International Journal of Basic and Applied Science 12 (3) (2023) 110-120

Published by: IOCSCIENCE

International Journal of Basic and Applied Science

Journal homepage: www.ijobas.pelnus.ac.id

Culture and heritage tourism sentiment classification through cross-industry standard process for data mining

Yerik Afrianto Singgalen

Tourism Department, Faculty of Business Administration and Communication, Atma Jaya Catholic University of Indonesia, Indonesia

Article Info

Article history:

Received Oct 9, 2023
Revised Nov 20, 2023
Accepted Dec 30, 2023

Keywords:

Culture and Heritage Tourism;
Machine learning algorithms;
Sentiment classification;
Support Vector Machine (SVM).

ABSTRACT

This study investigates the efficacy of machine learning algorithms in sentiment classification within the context of Culture and Heritage Tourism content analysis. This study adopts the CRISP-DM method, a comprehensive methodology encompassing distinct stages, including business understanding, data understanding, modeling, evaluation, and deployment. The k-nearest Neighbors, Decision Tree, Naive Bayes Classifier, and Support Vector Machine models are used. The performance of each model is scrutinized through confusion matrix analysis, encompassing metrics such as accuracy, precision, recall, and F-measure. Additionally, the impact of the Synthetic Minority Over-sampling Technique (SMOTE) implementation on addressing data imbalance is assessed. Leveraging data from the national geographic channel's youtube platform, with a focus on ma'nene content, results reveal SVM's consistent superiority, particularly with SMOTE integration, showcasing elevated accuracy (77.89%), precision (72.60%), recall (89.62%), and F-measure (80.21%) values. These findings underscore the importance of algorithm selection and data preprocessing methods in enhancing sentiment classification accuracy for culture and heritage tourism content, thus contributing quantifiable insights to the tourism research domain.

Corresponding Author:

Yerik Afrianto Singgalen,
Tourism Study Program, Faculty of Business Administration and Communication,
Atma Jaya Catholic University of Indonesia,
Jl. Jenderal Sudirman No. 51, Jakarta, 10220, Indonesia.
Email: yerik.afrianto@atmajaya.ac.id

1. INTRODUCTION

Heritage tourism plays a crucial role in preserving the socio-cultural values of communities [1]. By attracting visitors to historical sites, traditional villages, and cultural landmarks, heritage tourism serves as a primary vehicle for conserving and promoting cultural heritage [2]. This influx of tourists fosters increased awareness and appreciation for local customs, rituals, and traditions [3]. Moreover, the economic benefits derived from heritage tourism often translate into enhanced infrastructure, improved facilities, and increased funding for heritage conservation projects, further reinforcing the preservation efforts [4]. However, balancing tourism development and heritage preservation is imperative to prevent over-commercialization and cultural commodification [5]. In conclusion, heritage tourism is vital in safeguarding communities' social and cultural fabric while bolstering economic development [6].

In the era of digital communication, one of the primary challenges facing Culture and Heritage Tourism is the management of user sentiments and tourists' perceptions regarding collective cultural actions within communities [7]. The proliferation of digital platforms has facilitated instantaneous sharing of opinions and experiences, amplifying the impact of user sentiment on tourism perceptions and decision-making processes [8]. Consequently, cultural heritage sites and events are subject to scrutiny and evaluation by diverse individuals, influencing their attractiveness and reception among potential visitors [9]. Addressing these challenges necessitates proactive strategies for engaging with digital communities, fostering positive sentiment, and leveraging digital platforms to showcase heritage tourism destinations' cultural richness and authenticity [10]. In doing so, stakeholders can mitigate negative perceptions and capitalize on the opportunities afforded by digital communication to promote sustainable cultural tourism practices.

The urgency of this research lies in its potential to address critical gaps in the understanding and management of sentiment dynamics within the realm of culture and heritage Tourism [11]. As the digital landscape continues to evolve, with social media platforms and online communities playing an increasingly influential role in shaping tourist perceptions and behaviors, there is a pressing need to elucidate the intricate interplay between user sentiments and cultural tourism experiences [12]. By investigating these dynamics comprehensively, this research has the opportunity to yield valuable insights that can inform strategic interventions aimed at enhancing the sustainability, resilience, and appeal of cultural heritage destinations [13]. Furthermore, the timely exploration of sentiment trends and patterns in the digital sphere can equip tourism stakeholders with actionable knowledge to adapt their marketing strategies, engage with digital audiences effectively, and foster positive perceptions of cultural heritage assets [14]. Thus, the significance of this research extends beyond academic inquiry, holding implications for practical applications and policy development in cultural tourism management.

This research's practical and theoretical contribution is substantial, as it not only advances our empirical understanding of sentiment dynamics in the domain of culture and heritage tourism but also provides actionable insights for tourism management and policy formulation [15]. Through comprehensive analysis of user sentiments and perceptions within digital communication channels, this study contributes to developing practical strategies for enhancing destination marketing, visitor engagement, and cultural heritage preservation efforts [16]. Furthermore, by elucidating the theoretical underpinnings of sentiment analysis in the context of cultural tourism, this research enriches scholarly discourse [17]. It lays the groundwork for future investigations into the complex interplay between digital communication, cultural representation, and tourist behavior [18]. Overall, the findings of this study offer valuable guidance for stakeholders in the tourism industry seeking to leverage digital platforms effectively and sustainably to promote cultural heritage assets to diverse audiences [19].

