Enhancing Sentiment Analysis of Garden by the Bay Reviews on TripAdvisor Platform Using CRISP-DM through DT and SVM with SMOTE

Yerik Afrianto Singgelen
Faculty of Business Administration and Communication, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Email: yerik.afrianto@atmajaya.ac.id
Email Penulis Korespondensi: yerik.afrianto@atmajaya.ac.id

Abstract-This research aims to improve sentiment analysis of reviews related to Garden by the Bay, a prominent tourist destination in Singapore, by leveraging the CRISP-DM methodology and Synthetic Minority Over-sampling Technique (SMOTE). The study employs a comprehensive approach, integrating CRISP-DM phases to systematically collect, clean, and analyze data from online reviews. The dataset comprises a substantial number of reviews, reflecting diverse visitor experiences. Using SMOTE, class imbalance issues within the dataset are addressed, leading to enhanced performance of sentiment analysis algorithms. The evaluation of Decision Tree (DT) and Support Vector Machine (SVM) algorithms, both with and without SMOTE, reveals significant improvements in accuracy, precision, recall, and F-measure metrics when SMOTE is applied. These findings underscore the efficacy of SMOTE in optimizing sentiment analysis algorithms for the Garden by the Bay dataset, thereby facilitating a deeper understanding of visitor sentiments and experiences, which inform strategies for enhancing the tourism experience at Garden by the Bay.

Keywords: CRISP-DM; DT; SMOTE; SVM; TripAdvisor

1. INTRODUCTION

Promoting tourism activities is pivotal in bolstering a nation's foreign exchange reserves through tourist revenue. Tourism significantly contributes to a country's economic prosperity [1–3]. Firstly, by attracting a steady influx of visitors, nations capitalize on the expenditure of tourists on accommodation, transportation, and various goods and services, thus injecting substantial funds into the local economy [4–6]. Secondly, tourism fosters the development of related industries such as hospitality, transportation, and entertainment, creating employment opportunities and stimulating economic growth [7–10]. Undeniably, the sustained expansion of tourism activities directly correlates with an augmented influx of foreign currency, thereby fortifying the nation's financial standing on the global stage.

An enticing destination for travelers, Garden By the Bay in Singapore stands out as a captivating marvel of urban landscaping and ecological innovation. Situated in the heart of the city-state, this sprawling botanical garden showcases a harmonious blend of nature and technology, offering visitors a multifaceted experience [11, 12]. Firstly, the garden's expansive collection of diverse flora, meticulously curated across various themed zones, presents a rich tapestry of botanical wonders, captivating the imagination of nature enthusiasts and horticultural aficionados alike [13, 14]. Secondly, the iconic Supertree Grove, adorned with vertical gardens and illuminated by a dazzling display of lights, is a testament to Singapore's commitment to sustainability and environmental conservation [15, 16]. Arguably, Garden By the Bay represents Singapore's dedication to green urban development, making it a must-visit destination for travelers seeking natural beauty and architectural ingenuity.

Travelers frequently utilize TripAdvisor to share experiences and insights from visits to various tourist destinations. As a digital repository of user-generated reviews and ratings, TripAdvisor enables tourists to access firsthand information and recommendations, facilitating informed decision-making before embarking on the journeys [17–20]. Moreover, the platform's interactive features, such as forums and discussion boards, foster a sense of community among travelers, allowing them to exchange tips, advice, and anecdotes, enriching the overall travel experience [21–24]. Undoubtedly, the widespread use of TripAdvisor underscores its significance as a valuable tool for travelers seeking to navigate the vast landscape of global tourism, ultimately influencing destination choices and shaping travel trends.

This research aims to classify sentiment based on review data from the Garden by the Bay Singapore destination page on TripAdvisor to analyze tourists' preferences and behaviors. By leveraging natural language processing techniques, the study seeks to discern the underlying sentiments expressed within the plethora of user-generated reviews, thereby providing valuable insights into tourists' attitudes and experiences [25, 25, 26]. Through systematic analysis of sentiment patterns, this research endeavors to uncover nuanced aspects of visitor satisfaction, dissatisfaction, and overall perception, shedding light on key factors influencing destination appeal and visitor decision-making processes [27, 28]. Ultimately, the findings of this investigation are poised to inform strategic initiatives aimed at enhancing the tourism experience and bolstering the attractiveness of Garden by the Bay Singapore as a premier travel destination.

