KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 5, April 2024, Hal 2571-2579
DOI 10.30865/klik.v4i5.1826
https://djournals.com/klik

Implementation of SVM and DT for Sentiment Classification: Tempel Hamlet Content Reviews

Yerik Afrianto Singgalen*

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

Abstract−The study aims to investigate the effectiveness of sentiment analysis algorithms, specifically Support Vector Machine (SVM) and Decision Tree (DT), integrated with the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate class imbalance issues. Guided by the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, the research involves several stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The process begins with understanding the business objectives of sentiment analysis and proceeds to explore and prepare the dataset for analysis. SVM and DT algorithms, enhanced with SMOTE, are then implemented for sentiment classification. The study reveals promising results in sentiment analysis tasks. When integrated with SMOTE, SVM achieves an accuracy of 99.21%, while DT attains an accuracy of 98.33%. The Area Under the Curve (AUC) metrics indicate high confidence in classifying positive instances, with SVM and DT demonstrating AUC scores of 1.000 and 0.996, respectively. These findings underscore the efficacy of SVM and DT algorithms, enhanced with SMOTE, in accurately classifying sentiment within text data, thereby addressing class imbalance issues effectively.

Keywords: Classification; Content Analysis; Decision Tree; Sentiment; Support Vector Machine

1. INTRODUCTION

The response of viewers to documentary videos depicting rural settlements and community livelihoods in remote areas necessitates identification and analysis to provide recommendations for content creators regarding filming techniques and storyboard development. By systematically evaluating viewer feedback and engagement metrics, content creators discern audience preferences, narrative structures, and visual storytelling elements that resonate most effectively [1], [2]. This analytical approach enables the formulation of tailored recommendations to optimize the production process, ensuring that documentary videos authentically capture the essence of rural life while engaging and informing viewers [3]–[6]. Ultimately, this systematic analysis facilitates the creation of compelling and impactful content that fosters empathy, understanding, and appreciation for diverse societal experiences.

The digital era's advancement has ignited a burgeoning trend among content creators to craft videos tailored to viewer preferences. As technology proliferates, creators harness data analytics to discern audience tastes, aligning content with viewers' inclinations [7]–[10]. This data-driven approach optimizes content relevance and enhances viewer engagement, fostering a symbiotic relationship between creators and their audiences [11]–[14]. Consequently, content creators are incentivized to continually refine their craft, navigating the intricate landscape of digital platforms to deliver compelling and resonant content experiences, thereby perpetuating a cycle of innovation and consumer satisfaction in the digital ecosystem.

This study aims to analyze the performance of Support Vector Machine (SVM) and Decision Tree (DT) algorithms in classifying text data from reviews of the video content "Tempel Hamlet" with the ID CH2eJH1I2P4&t=995s, which has garnered 895,708 views since its publication on March 16, 2024. By subjecting the textual data to SVM and DT algorithms, researchers can gauge the efficacy of these methods in discerning sentiment and categorizing the content's reception [15]–[19]. Through comparative analysis, this research endeavors to elucidate the strengths and limitations of SVM and DT in text classification tasks, providing valuable insights into their applicability in processing large-scale textual datasets [20], [21].

The method employed for implementing the sentiment classification model is CRISP-DM, an established framework widely utilized in data mining projects. This methodology systematically guides the development of the sentiment analysis model, encompassing distinct stages such as business understanding, data understanding, data preparation, modeling, evaluation, and deployment [22]–[27]. By adhering to the CRISP-DM framework, practitioners can ensure a structured and efficient approach to sentiment classification model implementation, thereby facilitating robust and reliable analysis of textual data for sentiment assessment [28]–[32]. Consequently, using CRISP-DM fosters methodological rigor and enhances the reproducibility and scalability of sentiment analysis endeavors.