The limitation and opportunity for further research in this study are noteworthy, as they delineate both the boundaries of the current investigation and the potential avenues for expanding scholarly inquiry in the field of sentiment analysis in Culture and Heritage Tourism. While the present study provides valuable insights into user sentiments within digital communication channels, it is constrained by data availability, sample size limitations, and the dynamic nature of online discourse. However, these limitations also serve as catalysts for future research endeavors, presenting opportunities to explore emerging trends, refine analytical methodologies, and delve deeper into specific aspects of sentiment dynamics, such as the influence of cultural context, linguistic nuances, and user demographics on tourism perceptions. By addressing these challenges and capitalizing on emerging opportunities, future research can advance our understanding of sentiment analysis in cultural tourism and contribute to developing more robust, nuanced frameworks for managing and promoting cultural heritage destinations in the digital age.

2. RESEARCH METHOD

This study adopts the CRISP-DM method, a comprehensive methodology encompassing distinct stages, including business understanding, data understanding, modeling, evaluation, and deployment. CRISP-DM offers a structured and systematic framework guiding researchers through the various phases of the data mining process [20]. Commencing with the business understanding stage, researchers gain crucial insights into project objectives and requirements [21]. Subsequently, the data understanding stage involves thoroughly exploring and familiarizing the dataset [22]. The modeling stage encompasses constructing and refining predictive models, followed by the evaluation stage, which assesses the models' performance [23]. Finally, in the deployment stage, insights derived from the analysis are applied to real-world scenarios [24]. CRISP-DM provides researchers with a robust framework for effectively managing and executing data mining projects, facilitating valuable insights and knowledge generation [25].

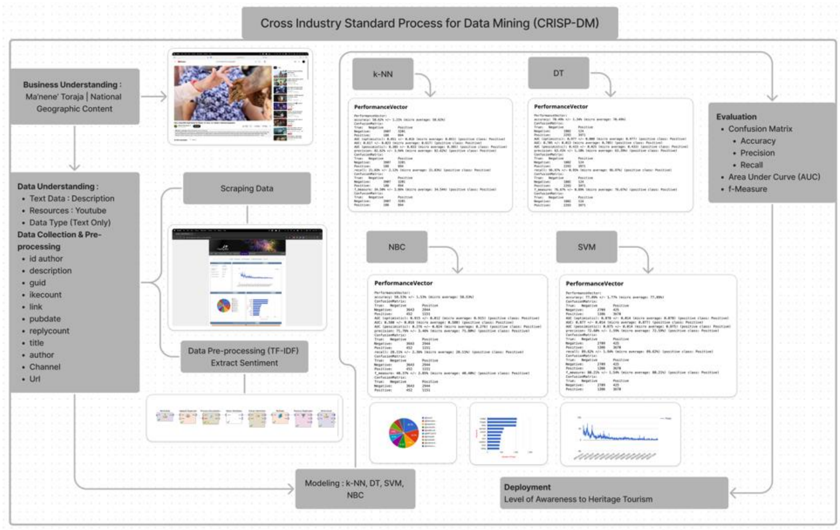

Figure 1. Implementation of CRISP-DM

In the modeling stage, the algorithms employed include k-nearest Neighbors (k-NN), Decision Tree (DT), Naive Bayes Classifier (NBC), and Support Vector Machine (SVM), with a division of the dataset into training data (20%) and testing data (80%). This stage involves implementing and evaluating various machine learning algorithms to build predictive models based on the available data [26]. By utilizing a diverse set of algorithms and appropriately splitting the dataset into training and testing subsets, researchers can assess the performance and effectiveness of each algorithm in classifying and predicting outcomes [27]. The modeling stage is a crucial component in data mining, enabling researchers to identify the most suitable algorithm for the given dataset and research objectives, thus facilitating the generation of accurate and reliable insights [28].

The Synthetic Minority Over-sampling Technique (SMOTE) operator is utilized as a solution to address the data imbalance [29]. Data imbalance, characterized by unequal distribution of classes within a dataset, can pose challenges in machine learning tasks, particularly in classification. SMOTE is a widely employed method designed to mitigate the effects of data imbalance by generating synthetic samples for the minority class [30]. By artificially increasing the representation of the minority class, SMOTE aims to create a more balanced dataset, thereby improving the performance and robustness of machine learning models [31]. This approach enhances the model's ability to accurately classify instances from majority and minority classes, ultimately leading to more reliable and unbiased predictions. Overall, integrating the SMOTE operator is a valuable strategy for addressing data imbalance and optimizing the performance of machine learning algorithms.