The urgency of this research lies in its capacity to provide actionable insights into tourist sentiment and behavior, which is crucial for informing strategic decision-making in the tourism industry. By systematically analyzing sentiment data derived from TripAdvisor reviews of the Garden by the Bay Singapore destination, this
study addresses a pressing need for a nuanced understanding of visitor preferences and perceptions. Such insights are instrumental in guiding destination management efforts, enabling stakeholders to tailor experiences, amenities, and marketing strategies to meet evolving tourist expectations. Consequently, this research enhances the tourist experience at Garden by the Bay Singapore. It is a valuable resource for advancing broader tourism management practices in an increasingly competitive global landscape.

The theoretical and practical implications of this research are manifold and extend across multiple domains within the field of tourism studies. Firstly, from a theoretical perspective, this study advances sentiment analysis methodologies by applying natural language processing techniques to analyze large-scale review data [29]–[31]. The research enriches our theoretical understanding of visitor behavior and decision-making processes by elucidating the complex interplay between tourist sentiment and destination attributes [32]–[36]. Secondly, in terms of practical implications, the insights gleaned from this study hold significant value for destination management organizations and tourism stakeholders [37]–[39]. Armed with a deeper understanding of tourist preferences and perceptions, practitioners devise targeted strategies to enhance the overall visitor experience at Garden by Bay Singapore, fostering sustainable tourism development and bolstering the destination's competitive edge in the global marketplace. This research's theoretical and practical contributions are poised to inform academic discourse and industry practices, facilitating more informed decision-making and ultimately enriching the tourism landscape.

The contribution to knowledge, when viewed from the perspective of ecotourism within the context of the Garden by the Bay destination, is significant and multifaceted. Primarily, this research sheds light on the intricate relationship between sustainable tourism practices and preserving natural ecosystems, thereby advancing our understanding of ecotourism principles in urban environments. By analyzing sentiment data from TripAdvisor reviews, the study elucidates how visitors interact with and perceive the ecological features of Garden by the Bay, offering valuable insights into the effectiveness of conservation efforts and ecotourism initiatives within the destination. Furthermore, the research provides actionable recommendations for enhancing the sustainability of tourist experiences at Garden by the Bay, fostering a symbiotic relationship between tourism development and environmental preservation. In essence, this contribution to knowledge enriches academic discourse on ecotourism and informs practical strategies for promoting responsible tourism practices and ensuring the long-term viability of urban green spaces.

2. RESEARCH METHODOLOGY

2.1 Gap Analysis

Gap analysis is essential for identifying relevant theoretical concepts within the context of this research. By systematically examining existing literature and theoretical frameworks, this research pinpoints areas where current knowledge falls short or fails to adequately address the specific dynamics of the study's focus. This process allows for identifying gaps or discrepancies in the theoretical understanding of the subject matter, paving the way for formulating research questions and hypotheses that contribute meaningfully to advancing knowledge in the field. Scholars ensure that research addresses critical literature gaps through rigorous gap analysis and enhances the findings' relevance and significance. In conclusion, gap analysis is vital for grounding research within a solid theoretical framework and guiding the exploration of new avenues for scholarly inquiry.

Figure 1. Network, Overlay, and Density Visualization

Figure 1 shows the network, overlay, and density visualization using VosViewer. Based on the gap analysis of theories relevant to the content and context, it is evident that further examination of sentiment analysis and tourist behavior regarding ecotourism is warranted within the context of tourist destinations. The gap analysis reveals a dearth of comprehensive understanding regarding how tourists perceive and interact with ecotourism initiatives within specific destinations. Therefore, conducting in-depth analyses of tourist sentiments and behaviors, particularly in ecotourism, is imperative for advancing our understanding of sustainable tourism
practices. By delving deeper into the nuances of tourist perceptions and behaviors within specific ecotourism destinations, this research contributes valuable insights that inform the development of tailored strategies to foster responsible tourism practices and promote the conservation of natural resources. Thus, this research gap underscores the importance of context-specific investigations to inform evidence-based decision-making in destination management and sustainable tourism development.

Based on the results of gap identification, there is a pressing need for a study on sentiment analysis utilizing algorithms with optimal performance aligned with the CRISP-DM framework. Identifying this gap underscores the importance of employing robust analytical techniques to extract meaningful insights from large volumes of tourist-generated data. By leveraging advanced sentiment analysis algorithms within the CRISP-DM framework, this research systematically preprocesses data, builds predictive models, and evaluates sentiment patterns with precision and efficiency. Consequently, this approach enables a more comprehensive understanding of tourist sentiments and behaviors, facilitating evidence-based decision-making in destination management and sustainable tourism development initiatives. Thus, integrating state-of-the-art analytical methodologies within established frameworks is crucial for addressing the identified research gap and advancing knowledge in tourism studies.