This research's theoretical and practical implications extend beyond the immediate scope, offering valuable insights into academic discourse and real-world applications. By elucidating the efficacy of machine learning algorithms, such as Support Vector Machine (SVM) and Decision Tree (DT), in text classification within the context of video content reviews, this study contributes to the theoretical understanding of computational linguistics and data mining [33]–[38]. Moreover, the findings hold significant practical implications for content creators and platform developers, informing strategies for content recommendation systems and audience engagement optimization [11], [39], [40]. This research underscores the transformative potential of data-driven approaches in enhancing decision-making processes across diverse domains by bridging the gap between theoretical inquiry and practical implementation.

The limitation of this research hinges on the methodology employed and the nature of the content analyzed. Methodologically, the study may encounter constraints stemming from the selection and implementation of machine learning algorithms, potentially influencing the accuracy and generalizability of the findings. Moreover, the scope of the content under examination, specifically video reviews of "Tempel Hamlet" with the ID CH2eJH1I2P4&t=995s, may introduce biases or limitations inherent to the subject matter itself, thus restricting the broader applicability of the results. Consequently, while this research provides valuable insights, carefully considering these methodological and content-related limitations is imperative for effectively interpreting and extrapolating its findings. The research contributes significantly to the advancement of knowledge by shedding light on the intricate dynamics of text classification algorithms in the context of video content analysis. Through rigorous experimentation and comparative analysis, this study elucidates the nuanced performance differences between Support Vector Machine (SVM) and Decision Tree (DT) algorithms, thereby providing valuable insights into their efficacy and limitations. By examining the classification accuracy and computational efficiency of these algorithms, the research not only enhances our understanding of text analysis methodologies but also offers practical implications for improving content recommendation systems and enhancing user experience in digital platforms.

2. RESEARCH METHODOLOGY

2.1 Gap Analysis

Gap analysis is conducted to identify the extent of progress in studies concerning data mining, particularly in testing Support Vector Machine (SVM) and Decision Tree (DT) algorithms regarding sentiment classification of video content. This systematic evaluation scrutinizes existing literature to ascertain the gaps and deficiencies in current research methodologies and findings related to applying SVM and DT algorithms in sentiment analysis within video content [41]–[44]. By pinpointing these gaps, researchers can delineate areas for further investigation and innovation, thereby fostering advancements in theoretical understanding and practical implementation of data mining techniques in sentiment analysis of video content.