The k-NN model possesses its advantages in sentiment classification. As a non-parametric and instance-based learning algorithm, k-NN classifies instances based on their similarity to neighboring instances in the feature space [32]. This approach allows the k-NN model to capture complex patterns and nuances in the data. It is particularly effective in sentiment analysis tasks where the relationships between features and sentiments may be nonlinear or context-dependent [33]. Additionally, k-NN's simplicity and ease of implementation make it a practical choice for sentiment classification applications, especially in scenarios where interpretability and flexibility are valued. In conclusion, the k-NN model stands out as a promising approach for sentiment classification, offering a blend of versatility, effectiveness, and interpretability in capturing sentiment patterns within textual data. The formula for the k-Nearest Neighbors (k-NN) algorithm is as follows:

$$[\hat{y}(x) = \frac{1}{k}\sum_{i \in N_k(x)} y_i] \tag{1}$$

Where:
$(\hat{y}(x))$is the predicted target for data point x
$N_k(x)$is the set of (k) data points nearest to x based on a specified distance function.
y_i is the target value of the (i^{th})data point included in the set $(N_k(x))$.

The Decision Tree (DT) model possesses distinct advantages in sentiment classification tasks. DT operates by recursively partitioning the feature space based on the most informative attributes, resulting in a hierarchical structure that can effectively capture complex decision boundaries [34]. This inherent interpretability of DT allows for easy visualization and understanding of the classification process, making it particularly advantageous in sentiment analysis, where interpretability is crucial for understanding the reasoning behind classification decisions [35]. Furthermore, DT is robust to noisy data and can handle numerical and categorical features, enhancing its applicability in diverse sentiment classification scenarios.

$$[H(T) = -\sum_{i=1}^{c} p(i|t) \log_2 p\ (i|t)] \tag{2}$$

Where:
$(H(T))$is the total entropy of decision tree (T).
$(p(i|t))$ is the probability that a tuple generated by node (t) belongs to class label (i).
(c) is the number of possible classes.

The Naive Bayes Classifier (NBC) model demonstrates robust performance in sentiment classification endeavors [36]. Leveraging probabilistic principles, particularly Bayes' theorem, NBC categorizes instances based on the conditional probability of each class given the input features. The NBC often yields commendable results, particularly in text classification tasks like sentiment analysis [37]. Its simplicity, computational efficiency, and adeptness at managing large datasets render it a compelling option for sentiment classification tasks, especially in resource-constrained environments. Additionally, NBC exhibits resilience to irrelevant features and adeptly manages missing data, further bolstering its applicability in practical sentiment analysis scenarios. The equation for NBC in sentiment classification can be represented as follows:

$$[\text{Sentiment} = \arg\max_{c \in C} P\ (c) \cdot \prod_{i=1}^{n} P(w_i|c)] \tag{3}$$

Where:
(Sentiment) is the predicted sentiment of the text.
c presents the sentiment classes
C is the set of sentiment classes.
$(P(c))$is the prior probability of class (c), representing the probability of a text belonging to sentiment class (c).
(n) is the number of words in the text.
(w_i) represents each word in the text.
$(P(w_i|c))$is the conditional probability of observing word (w_i) given the sentiment class (c), often calculated using the frequency or probability of (w_i) occurring in texts of class (c).

The Support Vector Machine (SVM) model exhibits distinct advantages in sentiment classification tasks. SVM operates by finding the optimal hyperplane that separates instances into different classes, maximizing the margin between the classes [38]. This margin maximization principle allows SVM to effectively capture complex decision boundaries and achieve high classification accuracy, particularly in scenarios with nonlinear relationships between features and sentiments [39]. Furthermore, SVM's ability to handle high-dimensional data and its flexibility in incorporating different kernel functions make it suitable for various sentiment analysis applications, including text classification [40]. Additionally, SVM is robust to overfitting and performs well even with small training datasets, enhancing its utility in sentiment classification tasks. In conclusion, the SVM model is a powerful tool for sentiment analysis, offering a blend of accuracy, flexibility, and robustness in capturing sentiment patterns within textual data. The equation of a support vector machine (SVM) is shown as the equation below:

$$[f(x) = \text{sign}(\textstyle\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b)] \quad (4)$$

Where:
$(f(x))$ is the decision function.
(x_i) are the training samples.
(y_i) are the class labels.
$(K(x_i, x))$ is the kernel function that measures the similarity between the input samples.
(α_i) are the coefficients obtained during training.
(b) is the bias term.

Based on the equations of the k-nearest Neighbors (k-NN), Decision Tree (DT), Naive Bayes Classifier (NBC), and Support Vector Machine (SVM) models, comprehensive analysis can be conducted regarding the confusion matrix of each model, as well as the values of Area Under Curve (AUC) and f-measure based on the context of sentiment classification of Culture and Heritage Tourism content through the case of ma'nene content published by National Geographic Channel on the YouTube platform. These metrics provide valuable insights into the performance of each model in accurately classifying sentiments expressed in cultural and heritage tourism content, facilitating a deeper understanding of their effectiveness and suitability for sentiment analysis tasks in this specific domain. Ultimately, such analysis contributes to advancing sentiment classification techniques tailored to the unique characteristics of Culture and Heritage Tourism content, enhancing the quality and relevance of sentiment analysis outcomes in this context.