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

The CRISP-DM framework is employed in this research to comprehend the destination context and data utilization in testing appropriate classification models. This framework, which stands for Cross-Industry Standard Process for Data Mining, offers a systematic approach to data mining projects, encompassing distinct phases such as business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Adhering to the CRISP-DM framework ensures methodological rigor and consistency throughout the research process, from defining business objectives to deploying predictive models for sentiment analysis. Consequently, using CRISP-DM facilitates a structured and comprehensive exploration of tourist sentiment and behavior, ultimately advancing knowledge in tourism studies.

Figure 2 shows the implementation of the CRISP-DM framework. Using CRISP-DM is relevant and highly effective in guiding the research process, ensuring efficiency and methodological rigor. This standardized framework offers a structured methodology for handling complex data mining projects, enabling this research to systematically navigate various stages of the research process, from initial data exploration to model deployment. By providing a clear roadmap for data understanding, preparation, modeling, evaluation, and deployment, CRISP-DM streamlines the research workflow, thereby enhancing efficiency and minimizing potential pitfalls or oversights. Moreover, the framework’s flexibility allows for adapting methodologies to suit the research context’s specific requirements and complexities, ensuring the findings’ relevance and applicability. In conclusion, using CRISP-DM enhances the effectiveness of research endeavors and contributes to producing high-quality, reliable outcomes in tourism studies.

Under the CRISP-DM framework, filtering is conducted based on the business understanding stage to comprehend the destination context. This initial phase involves establishing a clear understanding of the business objectives and requirements and identifying the key factors influencing the tourism landscape of the destination. Through rigorous analysis and stakeholder engagement, this research gains valuable insights into the unique characteristics, challenges, and opportunities associated with the destination, laying the foundation for subsequent data collection and analysis. By leveraging the business understanding stage of the CRISP-DM framework, this research ensures alignment between research objectives and the practical needs of destination management, thereby enhancing the relevance and applicability of the findings to real-world scenarios.

2.2.1 Business Understanding

During the business understanding stage, the destination context for Garden by the Bay is identified by analyzing the TripAdvisor platform. This pivotal phase involves a comprehensive examination of user-generated content, including reviews, ratings, and discussions, to discern prevailing sentiments, preferences, and experiences of tourists visiting the destination. By delving into the wealth of information available on TripAdvisor, this research gains valuable insights into the distinctive features, attractions, and amenities of Garden by the Bay and the perceived strengths and weaknesses from visitors’ perspectives. Consequently, this meticulous analysis facilitates a nuanced understanding of the destination context, laying a solid foundation for subsequent stages of the research process, including data collection, analysis, and interpretation.
Figure 3. Detail Information of Garden by the Bay on the TripAdvisor Platform

Figure 3 shows detailed information on Garden by the Bay destination on the TripAdvisor platform. An integral component of Singapore's "City in a Garden" initiative, Gardens by the Bay, is meticulously outlined through data gleaned from TripAdvisor reviews. Spanning 101 hectares in Marina Bay, this sprawling urban oasis comprises three distinct waterfront gardens: Bay South, Bay East, and Bay Central. Gardens by the Bay emerges as a vibrant showcase of horticultural ingenuity and garden artistry, encapsulating Singapore's commitment to sustainable urban development and environmental conservation. The destination showcases various plant species through its immersive botanical displays and innovative design. It is a testament to Singapore's status as a global leader in green initiatives. In essence, Gardens by the Bay epitomizes the seamless integration of nature and urban living, offering visitors a captivating glimpse into the intersection of botanical beauty and architectural innovation.

The results of the total review identification reveal a comprehensive overview of visitor sentiment towards the destination, with a total of 60,516 reviews analyzed. Among these, the majority, comprising 43,533 reviews, reflect an "Excellent" rating, indicating high visitor satisfaction. Additionally, 13,834 reviews classify the experience as "Very Good," further affirming the positive reception of the destination. However, it is noteworthy that a smaller proportion of reviews express less favorable opinions, with 2,542 reviews rated as "Average," 407 as "Poor," and 200 as "Terrible." Despite these varying perceptions, the predominance of positive ratings suggests an overall favorable impression of the destination among TripAdvisor users.

