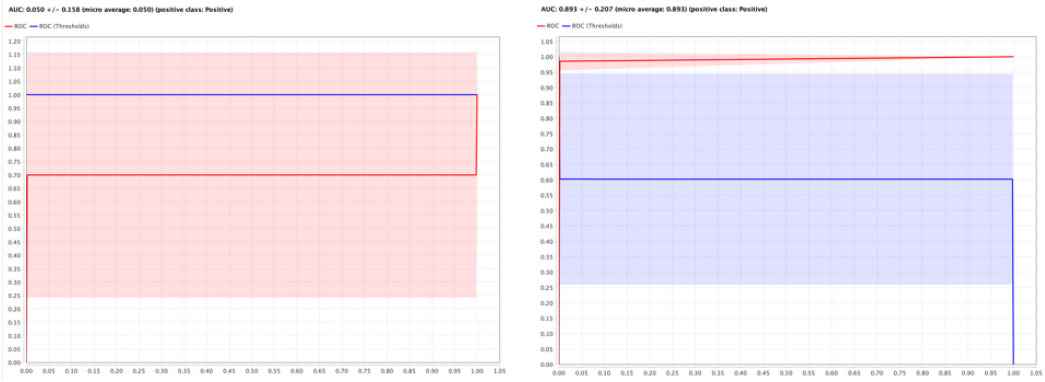

Figure 9. k-NN Performance

The performance of the k-Nearest Neighbors (k-NN) model with the application of SMOTE illustrates a significant improvement in handling class imbalances, achieving an accuracy of 97.35% and a precision of 99.23%. The confusion matrix shows that the model correctly identifies 130 true negatives and 125 true positives, with only a few misclassifications (6 false negatives and one false positive), reflecting its enhanced ability to discriminate between both classes effectively. The Area Under the Curve (AUC) values further support this enhanced performance, with an optimistic AUC of 1.000, a standard AUC of 0.893, and a pessimistic AUC of 0.985, indicating a strong and consistent capability to distinguish between positive and negative classes. The recall rate of 95.55% demonstrates that the model retains a high sensitivity to positive instances while minimizing false negatives. The F-measure of 97.26% confirms a balanced performance in precision and recall, showcasing the model's robustness in classification tasks. These results suggest that integrating SMOTE into the k-NN model significantly enhances its predictive accuracy and reliability across imbalanced datasets, making it a more effective tool for diverse analytical scenarios.

Support Vector Machine

The performance analysis of the Support Vector Machine (SVM) model without implementing SMOTE reveals a substantial accuracy of 94.73% and a precision of 97.75%, indicating the model's competence in correctly identifying positive instances. However, the confusion matrix highlights a significant drawback, with the model failing to predict any true negative cases, leading to four false positives and three false negatives. This outcome reflects a noticeable bias toward the positive class, further evidenced by the low AUC values, which remain at 0.024 across optimistic, standard, and pessimistic measures, suggesting a limited capability to distinguish between the two classes. The recall rate of 96.98% suggests that the model effectively captures most of the true positive cases, but its ability to identify negative cases is compromised. The F-measure of 97.26% indicates a satisfactory balance between precision and recall, primarily driven by the model's focus on the positive class. These findings suggest that while the SVM model maintains high precision and recall for optimistic predictions, its overall performance is hindered by an imbalance in handling negative instances. This underscores the necessity for techniques like SMOTE to enhance its discriminative power across all classes.

