Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

Sentiment Classification of Over-Tourism Issues in Responsible Tourism Content using Naïve Bayes Classifier

Yerik Afrianto Singgalen

Faculty of Business Administration and Communication, Tourism Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Email: yerik.afrianto@atmajaya.ac.id
Submitted: 03/02/2024; Accepted: 20/02/2024; Published: 20/02/2024

Abstract−The research problem addressed in this study is the analysis of public sentiment regarding over-tourism issues. Utilizing the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology and the Naive Bayes Classifier (NBC) algorithm, the study navigates through stages of business understanding, data processing, modeling, evaluation, and deployment. The central focus lies in understanding and classifying public sentiments surrounding the challenges associated with over-tourism. The findings reveal that the NBC algorithm, particularly when augmented with Synthetic Minority Over-sampling Technique (SMOTE), demonstrates superior performance metrics, showcasing an accuracy of 84.82%, precision of 91.69%, recall of 76.75%, f-measure of 83.47%, and AUC of 0.838. The comparison with NBC without SMOTE, which registers an accuracy of 78.16%, precision of 87.61%, recall of 74.56%, f-measure of 80.51%, and AUC of 0.745, underscores the significance of addressing class imbalance for improved predictive performance. Integrating CRISP-DM with the NBC algorithm and SMOTE proves instrumental in advancing sentiment analysis methodologies, providing nuanced insights into public perceptions and attitudes concerning the critical issue of over-tourism.

Keywords: Sentiment; Classification; Over-Tourism; Naïve Bayes Classifier; Responsible Tourism

1. INTRODUCTION

The sentiment classification of over-tourism issues in responsible tourism content represents a pertinent approach to comprehending public understanding and responses to information dissemination through digital media [1]. With the exponential growth of digital platforms, responsible tourism discourse has become increasingly prevalent, necessitating effective strategies for gauging public sentiment [2]. Through sentiment classification, researchers can discern prevailing attitudes towards over-tourism issues, refining communication strategies tailored to diverse audience perceptions [3]. This analytical framework not only facilitates the identification of critical concerns but also enables the formulation of targeted interventions to mitigate negative sentiments and foster constructive dialogue [4]. Consequently, applying sentiment classification in the context of responsible tourism content serves as a vital tool for enhancing public engagement and promoting sustainable tourism practices.

The utilization of digital media as a means of information mobility incites various responses and understandings among users of applications concerning sustainability issues in tourism. In today's digitally connected world, disseminating information through digital platforms has become instrumental in shaping public discourse on sustainability within the tourism sector [5]. Through the proliferation of applications tailored to tourism, users are exposed to various information regarding sustainability practices, ranging from environmental conservation to community engagement [6]. Consequently, users' responses and interpretations of such information are diverse and multifaceted, influenced by personal values, cultural backgrounds, and socioeconomic status [7]. Despite the potential for digital media to facilitate informed decision-making and promote sustainable tourism behaviors, the subjective nature of user responses underscores the need for tailored communication strategies to convey sustainability messages effectively [8]. In conclusion, while digital media is a powerful tool for fostering awareness and engagement with sustainability issues in tourism, nuanced approaches are required to navigate the complexities of user perceptions and foster meaningful dialogue toward sustainable practices.

The issue of over-tourism has emerged as a significant concern due to the disproportionate mobility of tourist access juxtaposed with the carrying capacity of destination sites, eliciting a myriad of comments and sentiments from both local communities and travelers seeking to derive meaning from their tourism experiences [9]. The rapid growth of tourism has led to overcrowding, environmental degradation, and socio-cultural disruption in popular destinations worldwide, underscoring the urgency of addressing this imbalance [10]. As destinations struggle to cope with the influx of visitors, tensions arise between the aspirations of tourists and the preservation of local heritage and livelihoods [11]. Furthermore, the discourse surrounding over-tourism reflects divergent perspectives and interests, with stakeholders offering nuanced opinions on potential solutions and management strategies [12]. In light of these complexities, effectively mitigating over-tourism necessitates collaborative efforts among stakeholders to ensure sustainable tourism practices prioritizing environmental stewardship, cultural preservation, and equitable economic benefits.

The urgency of this research lies in comprehending public responses to the issue of over-tourism and how campaigns stimulate public awareness regarding the ecological and tourism implications. With the escalating global phenomenon of over-tourism, there is a pressing need to delve into the multifaceted reactions of the public towards this phenomenon, which encompasses environmental degradation, cultural commodification, and

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

socio-economic disparities [13]. By examining public perceptions and attitudes towards over-tourism, this research aims to shed light on the complexities of tourist destination management and sustainable tourism promotion [14]. Furthermore, understanding how campaigns influence public consciousness regarding tourism's ecological and socio-cultural dimensions aids in crafting targeted interventions to foster responsible tourism behaviors and mitigate the adverse impacts of over-tourism [15]. In conclusion, this research is a pivotal step toward advancing sustainable tourism practices and fostering harmonious interactions between tourists, host communities, and natural environments.

The practical implication of this research pertains to exploring the digital media network and the efficacy of digital campaigns in augmenting ecological awareness amidst over-tourism concerns. In the contemporary era, digital media platforms are pivotal tools for disseminating information and shaping public perceptions on socio-environmental issues, including tourism sustainability [16]. By scrutinizing the effectiveness of digital campaigns in heightening ecological awareness within the context of over-tourism, this research offers valuable insights for stakeholders seeking to leverage digital platforms for sustainable tourism advocacy [17]. Additionally, findings from this study can inform the development of targeted digital strategies to foster responsible tourist behaviors and promote conservation initiatives [18]. In conclusion, this research contributes to the burgeoning discourse on sustainable tourism communication strategies, emphasizing the potential of digital media networks in catalyzing positive environmental outcomes amidst the challenges of over-tourism.

The theoretical implication of this research revolves around its contribution to understanding the sustainability of the tourism industry and the pivotal role of tourist behavior in the development process [19]. This study enriches theoretical frameworks concerning sustainable tourism management and policy formulation by delving into the intricate dynamics between tourism sustainability and tourist behavior [20]. Furthermore, it underscores the significance of aligning tourist behaviors with sustainability principles to ensure the long-term viability of tourism destinations. Insights gleaned from this research shed light on the interplay between individual tourist actions and broader socio-economic and environmental factors shaping the trajectory of tourism development [21]. Ultimately, this research underscores the imperative of adopting holistic approaches that integrate both supply-side interventions and demand-side strategies to advance sustainable tourism practices and mitigate the adverse impacts of tourism on host communities and natural environments [22].