2.2.2 Data Understanding

During the data understanding stage, a comprehensive study explores tourists' perceptions of the Garden by the Bay destination. This crucial phase involves thoroughly examining and analyzing various data sources, including TripAdvisor reviews, to uncover underlying trends, patterns, and sentiments visitors express. By meticulously scrutinizing the information available, this research provides valuable insights into tourists' experiences, preferences, and satisfaction levels, laying a solid foundation for subsequent stages of the research process. Consequently, this in-depth exploration of tourist perceptions contributes to a nuanced understanding of the destination's appeal and facilitates informed decision-making in destination management and marketing strategies.

Figure 4. Observation Result of in Garden by the Bay (April 2024)

Figure 4 shows the observation result on the field. Based on observations conducted at the Garden by the Bay destination, tourists' characteristics are classified based on travel companionship, including solo travelers, friends, family groups, couples, and business travelers. This classification scheme allows for a comprehensive understanding of the diverse demographic profiles and travel preferences exhibited by visitors to the destination. By discerning the varied needs and interests of different traveler types, destination management organizations tailor the offerings and services to cater to the specific requirements of each group, thereby enhancing the overall visitor experience and satisfaction levels. Consequently, this classification of tourist characteristics is a valuable
tool for informing strategic decision-making and resource allocation in destination management and marketing endeavors.

Furthermore, each type of visitor reviews the experiences at Garden by the Bay within different time frames, based on the month and year of visit. Consequently, the review data is classified according to visitor types and segmented based on the timing of visits. This multifaceted approach to data classification enables a more nuanced analysis of tourist sentiments and behaviors, taking into account temporal variations in visitor experiences and perceptions. By considering visitor demographics and temporal factors, this research uncovers more profound insights into the factors influencing visitor satisfaction and preferences, ultimately informing targeted strategies for destination management and marketing initiatives.

Figure 5. Filter Reviews based on Type of Visit and Time of Year (TripAdvisor)

Figure 5 shows the filter reviews based on the type of visit and time of year in the TripAdvisor platform. The review data is categorized based on visitor types and time of visit, allowing for the analysis of sentiments and factors driving tourists to visit Garden by the Bay. This systematic grouping facilitates a comprehensive examination of visitor experiences and perceptions, enabling this research to identify patterns, trends, and correlations between visitor demographics, temporal factors, and sentiment expressions. By dissecting the review data into meaningful segments, insights gleaned into the motivations and preferences of different visitor segments, shedding light on the key drivers influencing visitation to the destination. Consequently, this analytical approach empowers destination management organizations to better tailor the marketing strategies, amenities, and services to meet the target audience's diverse needs and expectations, ultimately enhancing the overall visitor experience at Garden by the Bay.

Figure 6. Scraping Reviews Data using Webharvy

Figure 6 shows the process of scraping review data using Webharvy. Based on the results of scraping review data from Garden by the Bay, 543 review entries have been collected, providing a rich dataset for identifying essential topics related to the tourist destination. This substantial volume of data offers this research a comprehensive overview of the various aspects of visitor experiences and perceptions at Garden by the Bay, enabling the identification of recurring themes, concerns, and highlights. By analyzing this dataset, this research extracts valuable insights into the factors influencing visitor satisfaction, preferences, and decision-making processes, informing strategic initiatives to enhance the overall visitor experience and destination management efforts.
Figure 7. Data Cleaning and Extract Sentiment

Figure 7 shows the data cleaning and extraction in Rapidminer. The collected data has undergone rigorous cleaning and extraction processes to identify negative and positive values from the English string scores. Through meticulous data preparation techniques, such as data cleaning and parsing, this research extracted meaningful sentiment indicators from the raw data, allowing for the classification of reviews into positive and negative categories based on linguistic attributes. This methodological approach enables a systematic analysis of sentiment patterns and trends within the dataset, providing valuable insights into visitor perceptions and experiences at the destination. Consequently, by discerning the nuances of sentiment expressed in the reviews, this research uncovers visitor satisfaction and dissatisfaction, thereby informing targeted strategies for enhancing the overall quality of the tourist experience.

2.2.3 Modeling

During the modeling stage, the data, extracted and classified into negative and positive classes, is subjected to modeling using Support Vector Machine (SVM) and Decision Tree (DT) algorithms. This phase aims to develop predictive models capable of discerning sentiment patterns and accurately classifying reviews based on linguistic features. By employing SVM and DT algorithms, renowned for efficacy in handling classification tasks, this research effectively captures the intricate relationships between input features and sentiment labels, thus facilitating the identification of nuanced sentiment patterns within the dataset. Consequently, this modeling approach enables this research to gain deeper insights into the factors driving positive and negative sentiment among visitors to Garden by the Bay, ultimately informing strategic decisions to enhance the overall visitor experience.