3. RESULTS AND DISCUSSIONS

Culture and heritage tourism extends beyond mere tourist attractions to encompass the social identity of a community, as evidenced by the tradition of ma'nene' in Toraja, Central Sulawesi, Indonesia. This tradition, which involves the ritualistic cleaning and dressing of deceased ancestors' bodies, not only serves as a cultural spectacle for visitors but also holds deep significance as a manifestation of the Torajan people's social and spiritual beliefs. The ma'nene' ceremony underscores the intimate connection between culture, heritage, and community identity, highlighting the importance of preserving and promoting such traditions in the context of tourism. In essence, culture and heritage tourism plays a pivotal role in showcasing and safeguarding communities' rich cultural heritage, fostering a deeper understanding and appreciation of their social fabric.

In the National Geographic Channel on the YouTube platform, there exists a documentary video showcasing the ma'nene' tradition of the Toraja people, which has garnered public opinion based on individual perceptions regarding culture and heritage tourism. The narration reflecting public opinions needs to be classified based on sentiment to guide the creation of video content related to culture and heritage tourism. This sentiment analysis allows content creators to understand better the audience's reactions and preferences towards cultural heritage presentations, thereby enabling the production of more engaging and relevant videos that resonate with viewers. Ultimately, such classification can enhance the effectiveness and impact of cultural heritage tourism content on platforms like YouTube, fostering greater appreciation and understanding of diverse cultural traditions among global audiences.

This research gathers data utilizing a data mining approach, employing the Netlytic platform based on the video code (hCKDsjLt_qU). Data mining is a valuable methodological tool for extracting meaningful patterns and insights from large datasets, enabling researchers to uncover hidden trends and relationships within the data. By leveraging Netlytic's capabilities, the research benefits from an efficient and systematic approach to data collection and analysis, enhancing the rigor and depth of the study's findings. In conclusion, using data mining techniques and platforms such as Netlytic provides researchers with valuable resources for conducting comprehensive and insightful investigations across various domains.

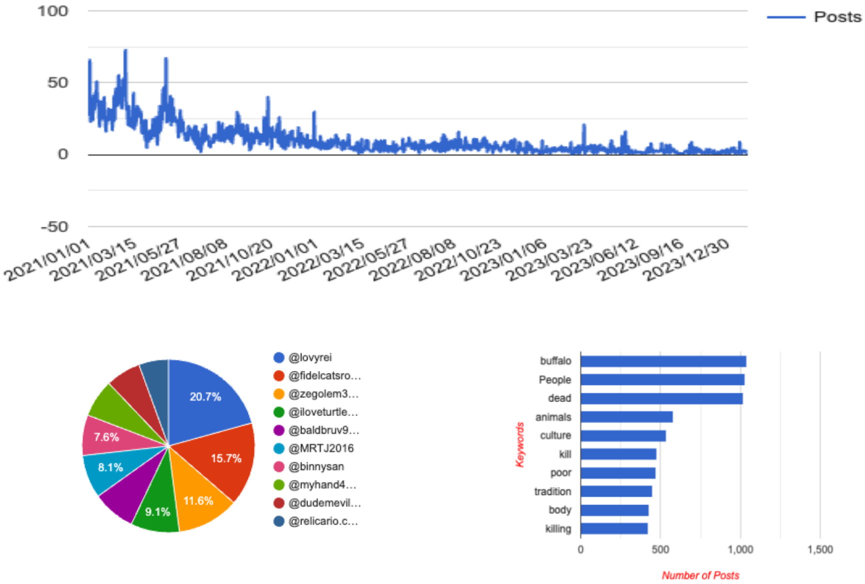

Figure 2. Post Over Time, Top Ten User, and Most Frequently Used Words

This research employs various data preprocessing techniques facilitated by operators, such as tokenizing, transforming cases, filtering tokens, and stopping word removal in English to clean the dataset and remove duplicate reviews. The instrument utilized for these tasks is the RapidMiner application. These preprocessing steps are essential in ensuring the quality and integrity of the dataset by standardizing text formatting, removing irrelevant tokens, and eliminating duplicate entries. By leveraging RapidMiner's functionalities, the research benefits from an efficient and systematic data-cleaning approach, laying a solid foundation for subsequent analysis and interpretation. In conclusion, using these preprocessing techniques enhances the reliability and validity of the research findings by ensuring the dataset's cleanliness and coherence.

Table 1. Extract Sentiment Result

Review	Score String	Total Score	Class
Disgusting terrible not about the dead but the way you kill the animals in a cruel manner when there are way less cruel faster ways 👎 👎	disgusting (-0.62) terrible (-0.54) dead (-0.85) kill (-0.95) cruel (-0.72) cruel (-0.72)	-4,38461538461539	Negative
I understand that people have diverse culture and tradition and people have different ways to view things or perceive it, n i respect that, but as an animal lover I can't take that. With all fue respect, its your tradition anyways, i don't have any part of it, I was just an expressor.	respect (0.54) lover (0.72) respect (0.54)	1,79487179487179	Positive

Based on the results of extracting sentiment from 9614 review data, it is apparent that 4495 reviews are classified as unfavorable, while 5119 are classified as positive. This classification is made possible by implementing the score string and total score derived from the extract sentiment operator within the RapidMiner application. These findings underscore the effectiveness of sentiment analysis techniques in discerning the sentiment polarity of large-scale review datasets, thereby providing valuable insights into the prevailing attitudes and opinions expressed within the analyzed data. In conclusion, using sentiment analysis tools such as RapidMiner facilitates the efficient classification and interpretation of sentiment within textual data, thereby enhancing our understanding of public sentiment and opinion.