This research employs the CRISP-DM methodology as a structured approach to address the research problem. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, provides a systematic framework comprising distinct stages such as business understanding, data preparation, modeling, evaluation, and deployment. By utilizing CRISP-DM, researchers can effectively navigate through the complexities of data analysis, ensuring a comprehensive and structured approach to problem-solving. This methodological choice not only enhances the rigor and reproducibility of the research process but also facilitates the identification of actionable insights and informed decision-making. Therefore, the adoption of CRISP-DM underscores the commitment to methodological excellence and scholarly integrity in conducting this research endeavor.

The exploration of similar research and the prospects of this study reveal a rich landscape of inquiry and potential contributions to the field. Existing research endeavors focusing on sentiment analysis and over-tourism provide valuable insights into public perceptions, sentiments, and behaviors related to sustainable tourism practices. However, the unique contribution of this research lies in its application of the CRISP-DM methodology to analyze sentiment data specifically within the context of responsible tourism channels. By adopting a structured approach, this study aims to offer a comprehensive understanding of public sentiments towards over-tourism issues and their implications for sustainable tourism management. Moving forward, the prospects of this research are promising, as it sets a foundation for further investigations into sentiment analysis methodologies, cross-media comparisons, and the development of targeted interventions to address over-tourism challenges effectively. Through rigorous inquiry and scholarly engagement, this research holds the potential to inform policy-making processes, enhance tourism practices, and contribute to the sustainable development of tourist destinations.

The limitation of this research is identified in the utilization of methodology and algorithm to assess performance, wherein the Cross Industry Standard Process for Data Mining (CRISP-DM) is employed alongside the Naive Bayes Classifier (NBC) [23]–[25]. While CRISP-DM provides a structured framework for data mining projects, and the NBC offers simplicity and efficiency in classification tasks, the application may encounter constraints in capturing the nuanced complexities of sentiment analysis within the context of over-tourism. Despite the utility, these methodologies and algorithms may overlook subtle nuances and contextual factors inherent in public perceptions of over-tourism issues, potentially limiting the comprehensiveness and accuracy of the research findings. Thus, future studies may benefit from incorporating complementary methodologies or hybrid approaches to enhance the robustness and validity of sentiment analysis in understanding public responses to over-tourism.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

2. RESEARCH METHODOLOGY

2.1 Cross-Industry Standard Process for Data Mining (CRISP-DM)

This study employs the CRISP-DM methodology and the NBC algorithm. The CRISP-DM stages encompass business understanding, data processing, modeling, evaluation, and deployment [26]. During the modeling stage, the NBC algorithm is adopted along with the Synthetic Minority Oversampling Technique (SMOTE). This approach enables the systematic exploration of public sentiment regarding over-tourism issues, facilitating a comprehensive understanding of the complexities involved [27]. By integrating CRISP-DM and NBC, supplemented by SMOTE, this research methodology offers a robust framework for analyzing sentiment data and deriving meaningful insights [28]. In conclusion, using CRISP-DM and NBC represents a methodologically sound approach to investigating public perceptions of over-tourism, contributing to advancing sentiment analysis methodologies in tourism research.

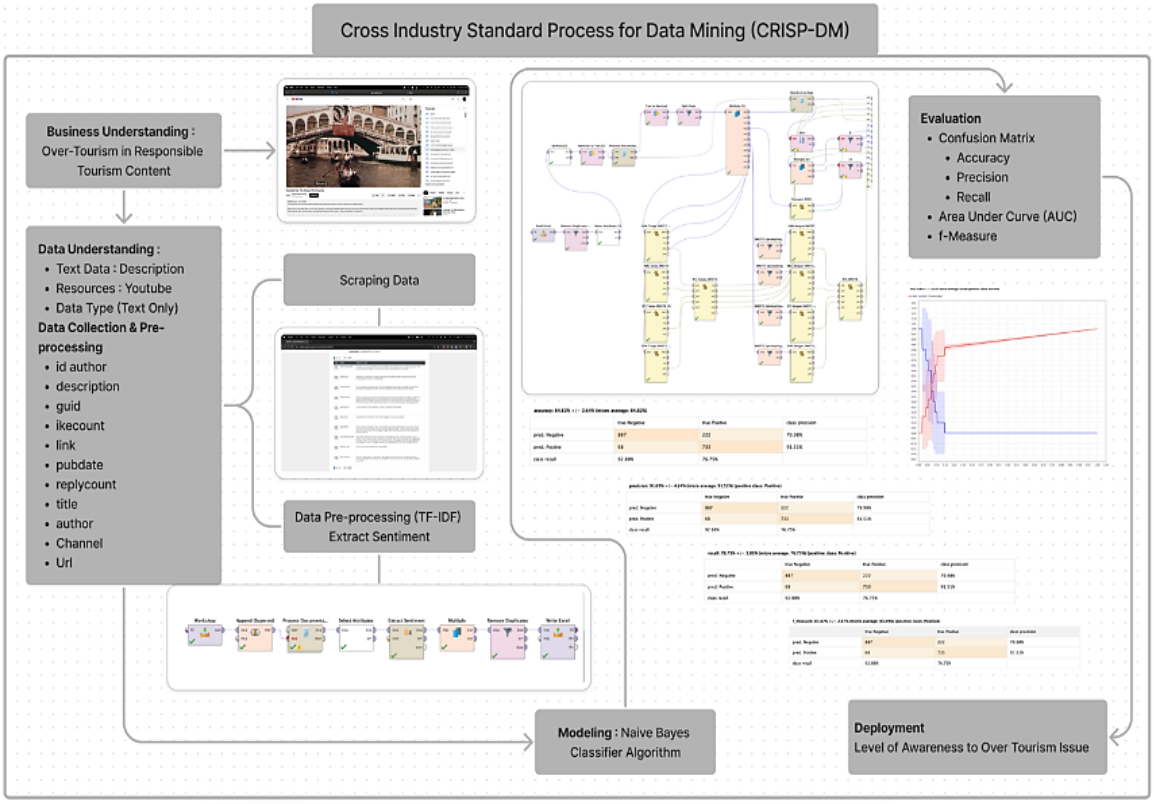


Figure 1. Implementaiton of CRISP-DM and NBC Algorithm

The CRISP-DM methodology and NBC algorithm are highly pertinent to the context of the dataset utilized in this research, where data obtained from the Netlytic website are extracted via the RapidMiner application and subsequently classified using the NBC algorithm and SMOTE to address data imbalance while also measuring algorithm performance based on accuracy, precision, recall, f-measure, and AUC values. This methodological framework offers a structured approach to analyzing sentiment data related to over-tourism issues, ensuring rigorous data processing and modeling procedures. Moreover, the integration of NBC and SMOTE enables the mitigation of class imbalance, enhancing the reliability and robustness of sentiment analysis outcomes. Applying the CRISP-DM and NBC algorithms in this study provides a comprehensive and systematic framework for investigating public perceptions of over-tourism, contributing valuable insights to tourism research.