Figure 8. Implementation of DT and SVM in Sentiment Classification

Figure 8 shows the implementation of DT and SVM in sentiment classification using SMOTE. To address the imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) operator is employed to enhance the performance of SVM and DT algorithms. This strategic approach aims to mitigate the impact of class imbalance by generating synthetic instances of the minority class, thereby achieving a more balanced distribution of samples across classes. By leveraging SMOTE, this research effectively bolsters the predictive capabilities of SVM and DT models, improving the ability to accurately classify sentiment in imbalanced datasets. Consequently, this utilization of SMOTE is a vital tool for enhancing the robustness and reliability of sentiment analysis models, ultimately contributing to more accurate insights into visitor perceptions and experiences at Garden by the Bay.

2.2.4 Evaluation

During the evaluation stage, algorithm performance is reassessed based on data division into training (30%) and testing (70%) subsets. This systematic approach enables this research to rigorously evaluate the generalization capabilities of the models by training them on a smaller portion of the data and then testing the performance on a
more significant, independent subset. By assessing model performance on unseen data, this research gauges the robustness and reliability of the sentiment analysis models, ensuring the ability to classify sentiments in real-world scenarios accurately. Consequently, this evaluation methodology serves as a critical step in validating the effectiveness and suitability of the models for informing decision-making processes in the context of Garden by the Bay destination management.

Figure 9. Plot View Performance Evaluation of SVM and DT in Sentiment Classification

Figure 9 shows the plot view performance evaluation of SVM and DT. Each model, whether SVM or DT, demonstrates improved performance after employing SMOTE compared to when SMOTE is not utilized. This performance enhancement is attributed to SMOTE's ability to address the class imbalance by generating synthetic instances of the minority class, thereby enabling the models to learn from a more balanced dataset. By augmenting the representation of the minority class, SMOTE facilitates better discrimination between positive and negative sentiment instances, resulting in more accurate classification outcomes. Consequently, integrating SMOTE into the modeling process is instrumental in enhancing the robustness and efficacy of sentiment analysis models, ultimately contributing to more accurate insights into visitor perceptions and experiences at Garden by the Bay.

2.2.5 Deployment

During the deployment stage, it becomes evident that review data are valuable recommendations for enhancing services and facilities tailored to tourists' preferences, particularly within Garden by the Bay. By leveraging insights from sentiment analysis models, destination management organizations identify areas for improvement and prioritize initiatives that align with visitor expectations and desires. This proactive approach to utilizing review data as actionable insights ensures that resources are allocated effectively, ultimately enhancing the overall visitor experience and bolstering Garden by the Bay's reputation as a premier tourist destination.

During the deployment stage, the most frequently discussed topics in the review data are analyzed based on tourists' needs and preferences to determine recommendations for improving facilities and optimizing tourism services. This meticulous analysis aims to identify recurring themes and areas of concern expressed by visitors, allowing destination management organizations to prioritize enhancements that align with visitor expectations. By addressing the most prominent topics identified through data analysis, tourism providers effectively tailor the offerings to meet the evolving needs of visitors, ultimately enhancing the overall quality of the tourism experience. Consequently, this strategic approach to leveraging review data ensures that resources are allocated efficiently to drive continuous improvement and maintain Garden by the Bay's position as a sought-after destination.

3. RESULT AND DISCUSSION

The discussion in this research is divided into two stages: firstly, the comparison results between the DT and SVM algorithms without employing SMOTE, and secondly, after employing SMOTE. Additionally, the frequently used words are visualized through a word cloud representation. This structured approach allows for a comprehensive examination of the performance of DT and SVM algorithms in handling imbalanced datasets, both before and after applying SMOTE. Furthermore, visualizing frequently used words provides a qualitative insight into the predominant topics and themes in the review data, enhancing the understanding of visitor sentiments and preferences. Consequently, by systematically dissecting the research findings, valuable insights are garnered into the efficacy of different analytical approaches and the key factors influencing visitor experiences at Garden by the Bay.

Based on the top ten frequently used words in the dataset and the respective frequencies, it is observed that specific terms have a higher occurrence within the document. For instance, "gardens" appears 335 times, followed by "dome" with 247 occurrences, indicating a significant presence in the dataset. Similarly, terms such as "visit," "cloud," and "Singapore" are mentioned 226, 219, and 213 times, respectively, reflecting the frequency within the
document. Additionally, words like "place," "flower," and "forest" are also prominent, with occurrences ranging from 188 to 205. The presence of these words in the dataset underscores the importance of capturing the essence of visitor experiences and perceptions at the destination; consequently, by analyzing the frequency of these terms, destination management organizations gain valuable insights into visitor preferences and areas requiring attention, enabling them to tailor strategies for enhancing visitor satisfaction and optimizing the overall tourism experience.