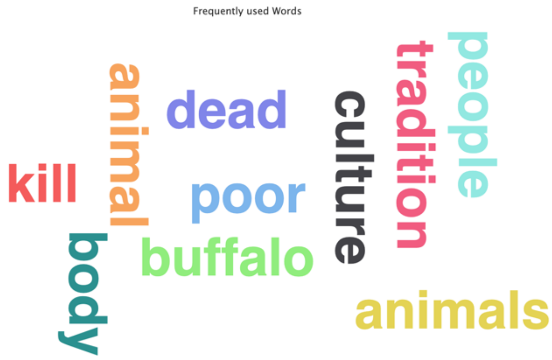

Figure 3. Frequently used Words in Rapidminer

Based on the analysis of frequently used words extracted from processed review data, it is evident that specific terms emerge as popular keywords. Among these, words such as "dead," "buffalo," "people," "animals," and "culture" appear prominently, with high frequencies ranging from 1012 to 1024 occurrences. Additionally, terms like "kill," "body," "animal," "tradition," and "poor" are also prevalent, with frequencies ranging from 415 to 473 occurrences. These findings suggest that the reviews often discuss topics related to death, buffalo, people, animals, and culture, indicating the significance of these themes in the analyzed data. The frequent occurrence of terms such as "kill," "body," and "poor" further suggests discussions around issues related to violence, anatomy, and socio-economic conditions. In conclusion, the identified famous words provide valuable insights into the prevalent topics and concerns expressed within the review data.

Table 2. Algorithm Performance Without SMOTE

Confusion Matrix	k-NN	DT	NBC	SVM
Accuracy	56.10%	71.62%	55.57%	74.82%
Precision	80.50%	65.76%	70.64%	69.63%
Recall	23.17%	97.46%	28.25%	93.58%
F_measure	35.95%	78.53%	40.31%	79.83%
AUC	0.604	0.698	0.565	0.867

The performance assessment of the k-NN algorithm without utilizing the SMOTE unveils specific metrics. As per the provided data, the algorithm exhibits an accuracy rate of 56.10%, indicating the proportion of correctly classified instances among all instances. Additionally, the precision metric stands at 80.50%, signifying the proportion of correctly predicted positive instances out of all instances classified as positive by the algorithm. However, the recall rate is relatively low at 23.17%, suggesting the algorithm's limited ability to identify all relevant instances correctly. Furthermore, the F-measure, amalgamating precision and recall, is computed at 35.95%, indicating a moderate balance between the two metrics. Lastly, the AUC value of 0.604 reflects the algorithm's capability to distinguish between positive and negative instances.

The evaluation of the DT algorithm without employing SMOTE reveals distinct performance metrics. As per the provided data, the algorithm achieves an accuracy rate of 71.62%, signifying the proportion of correctly classified instances among all instances. Moreover, the precision metric is calculated at 65.76%, denoting the proportion of correctly predicted positive instances out of all instances classified as positive by the algorithm. Notably, the recall rate is impressively high at 97.46%, suggesting the algorithm's robust ability to identify all relevant instances correctly. Furthermore, the F-measure, combining precision and recall, is computed at 78.53%, showcasing a balanced performance between the two metrics. Lastly, the AUC value of 0.698 reflects the algorithm's capability to distinguish between positive and negative instances.

The assessment of the NBC algorithm without employing the SMOTE unveils specific performance metrics. As per the provided data, the algorithm exhibits an accuracy rate of 55.57%, indicating the proportion of correctly classified instances among all instances. Additionally, the precision metric stands at 70.64%, signifying the proportion of correctly predicted positive instances out of all instances classified as positive by the algorithm. However, the recall rate is relatively low at 28.25%, suggesting the algorithm's limited ability to identify all relevant instances correctly. Furthermore, the F-measure, amalgamating precision and recall, is computed at 40.31%, indicating a moderate balance between the two metrics. Lastly, the AUC value of 0.565 reflects the algorithm's capability to distinguish between positive and negative instances.

Evaluating the SVM algorithm without employing the SMOTE reveals noteworthy performance metrics. According to the provided data, the algorithm achieves an accuracy rate of 74.82%, indicating the proportion of correctly classified instances among all instances. Moreover, the precision metric stands at 69.63%, denoting the proportion of correctly predicted positive instances out of all instances classified as positive by the algorithm. The recall rate is notably high at 93.58%, suggesting the algorithm's robust ability to identify all relevant instances correctly. Furthermore, the F-measure, combining precision and recall, is computed at 79.83%, showcasing a balanced performance between the two metrics. Lastly, the AUC value of 0.867 reflects the algorithm's capability to distinguish between positive and negative instances.