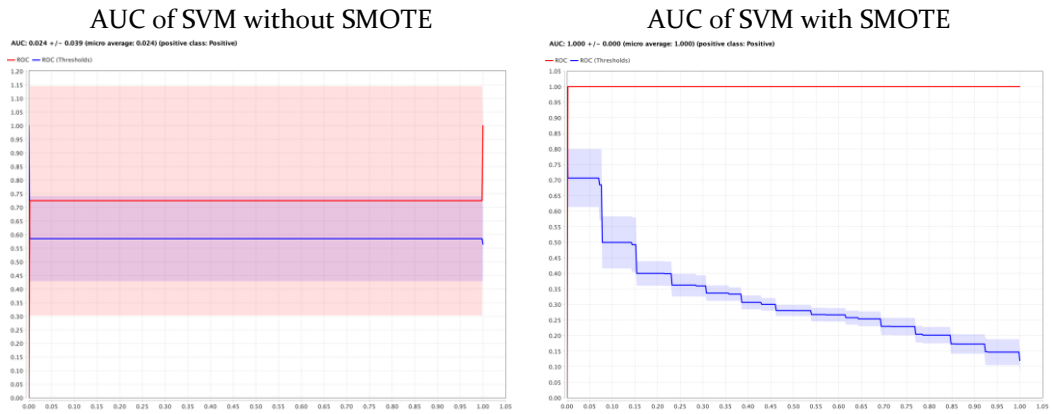

Figure 10. SVM Performance

After applying SMOTE, the performance evaluation of the Support Vector Machine (SVM) model demonstrates notable efficacy in managing class imbalances, achieving an impressive accuracy of 97.72% and a high recall rate of 99.29%. The confusion matrix shows that the model accurately predicts 126 true negatives and 130 true positives, with minimal misclassifications, including only one false negative and five false positives. The AUC values, consistently at 1.000 for optimistic, standard, and pessimistic scenarios, indicate the model's exceptional ability to distinguish between positive and negative classes, showcasing near-perfect classification performance. The precision of 96.48% reveals the model's high level of accuracy in identifying true positive cases. The F-measure of 97.80% reflects a balanced performance between precision and recall, confirming the model's robust predictive capability. These results underscore the effectiveness of SMOTE in enhancing the SVM model's performance, making it highly reliable for accurately classifying imbalanced datasets and highlighting its potential as a powerful tool in predictive analytics and decision-making processes.

ROC Comparison

The Receiver Operating Characteristic (ROC) comparison illustrates the discriminative abilities of various classification models, namely Decision Tree (DT), k-nearest Neighbors (k-NN), Support Vector Machine (SVM), and Naive Bayes Classifier (NBC). The ROC curves represent each model's actual positive rate against its false positive rate, revealing their performance across different thresholds. In this comparison, SVM, DT, and k-NN models exhibit superior performance, maintaining curves closely aligned near the top left corner of the graph, indicating high sensitivity and specificity. In contrast, the NBC model shows a lower ROC curve, reflecting a reduced ability to effectively distinguish between the positive and negative classes. This analysis suggests that while SVM, DT, and k-NN are robust in handling the classification tasks with a balanced approach, the NBC model might struggle with more nuanced or overlapping data distributions. The ROC comparison thus serves as a critical evaluation tool, highlighting the strengths and weaknesses of each model and guiding the selection of the most appropriate classifier based on the specific data characteristics and the desired balance between sensitivity and specificity.