During the data understanding stage, textual data was obtained from the Youtube platform with the ID U-52L7hYQiE, encompassing 1334 public comments on the video "Crowded Out: The Story of Overtourism." This stage involved the preliminary exploration of the raw data, aiming to comprehend the nature and characteristics of the comments to inform subsequent analytical processes. Using specific video content and associated comments from YouTube provides a rich source of information, offering a glimpse into public perceptions and sentiments surrounding the discussed topic of over-tourism. The selection of this video aligns with the research's focus on responsible tourism channels and adds a contextual layer to the sentiment analysis process. This strategic approach ensures that the data obtained is relevant and aligned with the research objectives, contributing to a more nuanced understanding of public sentiments towards over-tourism issues within the digital realm.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

In the modeling stage, textual data undergoes extraction and classification using the Naive Bayes Classifier algorithm, leveraging the SMOTE operator to address data imbalance. The central aspect of this stage involves transforming raw text data into a format suitable for analysis, enabling the algorithm to discern patterns and sentiments within the over-tourism discourse. Incorporating the Naive Bayes Classifier, known for its efficiency in text classification tasks, ensures a robust analytical foundation. Concurrently, integrating the SMOTE operator addresses the challenge of data imbalance by oversampling minority classes, enhancing the algorithm's ability to accurately classify sentiments within the dataset. This methodological choice reflects a comprehensive approach, combining the strengths of the Naive Bayes Classifier with the efficacy of SMOTE to improve the accuracy and reliability of sentiment analysis results related to over-tourism issues.

The algorithm's performance is assessed through critical metrics, including accuracy, precision, recall, f-measure, and AUC during the evaluation stage. These metrics serve as crucial benchmarks in gauging the effectiveness of the applied algorithm in sentiment analysis tasks related to over-tourism. Accuracy provides an overall measure of the algorithm's correctness in classifying sentiments, while precision quantifies the algorithm's ability to identify positive sentiments among the instances predicted as positive correctly. Recall measures the algorithm's capability to identify all positive instances within the dataset correctly. F-measure combines precision and recall, offering a balanced evaluation of the algorithm's performance. The Area Under the Curve (AUC) metric, derived from the Receiver Operating Characteristic (ROC) curve, also provides insights into the algorithm's discriminatory power. Carefully considering these metrics ensures a comprehensive evaluation, allowing for nuanced insights into the algorithm's performance and reliability in discerning sentiments related to over-tourism issues.

During the deployment stage, recommendations can be made based on the research findings to implement practical applications. The primary objective is to translate the insights gained from sentiment analysis of over-tourism issues into actionable strategies for sustainable tourism management. By deploying the knowledge derived from the algorithm's performance and sentiment classifications, stakeholders in the tourism industry, policymakers, and destination managers can make informed decisions to address the identified sentiments effectively. These recommendations bridge academic research and real-world applications, fostering a more sustainable and responsive approach to managing over-tourism challenges. This deployment stage underscores the significance of research in influencing positive changes within the tourism sector and contributing to the broader discourse on responsible and sustainable tourism practices.

2.2 Naïve Bayes Classifier (NBC)

The Naive Bayes Classifier (NBC) algorithm offers distinct advantages in sentiment classification, supported by several requisite stages. NBC's strength lies in its simplicity, efficiency, and robust performance in text classification tasks, particularly in sentiment analysis. The algorithm operates based on the Bayes theorem, assuming independence between features, which allows for rapid computation and scalability to large datasets. To effectively utilize NBC for sentiment classification, several steps are essential, including data preprocessing to clean and tokenize text, feature extraction to represent text as numerical vectors, model training using labeled data, and model evaluation to assess performance metrics such as accuracy, precision, recall, and F1-score. Additionally, NBC's ability to handle multiclass classification tasks and its resilience to overfitting make it a versatile and reliable choice for sentiment analysis applications.

In conclusion, with its straightforward implementation and firm performance, the NBC algorithm represents a valuable tool for sentiment classification tasks, contributing to natural language processing research and application advancements. The Bayesian Naive classifier requires only a relatively small amount of training data to determine the estimated parameters required for the classification process. At the classification stage, the class value is determined from data based on the term that occurs using the following equation.

$$P(X_k|Y) = \frac{P(Y|X_k)}{\sum iP(Y|X_i)} \quad (1)$$

The posterior state (X_k in Y) can be calculated from the prior state (Y in X_k) divided by the sum of all probability Y in all X_i.

$$\text{P (v1|C=c)} = \frac{CountTerms(v1,docsv(c))}{AllTerms(docs(c))} \quad (2)$$

Where v1 is one of the syllables that appear in over-tourism content reviews. While, $CountTerms(v1, docsv(c))$ Refers to the number of occurrences of a word labeled C ("positive" or "negative"). As for $AllTerms(docs(c))$ refers to the sum of all C-labeled words in the dataset. To avoid zero values in probability, place smoothing is implemented to reduce the probability of observed results and increase the probability of unobserved results. Thus, the equation used is as follows:

$$\text{P (v1|C=c)} = \frac{CountTerms(v1,docsv(c))+1}{AllTerms(docs(c))+|V|} \quad (3)$$

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

Where |V| refers to the sum of all words in the review data present in the dataset. Thus, classifying review data will show the word with the highest value to represent the reviewer's attention to the products and services obtained. The relevance of utilizing NBC in sentiment classification based on review data concerning over-tourism content is noteworthy. Given the proliferation of digital platforms and user-generated content related to tourism experiences, sentiment analysis is crucial in understanding public perceptions and attitudes towards over-tourism issues. NBC's ability to efficiently handle text classification tasks, particularly in the context of sentiment analysis, makes it well-suited for analyzing reviews and identifying sentiments expressed by tourists and other stakeholders. By leveraging NBC, researchers can effectively categorize reviews into positive, negative, or neutral sentiments, thereby gaining insights into the overarching sentiments surrounding over-tourism. This facilitates informed decision-making and targeted interventions to address concerns and promote sustainable tourism practices. In conclusion, the utilization of NBC in sentiment classification of over-tourism content underscores its relevance and efficacy in extracting valuable insights from textual data, thereby contributing to research in sustainable tourism management.