Table 3. Algorithm Performance With SMOTE

Confusion Matrix	k-NN	DT	NBC	SVM
Accuracy	58.62%	70.49%	58.53%	77.89%
Precision	82.62%	63.41%	71.76%	72.60%
Recall	21.83%	96.97%	28.11%	89.62%
F_measure	34.50%	76.67%	40.37%	80.21%
AUC	0.617	0.705	0.580	0.877

The performance evaluation of the k-NN algorithm utilizing the SMOTE reveals specific metrics. As per the provided data, the accuracy of the k-NN algorithm with SMOTE stands at 58.62%, indicating the proportion of correctly classified instances among all instances. Moreover, the precision rate of 82.62% illustrates the proportion of correctly predicted positive instances out of all instances classified as positive by the algorithm. However, the recall rate of 21.83% suggests the algorithm's relatively low ability to identify all relevant instances correctly. The F-measure, combining precision and recall into a single metric, is computed at 34.50%. Lastly, the AUC value of 0.617 showcases the algorithm's capability to distinguish between positive and negative instances.

The performance assessment of the DT algorithm integrated with the SMOTE yields notable metrics. According to the provided data, the algorithm achieves an accuracy rate of 70.49%, signifying the proportion of correctly classified instances among all instances. Additionally, the precision metric stands at 63.41%, indicating the proportion of correctly predicted positive instances of all instances classified as positive by the algorithm. Notably, the recall rate is impressively high at 96.97%, suggesting the algorithm's robust ability to identify all relevant instances correctly. Furthermore, combining precision and recall, the F-measure is computed at 76.67%, showcasing a balanced performance. Lastly, the AUC value 0.705 highlights the algorithm's effectiveness in distinguishing between positive and negative instances.

Evaluating the NBC algorithm integrated with the SMOTE reveals distinct performance metrics. According to the provided data, the algorithm demonstrates an accuracy rate of 58.53%, indicating the proportion of correctly classified instances among all instances. Moreover, the precision metric is calculated at 71.76%, denoting the proportion of correctly predicted positive instances out of all instances classified as positive by the algorithm. However, the recall rate appears relatively low at 28.11%, suggesting the algorithm's limited ability to identify all relevant instances correctly. The F-measure, amalgamating precision and recall, is computed at 40.37%, indicating a moderate balance between the two metrics. Lastly, the AUC value of 0.580 reflects the algorithm's capability to distinguish between positive and negative instances.

The performance evaluation of the SVM algorithm integrated with the SMOTE yields compelling results. According to the provided data, the algorithm achieves a remarkable accuracy rate of 77.89%, indicating the proportion of correctly classified instances among all instances. Additionally, the precision metric is notably high at 72.60%, denoting the proportion of correctly predicted positive instances out of all instances classified as positive by the algorithm. Furthermore, the recall rate is impressively high at 89.62%, indicating the algorithm's robust ability to identify all relevant instances correctly. The F-measure, combining precision and recall, is computed at 80.21%, showcasing a balanced performance between the two metrics. Lastly, the AUC value of 0.877 reflects the algorithm's effectiveness in distinguishing between positive and negative instances.

4. CONCLUSION

In comparing the confusion matrix values derived from sentiment classification models, including k-NN, DT, NBC, and SVM, with and without the SMOTE, it is evident that SMOTE generally enhances the models' performance metrics, particularly in terms of recall and F1-score, indicating a more robust ability to capture relevant instances accurately. For instance, the SVM model demonstrates notable improvement with SMOTE, as evidenced by increased accuracy from 74.82% to 77.89%, precision from 69.63% to 72.60%, recall from 93.58% to 89.62%, and F1-score from 79.83% to 80.21%. Similarly, models such as k-NN, DT, and NBC also exhibit enhancements across their performance metrics when applying SMOTE. These findings underscore the efficacy of SMOTE in mitigating data imbalance issues, thereby bolstering sentiment classification accuracy in the domain of Culture and Heritage Tourism content analysis. Further research may explore additional strategies or hybrid approaches to optimize model performance in this specific context.

ACKNOWLEDGEMENTS

I sincerely thank the Tourism Department, the Faculty of Business Administration and Communication at the Atma Jaya Catholic University of Indonesia, and the Research and Community Service Institute (LPPM).

REFERENCES

[1] Y. Zhao and E. Agyeiwaah, “Exploring value-based motivations for culture and heritage tourism using the means-end chain and laddering approach,” *J. Herit. Tour.*, vol. 18, no. 5, pp. 594–616, 2023, doi: 10.1080/1743873X.2023.2215933.

[2] T. D. Quang, W. B. Noseworthy, and D. Paulson, “Rising tensions: heritage-tourism development and the commodification of ‘Authentic’ culture among the Cham community of Vietnam,” *Cogent Soc. Sci.*, vol. 8, no. 1, pp. 1–23, 2022, doi: 10.1080/23311886.2022.2116161.

[3] R. Luekveerawattana, “Enhancing innovation in cultural heritage tourism: navigating external factors,” *Cogent Soc. Sci.*, vol. 10, no. 1, pp. 1–15, 2024, doi: 10.1080/23311886.2024.2301813.

[4] S. Jiang, B. Moyle, R. Yung, L. Tao, and N. Scott, “Augmented reality and the enhancement of memorable tourism experiences at heritage sites,” *Curr. Issues Tour.*, vol. 26, no. 2, pp. 242–257, 2023, doi: 10.1080/13683500.2022.2026303.