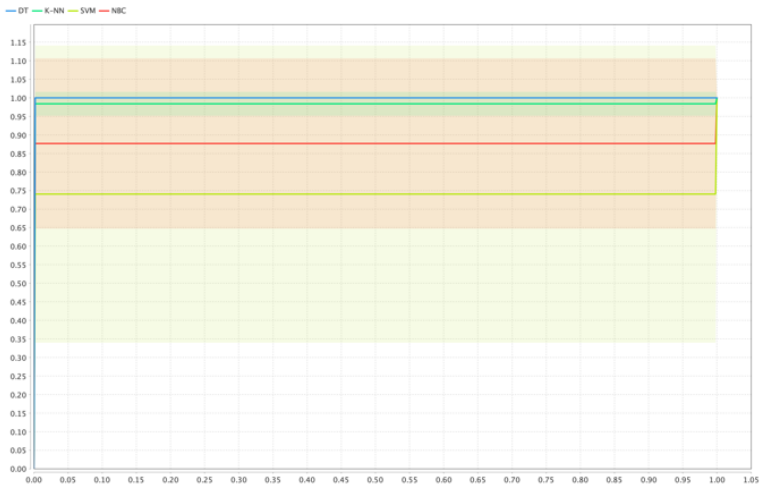

Figure 11. ROC Comparison

The analysis depicted in the graph provides a comparative evaluation of the performance of several machine learning models, including Decision Tree (DT), k-nearest Neighbors (k-NN), Support Vector Machine (SVM), and Naive Bayes Classifier (NBC). The ROC curves suggest that the SVM, DT, and k-NN models exhibit high levels of predictive accuracy and robust classification performance, as evidenced by their proximity to the top left corner, indicating excellent sensitivity and specificity. In contrast, the NBC model demonstrates relatively lower performance, indicated by a flatter curve, suggesting a diminished ability to discriminate between positive and negative cases effectively. This disparity among the models highlights the variations in their respective capacities to handle the dataset's complexities, with SVM and k-NN performing particularly well under the given conditions. The conclusion drawn from this analysis is that, for datasets with similar characteristics, SVM and k-NN are likely to provide more reliable and accurate classification outcomes, making them preferable choices over NBC for applications requiring high discrimination power.

Pairwise t-Test

The pairwise t-test results provide a statistical analysis of the differences in performance between various models, assessing whether the observed mean differences are significant. The probabilities indicate that several comparisons yield a p-value below the alpha threshold of 0.050, specifically between the mean performance values of models indexed as 0 and 1, 1 and 2, and 1 and 3, suggesting statistically significant differences in their performance. This finding implies that model 1 (0.810 ± 0.077) consistently performs differently from the others, which have higher mean values (0.974 ± 0.025, 0.989 ± 0.018, 0.977 ± 0.020). Conversely, the comparisons between models 0 and 3 (p=0.719) and 2 and 3 (p=0.197) show no statistically significant differences, indicating similar performance levels. These results emphasize the variability in model effectiveness, where model 1 is significantly less performant than the others, highlighting the importance of selecting the appropriate model based on empirical performance evidence for optimized outcomes in practical applications.

A	B	C	D	E
	0.974 +/- 0.025	0.810 +/- 0.077	0.989 +/- 0.018	0.977 +/- 0.020
0.974 +/- 0.025		0.000	0.144	0.719
0.810 +/- 0.077			0.000	0.000
0.989 +/- 0.018				0.197
0.977 +/- 0.020				

Figure 12. Pairwise t-Test Results

The analysis of the pairwise comparisons presented in the table highlights significant differences in model performance, as indicated by the calculated probabilities. A probability value lower than the alpha threshold 0.050 suggests a statistically significant difference between the compared mean performance values. Notably, model 1 (0.810 ± 0.077) shows significant differences when compared to models 0 (0.974 ± 0.025), 2 (0.989 ± 0.018), and 3 (0.977 ± 0.020), with p-values of 0.000, indicating that model 1 performs significantly worse than the others. In contrast, the comparisons between models 0 and 3 (p=0.719) and 2 and 3 (p=0.197) do not demonstrate significant differences, suggesting these models have comparable performance levels. This analysis implies that while some models exhibit equivalent efficacy, model 1's substantially lower performance warrants consideration in model selection. Consequently, these findings underscore the importance of selecting the most statistically robust model to ensure optimal performance and accuracy in predictive tasks.