3. RESULT AND DISCUSSION

Over-tourism emerges as a crucial issue within the context of sustainable tourism, hence underscoring the significance of this research in analyzing public responses and perceptions regarding it [29]. As tourism destinations grapple with the adverse effects of over-tourism, ranging from environmental degradation to cultural disruption, understanding public sentiments and paradigms becomes imperative for devising effective management strategies and fostering sustainable tourism practices [30]. By delving into public reactions towards over-tourism, this research contributes valuable insights to the discourse on sustainable tourism management, formulating informed policies and interventions to mitigate its negative impacts and promoting harmonious tourism development [31].

This research adopts CRISP-DM to classify public sentiment towards over-tourism issues through digital content published in responsible tourism channels. CRISP-DM, known for its structured approach to data mining projects, offers a systematic framework encompassing stages such as business understanding, data processing, modeling, evaluation, and deployment. By leveraging CRISP-DM, this study aims to comprehensively analyze public sentiments expressed in digital content related to responsible tourism, thereby contributing to a nuanced understanding of over-tourism perceptions and facilitating the development of targeted interventions to address these concerns. Through the adoption of CRISP-DM, this research endeavors to advance methodological approaches in sentiment analysis within the realm of sustainable tourism management. Based on the data obtained, the characteristics of respondents are known below.

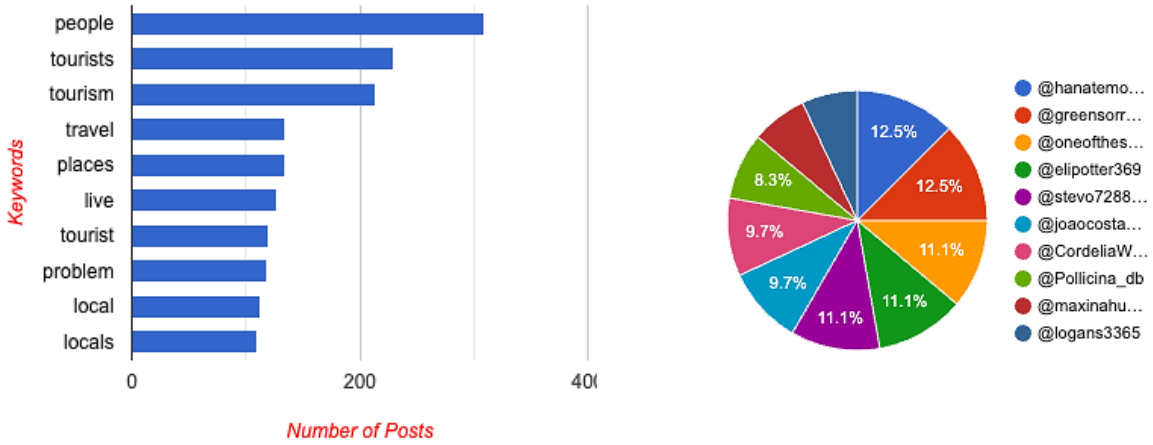


Figure 2. Most Frequently Used Words and Top Ten posters

Over-tourism issues commented upon during the business understanding stage will be analyzed using a sentiment analysis approach [32]. This initial phase of the CRISP-DM methodology is crucial for identifying key themes and concerns surrounding over-tourism within responsible tourism discourse [33]. Researchers can systematically categorize and evaluate public sentiments expressed in digital content by employing sentiment analysis, thereby gaining insights into prevailing attitudes and perceptions towards over-tourism issues [34]. This proactive approach facilitates the identification of relevant sentiment indicators [35]. It lays the groundwork for subsequent data processing and modeling stages, enabling a comprehensive exploration of public sentiment dynamics surrounding over-tourism. Based on the results of the sentiment extract, it can be seen that the rapid miner application is efficient in converting strings into scores, as shown in the following table.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

Table 1. Extract Sentiment in Rapidminer

Review	Score	Total Score
TikTok is responsible for a huge part of this .... I just hope that Safari tourism doesn't become a mainstream trend among Gen Z and millenials .... I don't want our national parks flooded with Gen Z tiktokers disturbing the animals and littering the parks...most of the Safari tourists we see are mostly Gen X and boomers, they are less obnoxious ,don't litter everywhere and crucially spend alot of money. I don't think it will happen though, safari drives are very thrilling but aren't exactly tiktok short material ,nat geo took care of that... Btw Rwanda and Uganda heavily restrict the no of tourists going to see the gorillas and for good reason. It's easy for a national park to do that but a city? Quite difficult but it can be done ... restricting the permits for short term rentals, hotels while encouraging residential buildings, restricting the no of listings in that city in airbnb and such and most of all restrict photography if people want a photograph license local photographers ....in short make going there difficult.	responsible (0.33) huge (0.33) hope (0.49) want (0.08) disturbing (-0.59) obnoxious (-0.51) thrilling (0.54) care (0.56) restrict (-0.41) no (-0.31) good (0.49) easy (0.49) difficult (-0.38) restricting (-0.41) encouraging (0.62) restricting (-0.41) no (-0.31) restrict (-0.41) want (0.08) difficult (-0.38)	-0,128205128205128
I loved this documentary. My wife and I love to travel and try really hard to be respectful of local cultures where we go. We just had a 3-week road trip around France; starting and ending in Paris. We saw the effects of over-tourism first-hand and how poorly some people act, especially Americans. Travel isn't going away but people could certainly be more considerate. Too many Americans are an embarrassment; we don't like telling people we're from the US because of the negative perception. Something needs to be done about AirBnB and how it is ruining life for locals. Sedona, AZ is a perfect example: the workers used to be able to live in town but have been pushed out by high rents because so many AirBnB's have taken over. It's toxic capitalism.	loved (0.74) love (0.82) hard (-0.10) respectful (0.51) hand (0.56) certainly (0.36) considerate (0.49) embarrassment (-0.49) like (0.38) negative (-0.69) ruining (-0.26) perfect (0.69)	3,02564102564103

This indicates that the sentiment extraction process is conducted by calculating the scores of each word with positive and negative connotations to facilitate classification using the NBC algorithm. By assigning sentiment scores to individual words, researchers can systematically quantify the emotional tone conveyed in digital content related to over-tourism, enabling the classification of sentiments as positive, negative, or neutral. This methodological approach underscores the importance of leveraging computational techniques like sentiment analysis to extract meaningful insights from textual data and inform decision-making processes in sustainable tourism management.