[5] W. Luo, Y. Lu, D. J. Timothy, and X. Zang, “Tourism and conserving intangible cultural heritage: Residents’ perspectives on protecting the nüshu female script,” *J. China Tour. Res.*, vol. 18, no. 6, pp. 1305–1329, 2022, doi: 10.1080/19388160.2022.2036663.

[6] E. Rindrasih and P. Witte, “Reinventing the post-disaster cultural landscape of heritage tourism in Kotagede, Yogyakarta, Indonesia,” *J. Herit. Tour.*, vol. 16, no. 2, pp. 136–150, 2021, doi: 10.1080/1743873X.2020.1781870.

[7] C. Dolezal and M. Novelli, “Power in community-based tourism: empowerment and partnership in Bali,” *J. Sustain. Tour.*, vol. 30, no. 10, pp. 2352–2370, 2022, doi: 10.1080/09669582.2020.1838527.

[8] D. Zheng, C. Huang, and B. Oraltay, "Digital cultural tourism: progress and a proposed framework for future research," *Asia Pacific J. Tour. Res.*, vol. 28, no. 3, pp. 234–253, 2023, doi: 10.1080/10941665.2023.2217958.

[9] M. H. Sumiaty, C. E. Randalele, R. Iye, and F. I. Nur Abida, "The value of Tallu Lolona and its influence to the life of Toraja people," *Cogent Soc. Sci.*, vol. 9, no. 2, pp. 1–18, 2023, doi: 10.1080/23311886.2023.2262775.

[10] R. A. Kinseng, A. Kartikasari, N. Aini, R. Gandi, and D. Dean, "COVID-19 and the emergence of virtual tourism in Indonesia: A sociological perspective," *Cogent Soc. Sci.*, vol. 8, no. 1, pp. 1–15, 2022, doi: 10.1080/23311886.2022.2026557.

[11] F. Mehraliyev, I. C. C. Chan, and A. P. Kirilenko, "Sentiment analysis in hospitality and tourism: a thematic and methodological review," *Int. J. Contemp. Hosp. Manag.*, vol. 34, no. 1, pp. 46–77, 2022, doi: 10.1108/IJCHM-02-2021-0132.

[12] A. Galiano-Coronil, S. Blanco-Moreno, L. B. Tobar-Pesantez, and G. A. Gutiérrez-Montoya, "Social media impact of tourism managers: a decision tree approach in happiness, social marketing and sustainability," *J. Manag. Dev.*, vol. 42, no. 6, pp. 436–457, 2023, doi: 10.1108/JMD-04-2023-0131.

[13] F. Mirzaalian and E. Halpenny, "Social media analytics in hospitality and tourism: A systematic literature review and future trends," *J. Hosp. Tour. Technol.*, vol. 10, no. 4, pp. 764–790, 2019, doi: 10.1108/JHTT-08-2018-0078.

[14] C. Zhu, M. U. Io, H. F. B. Ngan, and R. L. Peralta, "Interpreting the impact of augmented reality on heritage tourism: two empirical studies from World Heritage sites," *Curr. Issues Tour.*, pp. 1–15, 2024, doi: 10.1080/13683500.2023.2298349.

[15] E. Mele and L. Egberts, "Exploring travel blogs on tourism and landscape heritage: representations of the Swiss Alps," *J. Herit. Tour.*, vol. 18, no. 6, pp. 785–806, 2023, doi: 10.1080/1743873X.2023.2237617.

[16] A. Hausmann and S. Schuhbauer, "The role of information and communication technologies in cultural tourists' journeys: the case of a World Heritage Site," *J. Herit. Tour.*, vol. 16, no. 6, pp. 669–683, 2021, doi: 10.1080/1743873X.2020.1819300.

[17] C. E. Yu and X. Zhang, "The embedded feelings in local gastronomy: a sentiment analysis of online reviews," *J. Hosp. Tour. Technol.*, vol. 11, no. 3, pp. 461–478, 2020, doi: 10.1108/JHTT-02-2019-0028.

[18] A. Tuomi, E. Moreira Kares, and H. Zainal Abidin, "Digital cultural tourism: older adults' acceptance and use of digital cultural tourism services," *Scand. J. Hosp. Tour.*, vol. 23, no. 2–3, pp. 226–247, 2023, doi: 10.1080/15022250.2023.2256698.

[19] E. Gatelier, D. Ross, L. Phillips, and J. B. Suquet, "A business model innovation methodology for implementing digital interpretation experiences in European cultural heritage attractions," *J. Herit. Tour.*, vol. 17, no. 4, pp. 391–408, 2022, doi: 10.1080/1743873X.2022.2065920.

[20] Y. A. Singgalen, "Penerapan Metode CRISP-DM untuk Optimalisasi Strategi Pemasaran STP (Segmenting , Targeting , Positioning) Layanan Akomodasi Hotel , Homestay , dan Resort," *J. Media Inform. Budidarma*, vol. 7, no. 4, pp. 1980–1993, 2023, doi: 10.30865/mib.v7i4.6896.

[21] Y. A. Singgalen, "Penerapan Metode CRISP-DM dalam Klasifikasi Data Ulasan Pengunjung Destinasi Danau Toba Menggunakan Algoritma Naïve Bayes Classifier (NBC) dan Decision Tree (DT)," *J. Media Inform. Budidarma*, vol. 7, no. 3, pp. 1551–1562, 2023, doi: 10.30865/mib.v7i3.6461.