4. Conclusion

This research investigates the integration of Latent Dirichlet Allocation (LDA) for topic modeling with the performance evaluation of various classification algorithms—specifically, k-nearest Neighbors (k-NN), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), and Decision Trees (DT)—within the Digital Content Reviews and Analysis Framework. The framework systematically processes and analyzes digital content, including data cleaning, extraction, evaluation, and visualization techniques, to enhance machine learning models' interpretability and predictive accuracy. The study demonstrates that combining LDA with these classification algorithms significantly improves data interpretation and model performance, particularly in handling large-scale textual datasets. Notably, the Decision Tree algorithm achieved a 98.86% accuracy post-SMOTE. At the same time, the Support Vector Machine reached a near-perfect AUC of 1.000, highlighting the efficacy of these methods in managing imbalanced datasets. The findings provide valuable insights for optimizing model selection and developing more robust and adaptive machine-learning models across various applications. This research contributes to advancing the field of artificial intelligence by proposing a comprehensive framework that effectively addresses complex data-driven challenges, encouraging further exploration of more flexible and scalable models to accommodate evolving data environments.

References

[1] Y. Feng, H. Chen, and Q. Xie, "AI Influencers in Advertising: The Role of AI Influencer-Related Attributes in Shaping Consumer Attitudes, Consumer Trust, and Perceived Influencer–Product Fit," *J. Interact. Advert.*, vol. 24, no. 1, pp. 26–47, 2023, doi: 10.1080/15252019.2023.2284355.

[2] D. Madrid-Morales, "Using Computational Text Analysis Tools to Study African Online News Content," *African Journal. Stud.*, vol. 41, no. 4, pp. 68–82, 2020, doi: 10.1080/23743670.2020.1820885.

[3] H. N. T. Thu, "Measuring guest satisfaction from online reviews: Envidence in Vietnam," *Cogent Soc. Sci.*, vol. 6, no. 1, pp. 1–14, 2020, doi: 10.1080/23311886.2020.1801117.

[4] S. Gao, H. Wang, Y. Zhu, J. Liu, and O. Tang, "Comparative relation mining of customer reviews based on a hybrid CSR method," 2023, doi: 10.1080/09540091.2023.2251717.

[5] R. X. Nie, J. H. Hu, H. Y. Zhang, J. Q. Wang, K. S. Chin, and X. Bao, "Classifying Quality Attributes of Hotel Services Considering Review Characteristics and Semantic Consistency: A Review-Driven IPA," *J. Qual. Assur. Hosp. Tour.*, vol. 00, no. 00, pp. 1–30, 2023, doi: 10.1080/1528008X.2023.2259610.

[6] Y. Feng, H. Chen, and Q. Kong, "An expert with whom i can identify: the role of narratives in influencer marketing," *Int. J. Advert.*, vol. 40, no. 7, pp. 972–993, 2021, doi: 10.1080/02650487.2020.1824751.

[7] M. Brüggemann, J. Kunert, and L. Sprengelmeyer, "Framing Food in the News: Still Keeping the Politics out of the Broccoli," *Journal. Pract.*, pp. 1–23, 2022, doi: 10.1080/17512786.2022.2153074.

[8] F. Otay Demir, Ş. Yavuz Görkem, and G. Rafferty, "An inquiry on the potential of computational literary techniques towards successful destination branding and literary tourism," *Curr. Issues Tour.*, vol. 25, no. 5, pp. 764–778, 2022, doi: 10.1080/13683500.2021.1887100.

[9] Z. Kastrati, A. S. Imran, S. M. Daudpota, M. A. Memon, and M. Kastrati, "Soaring Energy Prices: Understanding Public Engagement on Twitter Using Sentiment Analysis and Topic Modeling with Transformers," *IEEE Access*, vol. 11, no. February, pp. 26541–26553, 2023, doi: 10.1109/ACCESS.2023.3257283.