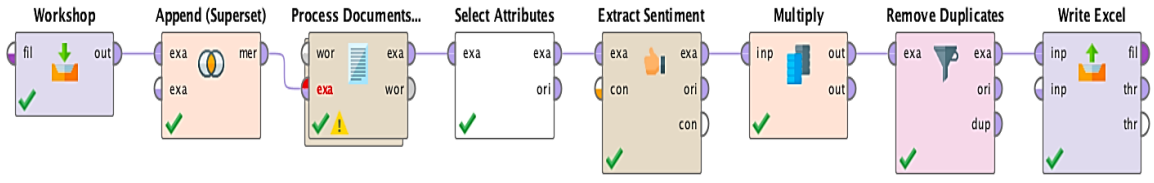


Figure 3. Extract Sentiment Process

The sentiment extraction process in the RapidMiner application yields optimal results, indicating its effectiveness in analyzing sentiments within digital content. Leveraging advanced data mining capabilities, RapidMiner facilitates the extraction of sentiment indicators from large datasets, enabling researchers to discern nuanced sentiment patterns and trends. This robust performance underscores the utility of RapidMiner as a valuable tool for sentiment analysis tasks, particularly in the context of understanding public perceptions towards complex issues such as over-tourism. As such, the application of RapidMiner contributes to advancing research methodologies in sentiment analysis and provides researchers with valuable insights into public sentiment dynamics.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

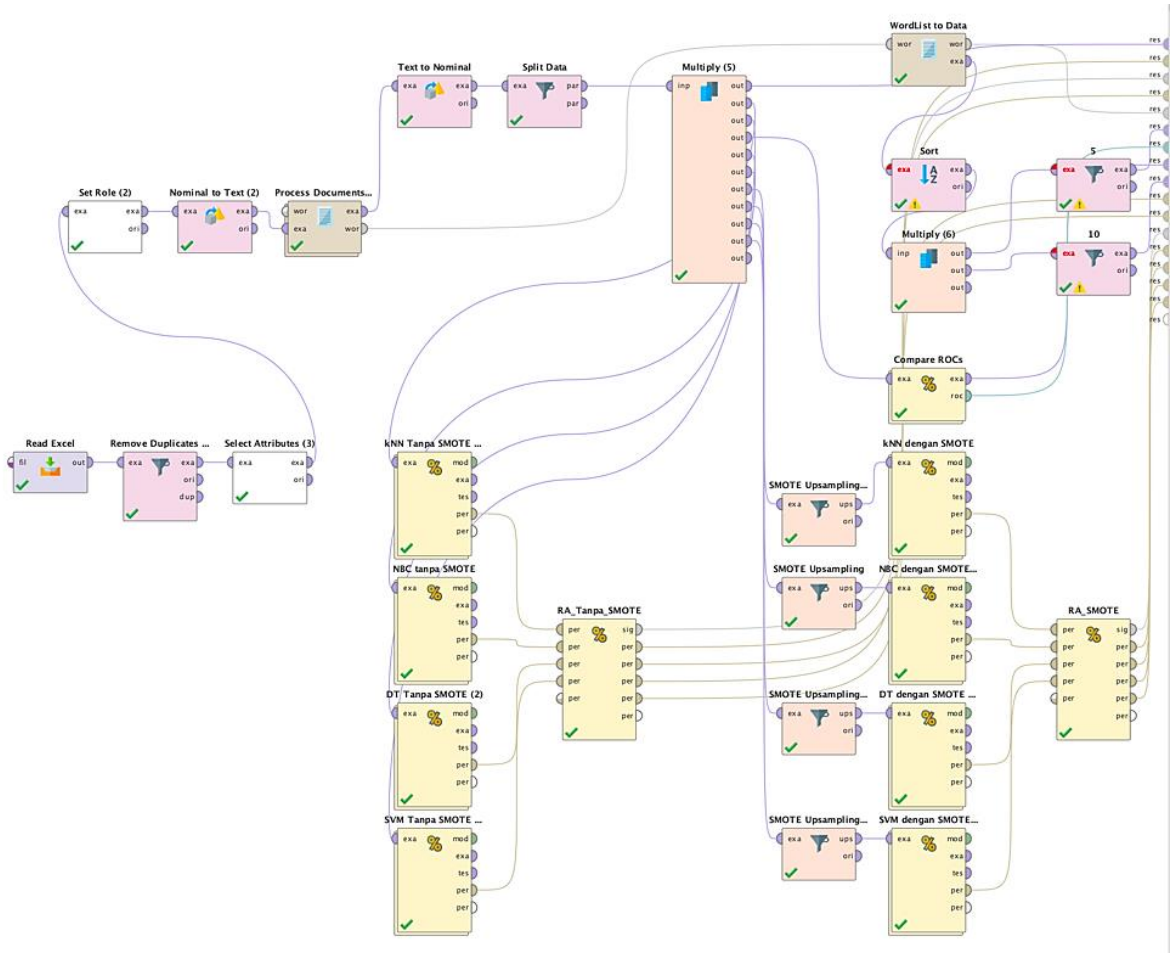


Figure 4. Modeling Process in Rapidminer

In the modeling stage, several processes are established to connect the sentiment extraction results with the NBC algorithm, aiming to optimize the classification process. One of the critical processes involves cleaning the extracted data using the TF-ID approach, which enhances the dataset's quality by removing irrelevant or redundant information and emphasizing the significance of terms based on their frequency and uniqueness within the corpus. By applying TF-IDF, the classification process becomes more efficient and accurate, focusing on relevant features while reducing noise and improving the algorithm's ability to discern sentiment patterns effectively. This systematic approach ensures the robustness and effectiveness of sentiment analysis in capturing public perceptions of over-tourism issues, thereby contributing to a comprehensive understanding of sentiment dynamics within sustainable tourism management. In addition, the results of the visualization of famous words in Word Cloud can be seen in the following figure.