Figure 10. Frequently Used Words based on Positive Class (Rapidminer)

Figure 10 shows the frequently used words based on positive class. Based on the top ten frequently used words in the dataset, it is evident that certain words are highly prevalent within the positive class. For example, "gardens" appeared 329 times, "dome" 242 times, and "visit" 218 times. These findings suggest that positive sentiments are often associated with the destination's ambiance and attractions. Moreover, words such as "Singapore" (212 times) and "place" (200 times) indicate a positive perception of the destination itself. Additionally, the recurrence of words like "cloud" (207 times) and "light" (174 times) suggests that favorable weather conditions and lighting may also influence positive experiences. The prominence of these words underscores the importance of enhancing and maintaining critical features of the destination to bolster positive sentiment among visitors.

Figure 11. Frequently Used Words based on Negative Class (Rapidminer)

Figure 11 shows the frequently used words in the dataset using Rapidminer. Based on the top ten frequently used words in the dataset, it is evident that specific terms appear more frequently in the negative class. Notably, "cloud" and "forest" are mentioned 12 and 13 times, respectively, indicating potential areas of concern or dissatisfaction among visitors. Additionally, terms such as "visit" (8 times) and "place" (5 times) are also prominent, suggesting aspects of the destination experience that may require attention or improvement. These words in the negative class highlight specific areas where interventions or enhancements could be targeted to address visitor concerns and optimize the tourism experience. Consequently, leveraging insights from frequently
used words allows destination management organizations to prioritize initiatives that align with visitor preferences and expectations, ultimately enhancing visitor satisfaction and destination competitiveness.

The subsequent stage involves evaluating the performance of DT and SVM models based on identifying frequently used words tailored to positive and negative classes. This structured approach enables a comprehensive assessment of how well the models classify sentiment based on the identified keywords, providing insights into the effectiveness of discerning positive and negative sentiments within the dataset. By systematically evaluating the performance of DT and SVM models, this research gauges the ability to classify sentiment accurately, thereby informing decision-making processes to improve sentiment analysis methodologies. Consequently, this evaluation stage plays a pivotal role in refining analytical approaches and enhancing the reliability of sentiment analysis models in capturing nuanced sentiment expressions.

### 3.1 Decision Tree Performance Evaluation

The decision tree algorithm's performance was compared before and after employing the SMOTE operator. This research assesses the impact of addressing class imbalance on the algorithm's classification accuracy and generalization capabilities by conducting a comparative analysis. This structured evaluation allows for a systematic examination of how utilizing SMOTE influences the decision tree model's ability to discriminate between different sentiment classes within the dataset effectively. Consequently, this comparative assessment provides valuable insights into the efficacy of employing SMOTE as a strategy for improving the performance of decision tree algorithms in handling imbalanced datasets, ultimately enhancing the reliability and robustness of sentiment analysis models.

#### Figure 12. Decision Tree With and Without SMOTE (Rapidminer)

<table>
<thead>
<tr>
<th>DT with SMOTE</th>
<th>DT without SMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PerformanceVector</td>
<td>PerformanceVector</td>
</tr>
<tr>
<td>accuracy: 98.70% ± 1.70% (micro average: 98.76%)</td>
<td>accuracy: 95.58% ± 1.72% (micro average: 95.50%)</td>
</tr>
<tr>
<td>ConfusionMatrix</td>
<td>ConfusionMatrix</td>
</tr>
<tr>
<td>True: Negative</td>
<td>True: Negative</td>
</tr>
<tr>
<td>Positive: Negative</td>
<td>Positive: Positive</td>
</tr>
<tr>
<td>Non-Equidistant: 9.90% vs. 0.0% (micro average: 9.90%) (positive class: Positive)</td>
<td>Non-Equidistant: 9.08% vs. 0.0% (micro average: 9.08%) (positive class: Positive)</td>
</tr>
<tr>
<td>Positive: Positive</td>
<td>Positive: Negative</td>
</tr>
<tr>
<td>Non-Equidistant: 0.0% vs. 9.80% (micro average: 0.0%) (positive class: Positive)</td>
<td>Non-Equidistant: 0.0% vs. 9.80% (micro average: 0.0%) (positive class: Positive)</td>
</tr>
<tr>
<td>Precision: 91.27% ± 0.78% (micro average: 91.45%) (positive class: Positive)</td>
<td>Precision: 92.25% ± 0.78% (micro average: 92.45%) (positive class: Positive)</td>
</tr>
</tbody>
</table>