[10] T. G. Thorley and E. Saltman, "GIFCT Tech Trials: Combining Behavioural Signals to Surface Terrorist and

Violent Extremist Content Online," *Stud. Confl. Terror.*, vol. 0, no. 0, pp. 1–26, 2023, doi: 10.1080/1057610X.2023.2222901.
[11] M. P. Mehta, G. Kumar, and M. Ramkumar, "Customer expectations in the hotel industry during the COVID-19 pandemic: a global perspective using sentiment analysis," *Tour. Recreat. Res.*, vol. 48, no. 1, pp. 110–127, 2023, doi: 10.1080/02508281.2021.1894692.
[12] Q. Yang, B. Zhu, H. Liao, and X. Wu, "Learning consumer preferences from online textual reviews and ratings based on the aggregation-disaggregation paradigm with attitudinal Choquet integral," *Econ. Res. Istraz.* , vol. 36, no. 1, pp. 3059–3086, 2023, doi: 10.1080/1331677X.2022.2106282.
[13] S. Zhou, X. Ye, J. Yang, and R. Sun, "Current Issues in Tourism From turbulence to recovery : tracking the cognition-sentiment-behaviour transformation among Chinese cruise industry stakeholders," *Curr. Issues Tour.*, pp. 1–21, 2024, doi: 10.1080/13683500.2024.2329778.
[14] J. Z. Maitama, N. Idris, A. Abdi, L. Shuib, and R. Fauzi, "A systematic review on implicit and explicit aspect extraction in sentiment analysis," *IEEE Access*, vol. 8, pp. 194166–194191, 2020, doi: 10.1109/ACCESS.2020.3031217.
[15] T. Falatouri, P. Brandtner, M. Nasseri, and F. Darbanian, "Service quality dimensions in Austrian food retailing – a text mining approach for physical retail stores," *Int. Rev. Retail. Distrib. Consum. Res.*, vol. 00, no. 00, pp. 1–36, 2024, doi: 10.1080/09593969.2024.2371456.
[16] J. Wu and N. Zhao, "What consumer complaints should hoteliers prioritize? Analysis of online reviews under different market segments," *J. Hosp. Mark. Manag.*, vol. 32, no. 1, pp. 1–28, 2023, doi: 10.1080/19368623.2022.2119187.
[17] T. Ginossar, I. J. Cruickshank, E. Zheleva, J. Sulskis, and T. Berger-Wolf, "Cross-platform spread: vaccine-related content, sources, and conspiracy theories in YouTube videos shared in early Twitter COVID-19 conversations," *Hum. Vaccines Immunother.*, vol. 18, no. 1, pp. 1–13, 2022, doi: 10.1080/21645515.2021.2003647.
[18] S. M. Al-Ghuribi, S. A. Mohd Noah, and S. Tiun, "Unsupervised Semantic Approach of Aspect-Based Sentiment Analysis for Large-Scale User Reviews," *IEEE Access*, vol. 8, pp. 218592–218613, 2020, doi: 10.1109/ACCESS.2020.3042312.
[19] K. Rauniyar *et al.*, "Multi-Aspect Annotation and Analysis of Nepali Tweets on Anti-Establishment Election Discourse," *IEEE Access*, vol. 11, no. November, pp. 143092–143115, 2023, doi: 10.1109/ACCESS.2023.3342154.
[20] H. N. T. Thu, "Measuring guest satisfaction from online reviews: Envidence in Vietnam," *Cogent Soc. Sci.*, vol. 6, no. 1, 2020, doi: 10.1080/23311886.2020.1801117.
[21] J. Khan, A. Alam, and Y. Lee, "Intelligent Hybrid Feature Selection for Textual Sentiment Classification," *IEEE Access*, vol. 9, pp. 140590–140608, 2021, doi: 10.1109/ACCESS.2021.3118982.
[22] D. Buenano-Fernandez, M. Gonzalez, D. Gil, and S. Lujan-Mora, "Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach," *IEEE Access*, vol. 8, pp. 35318–35330, 2020, doi: 10.1109/ACCESS.2020.2974983.
[23] M. Zheng, K. Jiang, R. Xu, and L. Qi, "An Adaptive LDA Optimal Topic Number Selection Method in News Topic Identification," *IEEE Access*, vol. 11, pp. 92273–92284, 2023, doi: 10.1109/ACCESS.2023.3308520.
[24] P. Yang, Y. Yao, and H. Zhou, "Leveraging Global and Local Topic Popularities for LDA-Based Document Clustering," *IEEE Access*, vol. 8, pp. 24734–24745, 2020, doi: 10.1109/ACCESS.2020.2969525.
[25] Y. Zheng, Y. Long, and H. Fan, "Identifying Labor Market Competitors with Machine Learning Based on Maimai Platform," *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2064047.
[26] Z. Wang, P. Udomwong, J. Fu, and P. Onpium, "Destination image analysis and marketing strategies in emerging panda tourism: a cross-cultural perspective," *Cogent Bus. Manag.*, vol. 11, no. 1, p., 2024, doi: 10.1080/23311975.2024.2364837.
[27] M. Rodriguez-Ibanez, F. J. Gimeno-Blanes, P. M. Cuenca-Jimenez, C. Soguero-Ruiz, and J. L. Rojo-Alvarez, "Sentiment Analysis of Political Tweets from the 2019 Spanish Elections," *IEEE Access*, vol. 9, pp. 101847–101862, 2021, doi: 10.1109/ACCESS.2021.3097492.
[28] S. Moral-Garcia and J. Abellan, "Improving the Results in Credit Scoring by Increasing Diversity in Ensembles of Classifiers," *IEEE Access*, vol. 11, no. May, pp. 58451–58461, 2023, doi: 10.1109/ACCESS.2023.3284137.
[29] C. Kaveski Peres and E. Pacheco Paladini, "Exploring the attributes of hotel service quality in Florianópolis-SC, Brazil: An analysis of tripAdvisor reviews," *Cogent Bus. Manag.*, vol. 8, no. 1, pp. 1–19, 2021, doi: 10.1080/23311975.2021.1926211.
[30] S. M. A. H. Shah, S. F. H. Shah, A. Ullah, A. Rizwan, G. Atteia, and M. Alabdulhafith, "Arabic Sentiment Analysis and Sarcasm Detection Using Probabilistic Projections-Based Variational Switch Transformer,"