Figure 5. Popular Words

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

Based on the performance measurements of the NBC algorithm, differences in accuracy, precision, recall, f-measure, and AUC values are evident when comparing the utilization of the SMOTE operator with its absence in the RapidMiner application. The integration of SMOTE addresses class imbalance issues by oversampling minority classes, thereby enhancing the algorithm's ability to classify sentiments within the dataset accurately. These findings underscore the importance of addressing class imbalance in sentiment analysis tasks and highlight the efficacy of SMOTE in enhancing the predictive capabilities of the NBC algorithm within the RapidMiner environment.

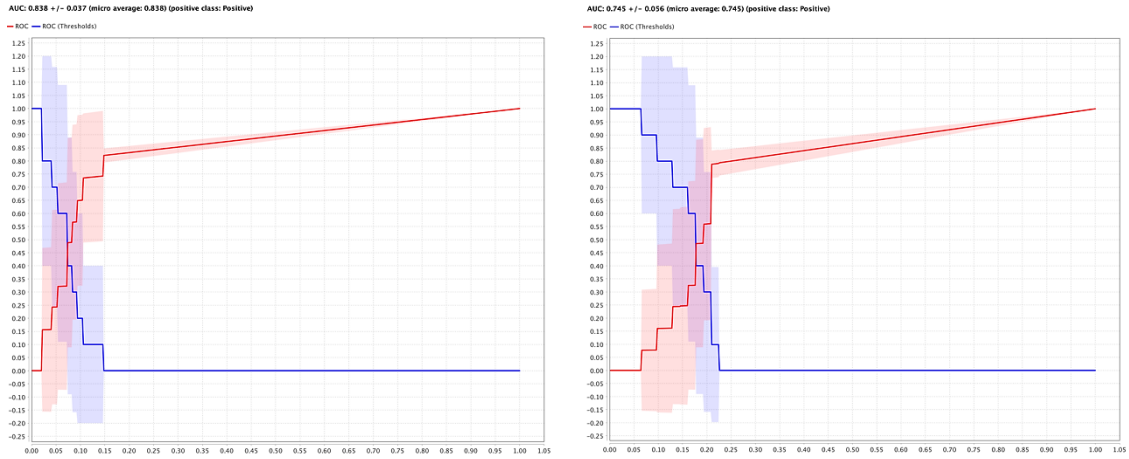
Table 2. NBC Performance in Sentiment Classification

Using SMOTE	Without using SMOTE
PerformanceVector: Accuracy: 84.82% +/- 2.64% (micro average: 84.82%) ConfusionMatrix: True: Negative Positive Negative: 887 222 Positive: 68 733 AUC (optimistic): 0.983 +/- 0.010 (micro average: 0.983) (positive class: Positive) AUC: 0.838 +/- 0.037 (micro average: 0.838) (positive class: Positive) AUC (pessimistic): 0.748 +/- 0.041 (micro average: 0.748) (positive class: Positive) precision: 91.69% +/- 4.24% (micro average: 91.51%) (positive class: Positive) ConfusionMatrix: True: Negative Positive Negative: 887 222 Positive: 68 733 recall: 76.75% +/- 3.81% (micro average: 76.75%) (positive class: Positive) ConfusionMatrix: True: Negative Positive Negative: 887 222 Positive: 68 733 f_measure: 83.47% +/- 2.87% (micro average: 83.49%) (positive class: Positive) ConfusionMatrix: True: Negative Positive Negative: 887 222 Positive: 68 733	PerformanceVector: Accuracy: 78.16% +/- 3.84% (micro average: 78.16%) ConfusionMatrix: True: Negative Positive Negative: 519 243 Positive: 101 712 AUC (optimistic): 0.955 +/- 0.018 (micro average: 0.955) (positive class: Positive) AUC: 0.745 +/- 0.056 (micro average: 0.745) (positive class: Positive) AUC (pessimistic): 0.655 +/- 0.065 (micro average: 0.655) (positive class: Positive) precision: 87.61% +/- 3.82% (micro average: 87.58%) (positive class: Positive) ConfusionMatrix: True: Negative Positive Negative: 519 243 Positive: 101 712 recall: 74.56% +/- 4.28% (micro average: 74.55%) (positive class: Positive) ConfusionMatrix: True: Negative Positive Negative: 519 243 Positive: 101 712 f_measure: 80.51% +/- 3.63% (micro average: 80.54%) (positive class: Positive) ConfusionMatrix: True: Negative Positive Negative: 519 243 Positive: 101 712

Based on the performance measurements of the NBC algorithm using the SMOTE operator, it is evident that it achieves high accuracy, precision, recall, and f-measure values, particularly in classifying positive sentiments related to over-tourism issues. The accuracy of 84.82%, precision of 91.69%, recall of 76.75%, and f-measure of 83.47% reflect the algorithm's robust performance in accurately identifying positive sentiment instances within the dataset. These results underscore the efficacy of employing the SMOTE operator in enhancing the predictive capabilities of the NBC algorithm, thereby contributing to more accurate sentiment analysis outcomes in the context of over-tourism.

The significance of the AUC value lies in its role as a critical indicator for measuring the performance of the NBC algorithm. AUC, or Area Under the Curve, represents the area under the Receiver Operating Characteristic (ROC) curve and provides insights into the algorithm's ability to discriminate between positive and negative classes. A higher AUC value indicates better discrimination power, suggesting that the algorithm can effectively distinguish between positive and negative sentiment instances. Additionally, AUC is a comprehensive metric considering both sensitivity (recall) and specificity, offering a holistic assessment of the algorithm's performance. Thus, AUC serves as a valuable benchmark for evaluating the effectiveness of the NBC algorithm in sentiment analysis tasks, particularly in capturing nuanced sentiments related to over-tourism issues. Meanwhile, the comparison of AUC charts can be seen in the following figure. Meanwhile, the comparison of AUC charts can be seen in the following figure.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904



(a) (b)

Figure 6. Area Under Curve (AUC) of the NBC Algorithm: (a) Using SMOTE, (b) Without using SMOTE