In the context of the Garden by the Bay review dataset, analyzing the Decision Tree (DT) algorithm's performance provides insights into its effectiveness for sentiment analysis tasks. The dataset likely contains reviews or feedback from visitors to the Garden by the Bay destination. By applying the DT algorithm to this dataset, this research categorizes sentiments expressed in the reviews, such as positive or negative opinions about

Yerik Afrianto Singgalen, Copyright © 2024, MIB, Page 1196
Submitted: 16/02/2024; Accepted: 29/04/2024; Published: 30/04/2024
the destination. The evaluation of the DT algorithm's performance, both with and without SMOTE, sheds light on its ability to accurately classify sentiments and discriminate between different classes within the dataset. The improvement observed in performance metrics after incorporating SMOTE suggests that addressing class imbalance issues leads to better predictive capabilities of the DT algorithm when applied to sentiment analysis tasks using the Garden by the Bay review dataset. This academic analysis provides valuable insights into the practical application of machine learning techniques for understanding and analyzing sentiment in real-world datasets related to tourism destinations like Garden by the Bay.

3.2 Support Vector Machine Performance

The performance of the Support Vector Machine (SVM) algorithm was compared both before and after applying the SMOTE operator. This comparative analysis enables this research to evaluate how addressing class imbalance impacts the SVM algorithm's classification accuracy and generalization ability. By systematically assessing the SVM model's performance under different conditions, this research discerns the effectiveness of SMOTE in accurately improving the algorithm's capability to classify sentiment in imbalanced datasets. Consequently, this comparative evaluation offers valuable insights into the efficacy of utilizing SMOTE to enhance the performance and reliability of SVM algorithms for sentiment analysis tasks, thereby contributing to more accurate and robust analytical methodologies.

<table>
<thead>
<tr>
<th>SVM with SMOTE</th>
<th>SVM without SMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance</strong>:</td>
<td><strong>Performance</strong>:</td>
</tr>
<tr>
<td><strong>Accuracy</strong>: 98.92% +/- 1.26% (micro average: 96.02%)</td>
<td><strong>Accuracy</strong>: 97.37% +/- 0.48% (micro average: 97.37%)</td>
</tr>
<tr>
<td><strong>Confusion Matrix</strong>:</td>
<td><strong>Confusion Matrix</strong>:</td>
</tr>
<tr>
<td>True: Negative Positive</td>
<td>True: Negative Positive</td>
</tr>
<tr>
<td>Negative: 370 0</td>
<td>Negative: 370 0</td>
</tr>
<tr>
<td>Positive: 0 382</td>
<td>Positive: 0 382</td>
</tr>
<tr>
<td><strong>AUC (Area Under the Curve)</strong>: 0.989 (micro average: 0.989) <strong>(positive class: Positive)</strong></td>
<td><strong>AUC (Area Under the Curve)</strong>: 0.951 (micro average: 0.951) <strong>(positive class: Positive)</strong></td>
</tr>
<tr>
<td><strong>AUC (macro average)</strong>: 0.989 (micro average: 1.000) <strong>(positive class: Positive)</strong></td>
<td><strong>AUC (macro average)</strong>: 0.951 (micro average: 1.000) <strong>(positive class: Positive)</strong></td>
</tr>
<tr>
<td><strong>Precision</strong>: 0.999 (micro average: 0.999) <strong>(positive class: Positive)</strong></td>
<td><strong>Precision</strong>: 0.999 (micro average: 0.999) <strong>(positive class: Positive)</strong></td>
</tr>
<tr>
<td><strong>Recall</strong>: 0.954 (micro average: 0.954) <strong>(positive class: Positive)</strong></td>
<td><strong>Recall</strong>: 0.795 (micro average: 0.795) <strong>(positive class: Positive)</strong></td>
</tr>
<tr>
<td><strong>F-measure</strong>: 0.985 (micro average: 0.985) <strong>(positive class: Positive)</strong></td>
<td><strong>F-measure</strong>: 0.765 (micro average: 0.765) <strong>(positive class: Positive)</strong></td>
</tr>
</tbody>
</table>

Figure 13. Support Vector Machine With and Without SMOTE (Rapidminer)