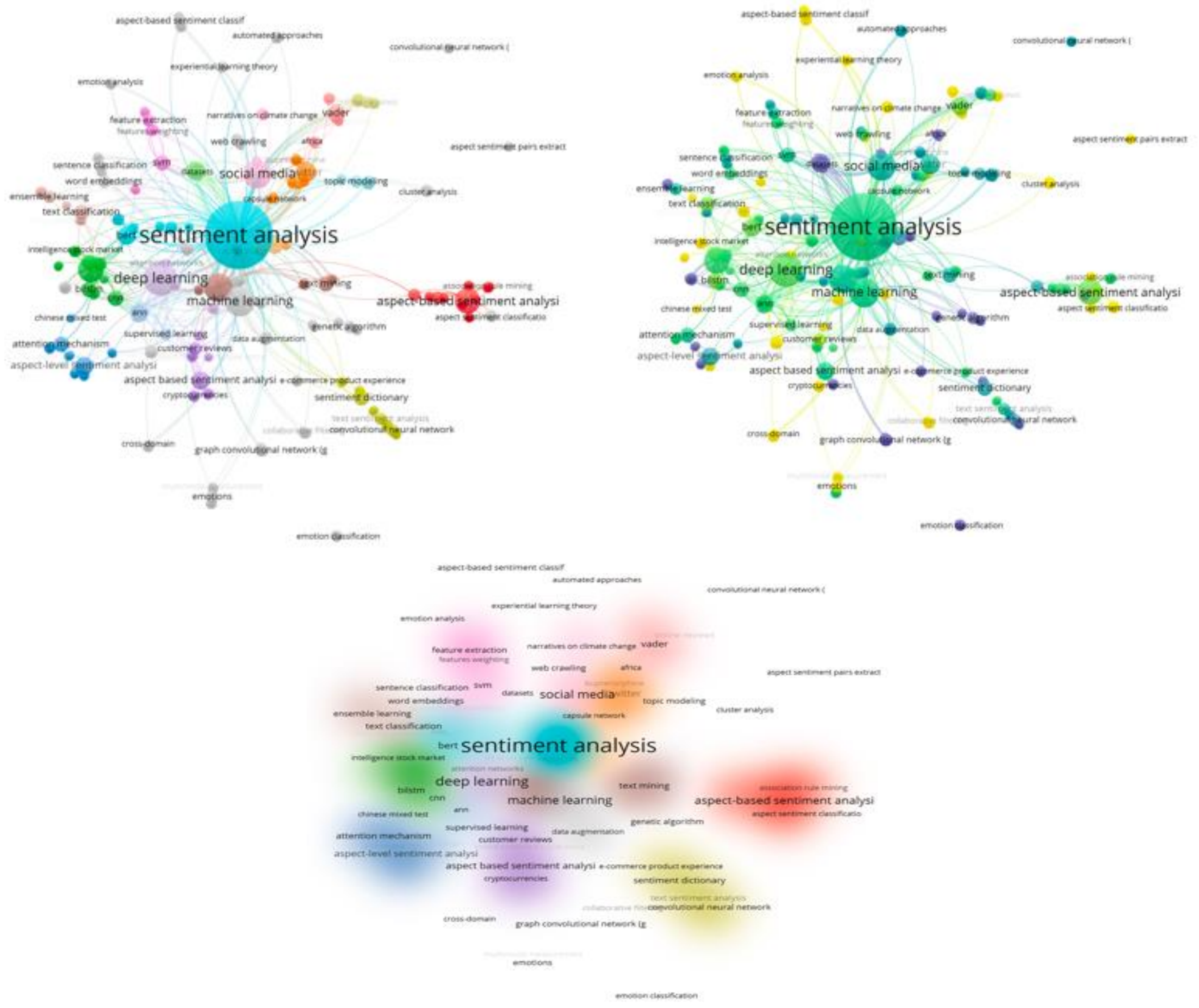

Figure 1. Gap Analysis using Vosviewer

Figure 1 shows the network, overlay, and density visualization of current research related to sentiment analysis. Based on the results of identifying and analyzing gaps in studies related to sentiment analysis, it is evident that specific content analysis focusing on community settlements and livelihoods warrants further examination to depict public responses to educational content in the form of documentary videos [45]–[48]. By delving into the nuances of community dynamics and societal interactions portrayed in documentary videos, researchers can glean more profound insights into how audiences perceive and engage with educational content, thus facilitating the development of more targeted and impactful content-creation strategies. This specialized analysis holds the potential to enrich our understanding of audience preferences and enhance the effectiveness of educational video content in fostering positive societal outcomes.

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

The framework utilized in this research is CRISP-DM, a widely recognized methodology for data mining projects. The implementation stages of this method include business understanding, data understanding, modeling, evaluation, and deployment. By adhering to this structured approach, researchers can systematically navigate through each phase of the data mining process, ensuring alignment with project objectives and facilitating efficient decision-making. Through the iterative nature of CRISP-DM, insights gleaned from earlier stages inform subsequent steps, culminating in the deployment of robust models and actionable findings that contribute to informed decision-making and practical applications in real-world contexts.

2.2.1 Business Understanding

During the business understanding stage, the specific video context under analysis is identified by its unique ID, CH2eJH1I2P4&t=995s, which has garnered 895,708 views since its publication on Mar 16, 2024, along with 1,756 comments. This phase entails a comprehensive examination of the objectives, requirements, and constraints associated with the video content, laying the groundwork for subsequent data understanding and modeling processes. By delineating the key characteristics and metrics of the video, researchers can tailor their analytical approach to extract meaningful insights that align with the overarching goals of the analysis, ultimately contributing to informed decision-making and strategic content development initiatives.