Based on the performance measurements of the NBC algorithm, it is evident that the AUC values vary depending on the utilization of the SMOTE operator. Specifically, when employing SMOTE, the algorithm demonstrates significantly higher AUC values, indicating superior discriminative power in classifying positive sentiments related to over-tourism issues. The optimistic AUC value of 0.983 and the standard AUC value of 0.838 suggest strong predictive capabilities in distinguishing between positive and negative sentiment classes. Conversely, the AUC value without SMOTE drops to 0.748, indicating reduced discriminative power. These findings highlight the importance of employing techniques like SMOTE to enhance the performance of sentiment analysis algorithms, ultimately contributing to more accurate and reliable analyses of public sentiment towards over-tourism. The accuracy, precision, recall, and f-measure metrics demonstrate consistent and high performance in sentiment classification, particularly in identifying positive sentiments related to over-tourism issues. The AUC values highlight the algorithm's robustness, indicating strong discriminative power in distinguishing between positive and negative sentiment classes. These findings underscore the effectiveness of employing the SMOTE operator in enhancing the predictive capabilities of the NBC algorithm, thereby contributing to more accurate sentiment analysis outcomes in the context of over-tourism.

The findings of this research reveal certain limitations, primarily focusing on testing the NBC algorithm through the CRISP-DM methodology. At the same time, other over-tourism content warrants comprehensive analysis to enable cross-media perspective comparisons. While the application of CRISP-DM provides a structured framework for sentiment analysis, its scope may not encompass all facets of over-tourism discourse across various digital media platforms. Therefore, future research endeavors should extend beyond algorithm testing to encompass a broader range of over-tourism content, allowing for a more comprehensive understanding of public perceptions and sentiments. By adopting a cross-media approach, researchers can gain insights into how over-tourism issues are portrayed and perceived across different digital platforms, thus enriching the discourse on sustainable tourism management.

Further research recommendations are proposed to investigate over-tourism issues within the context of Indonesia, mainly focusing on tourist destinations [36]. Exploring over-tourism dynamics presents an invaluable opportunity to delve into the intricate interplay between tourism development, environmental conservation, socio-cultural impacts, and community livelihoods [36]. By conducting in-depth studies specific to the destination area in Indonesia, researchers can offer insights tailored to this renowned tourist destination's unique challenges and opportunities [37]. Such research endeavors hold significant potential for informing policy-making processes, facilitating sustainable tourism practices, and fostering community resilience in the face of burgeoning tourist arrivals [38]. Therefore, future research initiatives should prioritize the examination of over-tourism to contribute meaningfully to the discourse on sustainable tourism management in Indonesia.

4. CONCLUSION

The findings of this study reveal that the NBC algorithm with SMOTE demonstrates commendable performance metrics, including an accuracy of 84.82%, precision of 91.69%, recall of 76.75%, f-measure of 83.47%, and AUC of 0.838. Compared to the performance without SMOTE, which yields an accuracy of 78.16%, precision of 87.61%, recall of 74.56%, f-measure of 80.51%, and AUC of 0.745, the utilization of SMOTE significantly enhances the predictive capabilities of the NBC in sentiment analysis of over-tourism issues, emphasizing its effectiveness in addressing class imbalance. Moreover, the practical implications of this research underscore the importance of analyzing algorithm performance through the CRISP-DM methodology in gauging public

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

sentiment towards over-tourism issues within responsible tourism content. These insights contribute to advancing methodological approaches in sentiment analysis and informing decision-making processes in sustainable tourism management.

REFERENCES

[1] P. Madzík, L. Falát, L. Copuš, and M. Valeri, "Digital transformation in tourism: bibliometric literature review based on machine learning approach," *Eur. J. Innov. Manag.*, vol. 26, no. 7, pp. 177–205, 2023, doi: 10.1108/EJIM-09-2022-0531.

[2] M. R. Khomsi, L. Fernandez-Aubin, and L. Rabier, "A prospective analysis of overtourism in Montreal," *J. Travel Tour. Mark.*, vol. 37, no. 8–9, pp. 873–886, 2020, doi: 10.1080/10548408.2020.1791782.

[3] M. O' Regan, N. B. Salazar, J. Choe, and D. Buhalis, "Unpacking overtourism as a discursive formation through interdiscursivity," *Tour. Rev.*, vol. 77, no. 1, pp. 54–71, 2022, doi: 10.1108/TR-12-2020-0594.

[4] H. Jang and M. Park, "Social media, media and urban transformation in the context of overtourism," *Int. J. Tour. Cities*, vol. 6, no. 1, pp. 233–260, 2020, doi: 10.1108/IJTC-08-2019-0145.

[5] A. Guizi, Z. Breda, and R. Costa, "How are overtourism and host–guest relationships portrayed by the Portuguese print media?," *Int. J. Tour. Cities*, vol. 6, no. 1, pp. 215–232, 2020, doi: 10.1108/IJTC-06-2019-0081.

[6] A. Walmsley, K. Koens, and C. Milano, "Overtourism and employment outcomes for the tourism worker: impacts to labour markets," *Tour. Rev.*, vol. 77, no. 1, pp. 1–15, 2022, doi: 10.1108/TR-07-2020-0343.

[7] R. Peterson and R. B. DiPietro, "Is Caribbean tourism in overdrive? Investigating the antecedents and effects of overtourism in sovereign and nonsovereign small island tourism economies (SITEs)," *Int. Hosp. Rev.*, vol. 35, no. 1, pp. 19–40, 2021, doi: 10.1108/ihr-07-2020-0022.

[8] C. Pasquinelli and M. Trunfio, "The missing link between overtourism and post-pandemic tourism. Framing Twitter debate on the Italian tourism crisis," *J. Place Manag. Dev.*, vol. 15, no. 3, pp. 229–247, 2022, doi: 10.1108/JPMD-07-2020-0073.

[9] E. Agyeiwaah, "Over-tourism and sustainable consumption of resources through sharing: the role of government," *Int. J. Tour. Cities*, vol. 6, no. 1, pp. 99–116, 2020, doi: 10.1108/IJTC-06-2019-0078.

[10] F. P. Ribeiro and K. Torkington, "Conflicting discursive representations of overtourism in Lisbon in the Portuguese digital press," *Int. J. Tour. Cities*, vol. 9, no. 1, pp. 286–301, 2023, doi: 10.1108/IJTC-05-2022-0108.

[11] S. Çelik and S. M. Rasoolimanesh, "Residents' Attitudes towards Tourism, Cost–Benefit Attitudes, and Support for Tourism: A Pre-development Perspective," *Tour. Plan. Dev.*, vol. 20, no. 4, pp. 522–540, 2023, doi: 10.1080/21568316.2021.1873836.