Figure 13 compares SVM with and without the SMOTE operator in Rapidminer. Based on the evaluation of the SVM algorithm's performance without employing SMOTE, it is evident that the model achieved an accuracy of 97.37%, an AUC (Area Under the Curve) of 0.811, precision of 97.37%, recall of 100.00%, and an F-measure of 98.67%. In contrast, when SMOTE was utilized, the SVM algorithm exhibited improved performance metrics, with an accuracy of 98.92%, AUC of 1.000, precision of 100.00%, recall of 97.84%, and an F-measure of 98.89%. These results demonstrate a significant enhancement in the model's predictive capabilities upon integration with SMOTE, leading to higher accuracy and performance metrics across various evaluation criteria. Consequently, the implementation of SMOTE proves to be effective in addressing class imbalance issues and optimizing the overall performance of the SVM algorithm for sentiment analysis tasks.

Evaluating the Support Vector Machine (SVM) algorithm's performance, both with and without using the Synthetic Minority Over-sampling Technique (SMOTE), reveals substantial improvements in predictive accuracy and overall performance metrics. Without SMOTE, the SVM model achieved a commendable accuracy of 97.37%, along with high precision and recall rates, indicating its effectiveness in classifying sentiment in the dataset. However, upon integrating SMOTE, the SVM algorithm exhibited notable enhancements across all metrics, with accuracy reaching 98.92% and achieving a perfect AUC score of 1.000. This enhancement underscores the efficacy of SMOTE in mitigating class imbalance issues within the dataset, resulting in more robust predictive capabilities and improved performance of the SVM algorithm for sentiment analysis tasks. The findings underscore the importance of addressing class imbalances and the potential of SMOTE to optimize SVM algorithm performance in real-world applications, such as sentiment analysis in tourism-related datasets.

The analysis of the SVM algorithm performance, particularly in the context of sentiment analysis applied to the Garden by the Bay dataset, is pivotal for understanding the algorithm's effectiveness in classifying sentiment within the reviews. The dataset likely consists of many reviews reflecting visitors' experiences, opinions, and sentiments regarding the time at the Garden by the Bay destination. By evaluating the SVM algorithm's performance, this research gauges its ability to accurately classify these sentiments, providing insights into visitors' perceptions and satisfaction levels. Moreover, utilizing SMOTE to address class imbalance within the dataset enhances the reliability of the SVM model's predictions, ensuring that sentiments across various classes are adequately represented and analyzed. This integration of SMOTE improves the accuracy and performance metrics of the SVM algorithm. It enhances its applicability and reliability in real-world scenarios, such as analyzing sentiment in tourism-related datasets like the Garden by the Bay reviews. Overall, the analysis enhances our understanding of the effectiveness of integrating machine learning techniques for sentiment analysis in tourism-related datasets, thereby contributing to more accurate and robust analytical methodologies.
understanding of visitor sentiment and satisfaction levels, thereby informing decision-making processes to optimize the tourist experience at Garden by the Bay.

4. CONCLUSION

In conclusion, employing the CRISP-DM methodology, the research findings underscore the efficacy of both Decision Tree (DT) and Support Vector Machine (SVM) algorithms in sentiment analysis tasks, particularly concerning the Garden by the Bay dataset. The evaluation of these algorithms’ performance reveals notable improvements upon integration with the Synthetic Minority Over-sampling Technique (SMOTE), which addresses class imbalance issues within the dataset. Specifically, DT exhibited enhanced accuracy (96.58% to 98.78%), AUC (0.496 to 0.988), precision (97.35% to 98.72%), recall (99.19% to 98.92%), and F-measure (98.26% to 98.80%) metrics when SMOTE was applied, indicating its improved predictive capabilities. Similarly, SVM showcased enhanced performance metrics, including accuracy (97.37% to 99.92%), AUC (0.811 to 1.000), precision (97.37% to 100.00%), recall (100.00% to 97.84%), and F-measure (98.67% to 98.89%), following the integration with SMOTE. These results emphasize the significance of SMOTE in optimizing algorithm performance, thereby enabling more accurate sentiment classification within the Garden by the Bay dataset. Ultimately, these insights contribute to a better understanding of visitor sentiments and experiences, facilitating informed decision-making processes to enhance the tourism experience at Garden by the Bay.

ACKNOWLEDGMENT

Thanks to the Tourism Department, Faculty of Business Administration and Communication, Atma Jaya Catholic University of Indonesia.

REFERENCES


Yerik Afrianto Singgelen, Copyright © 2024, MIB, Page 1198
Submitted: 16/02/2024; Accepted: 29/04/2024; Published: 30/04/2024