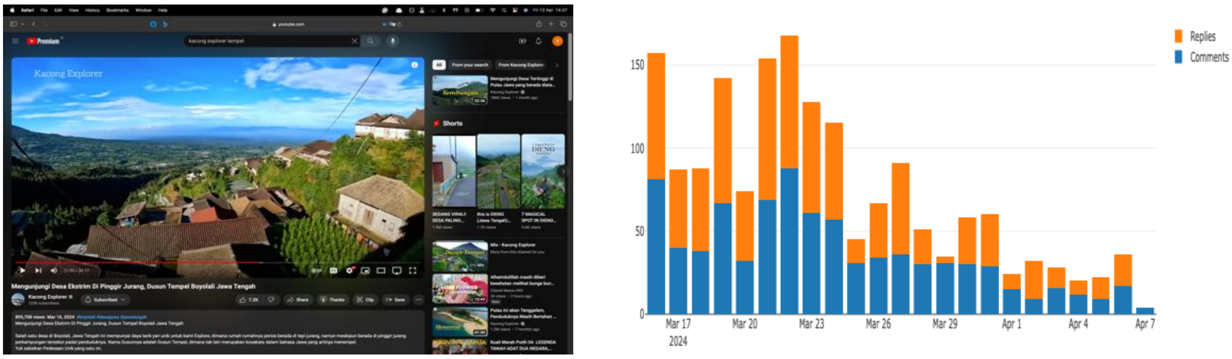

Figure 2. Video and Post-Per-Day Statistic (Communalytic)

Figure 2 shows the video content and post-per-day statistics. Based on the post-per-day statistics, it is evident that both exploratory and educational content resonates with viewers, as reflected in the level of viewer engagement, particularly in the form of comments. The data illustrates that posts on March 22 and 16 received the highest number of comments, indicating heightened viewer interest and interaction with the content. This trend underscores the effectiveness of delivering content that combines exploration and education elements, captures viewers' attention, and prompts them to engage with the material actively. Such insights from viewer comments provide valuable feedback for content creators, guiding them in crafting future content strategies that cater to audience preferences and foster continued viewer engagement and participation.

The interaction between the channel owner and viewers in the comment section demonstrates effective content management by Kacong Explorer. Kacong Explorer fosters community and rapport by engaging with viewers through comments, enhancing viewer satisfaction and loyalty. This proactive approach facilitates direct communication and feedback exchange and reinforces the channel's credibility and authenticity. Such management practices exemplify the importance of nurturing audience relationships in fostering a supportive and engaged viewer base, ultimately contributing to the channel's success and longevity in the digital landscape.

Figure 3. Top Ten Poster

Figure 3 shows the top ten posters of the dataset. Based on the data of the top ten posters, it is evident that @KacongExplorer maintains a significant lead with 414 posts, indicating a dominant presence and active participation within the community. This consistent engagement reflects a solid commitment to content dissemination and audience interaction, crucial to building and sustaining an online presence. The disparity in post frequency between @KacongExplorer and other users underscores the channel's effective content strategy and community management efforts, positioning it as a critical influencer and authority within its niche.

2.2.2 Data Understanding

During the data understanding stage, it is imperative to cleanse and extract the characteristics of the data that will proceed to the modeling process. This initial step is essential to ensure the integrity and quality of the data and facilitate accurate analysis and modeling. By identifying and addressing inconsistencies, errors, and irrelevant information within the dataset, researchers can enhance the reliability and validity of their findings. Furthermore, data cleansing and extraction enable researchers to focus on pertinent variables and features conducive to effective modeling and predictive analytics, thereby optimizing the outcomes of the data mining process.