[12] E. Koh, "The end of over-tourism? Opportunities in a post-Covid-19 world," *Int. J. Tour. Cities*, vol. 6, no. 4, pp. 1015–1023, 2020, doi: 10.1108/IJTC-04-2020-0080.

[13] I. Purwandani and S. P. Pakan, "Local habitus and temporal overtourism in Yogyakarta," *Consum. Behav. Tour. Hosp.*, vol. 17, no. 4, pp. 544–560, 2022, doi: 10.1108/CBTH-07-2021-0177.

[14] A. Amore, M. Falk, and B. A. Adie, "One visitor too many: assessing the degree of overtourism in established European urban destinations," *Int. J. Tour. Cities*, vol. 6, no. 1, pp. 117–137, 2020, doi: 10.1108/IJTC-09-2019-0152.

[15] M. R. Barbhuiya, "Overtourism in Indian cities: a case study of Nainital," *Int. J. Tour. Cities*, vol. 7, no. 3, pp. 702–724, 2021, doi: 10.1108/IJTC-08-2019-0148.

[16] J. H. Nilsson, "Conceptualizing and contextualizing overtourism: the dynamics of accelerating urban tourism," *Int. J. Tour. Cities*, vol. 6, no. 4, pp. 657–671, 2020, doi: 10.1108/IJTC-08-2019-0117.

[17] R. W. Butler and R. Dodds, "Overcoming overtourism: a review of failure," *Tour. Rev.*, vol. 77, no. 1, pp. 35–53, 2022, doi: 10.1108/TR-04-2021-0215.

[18] T. Mihalic and K. Kuščer, "Can overtourism be managed? Destination management factors affecting residents' irritation and quality of life," *Tour. Rev.*, vol. 77, no. 1, pp. 16–34, 2022, doi: 10.1108/TR-04-2020-0186.

[19] G. Wall, "From carrying capacity to overtourism: a perspective article," *Tour. Rev.*, vol. 75, no. 1, pp. 212–215, 2020, doi: 10.1108/TR-08-2019-0356.

[20] P. Guimarães, "Retail change in a context of an overtourism city. The case of Lisbon," *Int. J. Tour. Cities*, vol. 7, no. 2, pp. 547–564, 2021, doi: 10.1108/IJTC-11-2020-0258.

[21] S. M. Rasoolimanesh and S. Seyfi, "Residents' perceptions and attitudes towards tourism development: a perspective article," *Tour. Rev.*, vol. 76, no. 1, pp. 51–57, 2021, doi: 10.1108/TR-11-2019-0461.

[22] R. Martínez Suárez, J. A. Castañeda García, and M. Á. Rodríguez Molina, "Identifying tourist profiles to reduce overtourism: the case of a cultural destination," *Int. J. Tour. Cities*, vol. 7, no. 4, pp. 962–985, 2021, doi: 10.1108/IJTC-08-2020-0153.

[23] 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.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 2, February 2024, Page 275-285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i2.4904

[24] 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.

[25] Y. A. Singgalen, "Analisis Sentimen Wisatawan Melalui Data Ulasan Candi Borobudur di Tripadvisor Menggunakan Algoritma Naïve Bayes Classifier," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, p. 1343−1352, 2022, doi: 10.47065/bits.v4i3.2486.

[26] 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.

[27] 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.

[28] 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.

[29] F. Higgins-Desbiolles, "The 'war over tourism': challenges to sustainable tourism in the tourism academy after COVID-19," *J. Sustain. Tour.*, vol. 29, no. 4, pp. 551–569, 2020, doi: 10.1080/09669582.2020.1803334.

[30] P. Mohan and E. Strobl, "Tourism and tax revenue: evidence from stay-over tourists in the Eastern Caribbean," *Curr. Issues Tour.*, pp. 1–23, 2023, doi: 10.1080/13683500.2023.2202307.

[31] E. Koh and P. Fakfare, "Overcoming 'over-tourism': the closure of Maya Bay," *Int. J. Tour. Cities*, vol. 6, no. 2, pp. 279–296, 2020, doi: 10.1108/IJTC-02-2019-0023.

[32] N. Camatti, D. Bertocchi, H. Carić, and J. van der Borg, "A digital response system to mitigate overtourism. The case of Dubrovnik," *J. Travel Tour. Mark.*, vol. 37, no. 8–9, pp. 887–901, 2020, doi: 10.1080/10548408.2020.1828230.

[33] M. O'Regan and J. Choe, "#overtourism on Twitter: a social movement for change or an echo chamber?," *Curr. Issues Tour.*, vol. 26, no. 7, pp. 1082–1095, 2023, doi: 10.1080/13683500.2022.2047161.

[34] M. Blázquez-Salom, M. Cladera, and M. Sard, "Identifying the sustainability indicators of overtourism and undertourism in Majorca," *J. Sustain. Tour.*, vol. 31, no. 7, pp. 1694–1718, 2023, doi: 10.1080/09669582.2021.1942478.

[35] B. Szuster, M. D. Needham, L. Lesar, and Q. Chen, "From a drone's eye view: indicators of overtourism in a sea, sun, and sand destination," *J. Sustain. Tour.*, vol. 31, no. 7, pp. 1538–1555, 2023, doi: 10.1080/09669582.2020.1866586.

[36] S. Abbasian, G. Onn, and D. Arnautovic, "Overtourism in Dubrovnik in the eyes of local tourism employees: A qualitative study," *Cogent Soc. Sci.*, vol. 6, no. 2, pp. 1–15, 2020, doi: 10.1080/23311886.2020.1775944.

[37] I. Diaz-Parra and J. Jover, "Overtourism, place alienation and the right to the city: insights from the historic centre of Seville, Spain," *J. Sustain. Tour.*, vol. 29, no. 2–3, pp. 158–175, 2021, doi: 10.1080/09669582.2020.1717504.

[38] T. Pereira, C. Berselli, L. A. Pereira, and P. F. Limberger, "Overtourism: An Analysis of Demographic and Socioeconomic Factors with the Evasion Indicators of Residents in Brazilian Coastal Destinations," *Tour. Plan. Dev.*, vol. 19, no. 6, pp. 526–549, 2022, doi: 10.1080/21568316.2022.2027510.