[22] 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.

[23] Y. A. Singgalen, "Penerapan CRISP-DM dalam Klasifikasi Sentimen dan Analisis Perilaku Pembelian Layanan Akomodasi Hotel Berbasis Algoritma Decision Tree (DT)," *J. Sist. Komput. dan Inform.*, vol. 5, no. 2, pp. 237–248, 2023, doi: 10.30865/json.v5i2.7081.

[24] Y. A. Singgalen, "Analisis Perilaku Wisatawan Berdasarkan Data Ulasan di Website Tripadvisor Menggunakan CRISP-DM : Wisata Minat Khusus Pendakian Gunung Rinjani dan Gunung Bromo," *J. Comput. Syst. Informatics*, vol. 4, no. 2, pp. 326–338, 2023, doi: 10.47065/josyc.v4i2.3042.

[25] Y. A. Singgalen, "Analisis Sentimen Wisatawan terhadap Kualitas Layanan Hotel dan Resort di Lombok Menggunakan SERVQUAL dan CRISP-DM," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1870–1882, 2023, doi: 10.47065/bits.v4i4.3199.

[26] Y. A. Singgalen, "Analisis Sentimen Konsumen terhadap Food , Services , and Value di Restoran dan Rumah Makan Populer Kota Makassar Berdasarkan Rekomendasi Tripadvisor Menggunakan Metode CRISP-DM dan," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1899–1914, 2023, doi: 10.47065/bits.v4i4.3231.

[27] Y. A. Singgalen, "Analisis Sentimen Wisatawan terhadap Taman Nasional Bunaken dan Top 10 Hotel Rekomendasi Tripadvisor Menggunakan Algoritma SVM dan DT berbasis CRISP-DM," *J. Comput. Syst. Informatics*, vol. 4, no. 2, pp. 367–379, 2023, doi: 10.47065/josyc.v4i2.3092.

[28] Y. A. Singgalen, “Analisis Performa Algoritma NBC , DT , SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM,” *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1634–1646, 2022, doi: 10.47065/bits.v4i3.2766.

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

[30] A. Andreyestha and Q. N. Azizah, “Analisa Sentimen Kicauan Twitter Tokopedia Dengan Optimalisasi Data Tidak Seimbang Menggunakan Algoritma SMOTE,” *Infotek J. Inform. dan Teknol.*, vol. 5, no. 1, pp. 108–116, 2022, doi: 10.29408/jit.v5i1.4581.

[31] M. I. Putri and I. Kharisudin, “Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Analisis Sentimen Data Review Pengguna Aplikasi Marketplace Tokopedia,” in *PRISMA, Prosiding Seminar Nasional Matematika*, 2022, vol. 5, pp. 759–766. [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[32] N. Ginantra, C. P. Yanti, and ..., “Analisis Sentimen Ulasan Villa di Ubud Menggunakan Metode Naive Bayes, Decision Tree, dan K-NN,” *J. Nas. ...*, 2022, [Online]. Available: https://ejournal.undiksha.ac.id/index.php/janapati/article/view/49450

[33] H. J. Christanto and Y. A. Singgalen, “Sentiment Analysis on Customer Perception towards Products and Services of Restaurant in Labuan Bajo,” *J. Inf. Syst. Informatics*, vol. 4, no. 3, pp. 511–523, 2022, doi: 10.51519/journalisi.v4i3.276.

[34] Rousyati, W. Gata, D. Pratmanto, and N. K. Warchani, “Analisis Sentimen Financial Technology Peer to Peer Lending Pada Aplikasi Koinworks,” *J. Teknol. Infor*, vol. 9, no. 6, pp. 1167–1176, 2022, doi: 10.25126/jtiik.202294409.

[35] B. Noori, “Classification of Customer Reviews Using Machine Learning Algorithms,” *Appl. Artif. Intell.*, vol. 35, no. 8, pp. 567–588, 2021, doi: 10.1080/08839514.2021.1922843.

[36] G. Cahyani, W. Widayani, S. D. Anggita, and ..., “Klasifikasi Data Review IMDb Berdasarkan Analisis Sentimen Menggunakan Algoritma Support Vector Machine,” *J. MEDIA ...*, 2022, [Online]. Available: http://stmik-budidarma.ac.id/ejurnal/index.php/mib/article/view/4023

[37] Q. A’yuniyah *et al.*, “Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik,” *J. Sist. Komput. dan Inform.*, vol. 4, no. 1, pp. 72–76, 2022, doi: 10.30865/json.v4i1.4781.

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

[39] H. Syah and A. Witanti, “Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (Svm),” *J. Sist. Inf. Dan ...*, 2022, [Online]. Available: http://ejournal.lppm-unbaja.ac.id/index.php/jsii/article/view/1411

[40] F. Z. Ruskanda, M. R. Abiwardani, R. Mulyawan, I. Syafalni, and H. T. Larasati, “Quantum-Enhanced Support Vector Machine for Sentiment Classification,” *IEEE Access*, vol. 11, no. July, pp. 87520–87532, 2023, doi: 10.1109/ACCESS.2023.3304990.