IEEE Access, vol. 11, no. June, pp. 67865–67881, 2023, doi: 10.1109/ACCESS.2023.3289715.

[31] Y. A. Singgalen, "Analisis Sentimen Pengunjung Pulau Komodo dan Pulau Rinca di Website Tripadvisor Berbasis CRISP-DM," *J. Inf. Syst. Res.*, vol. 4, no. 2, pp. 614–625, 2023, doi: 10.47065/josh.v4i2.2999.

[32] Y. A. Singgalen, "Sentiment Classification of Over-Tourism Issues in Responsible Tourism Content using Naïve Bayes Classifier," *J. Comput. Syst. Informatics*, vol. 5, no. 2, pp. 275–285, 2024, doi: 10.47065/josyc.v5i2.4904.

[33] S. A. Azzahra and A. Wibowo, "Analisis Sentimen Multi-Aspek Berbasis Konversi Ikon Emosi dengan Algoritme Naïve Bayes untuk Ulasan Wisata Kuliner Pada Web Tripadvisor," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 4, pp. 737–743, 2020, doi: 10.25126/jtiik.2020731907.

[34] M. S. Rahman and H. Reza, "A Systematic Review Towards Big Data Analytics in Social Media," *Big Data Min. Anal.*, vol. 5, no. 3, pp. 228–244, 2022, doi: 10.26599/BDMA.2022.9020009.

[35] Y. Du, Y. Liu, Y. Yan, J. Fang, and X. Jiang, "Risk Management of Weather-Related Failures in Distribution Systems Based on Interpretable Extra-Trees," *J. Mod. Power Syst. Clean Energy*, vol. 11, no. 6, pp. 1868–1877, 2023, doi: 10.35833/MPCE.2022.000430.