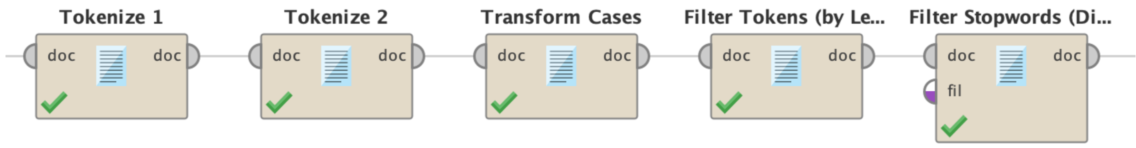

Figure 4. Data Cleaning Process (Rapidminer)

Figure 4 shows the data cleaning process in Rapidminer using operator tokenize, transforms cases, filter tokens, and filter stopwords. In addition, the data cleansing process for text reviews entails utilizing the tokenize operator to disregard non-letters, eliminating expressions such as symbols ([-!"#$%&'()*+,./:;<=>?@\[\\\]_`{|}~]), converting uppercase letters to lowercase, restricting characters to 4-25, and removing stopwords from both the Indonesian and English language dictionaries by the context of the text data to be analyzed. This systematic approach ensures the refinement and standardization of the text data, preparing it for subsequent analysis and modeling. By employing these techniques, researchers can effectively enhance the quality and relevance of the dataset, facilitating more accurate and insightful interpretations of the text data.

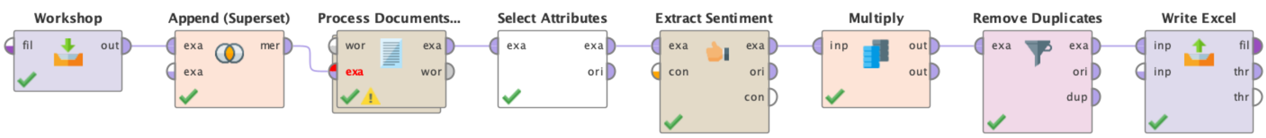

Figure 5. Extract Sentiment Process (Rapidminer)

Figure 5 shows the extract sentiment process in Rapidminer. After the data cleaning process, the subsequent step involves sentiment extraction using the Vader algorithm to obtain sentiment scores before labeling them into negative and positive classes. This procedure enables researchers to quantify the polarity of the text data, providing valuable insights into the overall sentiment conveyed by the reviews. By employing the Vader algorithm designed explicitly for sentiment analysis, researchers can efficiently categorize the text data into distinct sentiment classes, facilitating further analysis and interpretation of the sentiment dynamics within the dataset.

2.2.3 Modeling

During the modeling stage, the classification outcomes of negative and positive classes based on the Vader algorithm are tested through the Support Vector Machine (SVM) and Decision Tree (DT) algorithms. This phase aims to evaluate the performance and effectiveness of different classification techniques in discerning sentiment within the dataset. By employing SVM and DT algorithms, researchers can assess the robustness and generalizability of the sentiment classification model, thereby enhancing the reliability and accuracy of the analysis results. This systematic approach facilitates informed decision-making and enables researchers to draw meaningful conclusions regarding the sentiment dynamics of the text data under study.

The utilization of the Synthetic Minority Over-sampling Technique (SMOTE) serves the purpose of addressing dataset imbalance issues, thereby enabling the assessment of model performance tailored to the dataset's characteristics. Through its synthetic data generation approach, SMOTE effectively balances the distribution of minority and majority classes within the dataset, mitigating the biases introduced by imbalanced data. This methodology enhances the reliability and accuracy of model evaluation by ensuring that the model's predictive capabilities are not skewed towards the majority class, thus facilitating more robust and representative performance measurements. Consequently, incorporating SMOTE is a crucial step in the data preprocessing phase, contributing to the overall effectiveness and integrity of the analytical process.

Thus, identifying the top-performing model refers to both Support Vector Machine (SVM) and Decision Tree (DT) models, which exhibit superior performance metrics in the sentiment analysis task. Through comprehensive evaluation and comparison, these models demonstrate robustness and effectiveness in accurately classifying sentiment within the dataset. Their ability to handle class imbalance and produce high-quality predictions underscores their suitability for deployment in practical applications. As such, selecting SVM and DT as the optimal models reflects their capability to effectively meet the sentiment analysis task requirements, offering valuable insights for decision-making and enhancing user experiences in various domains.

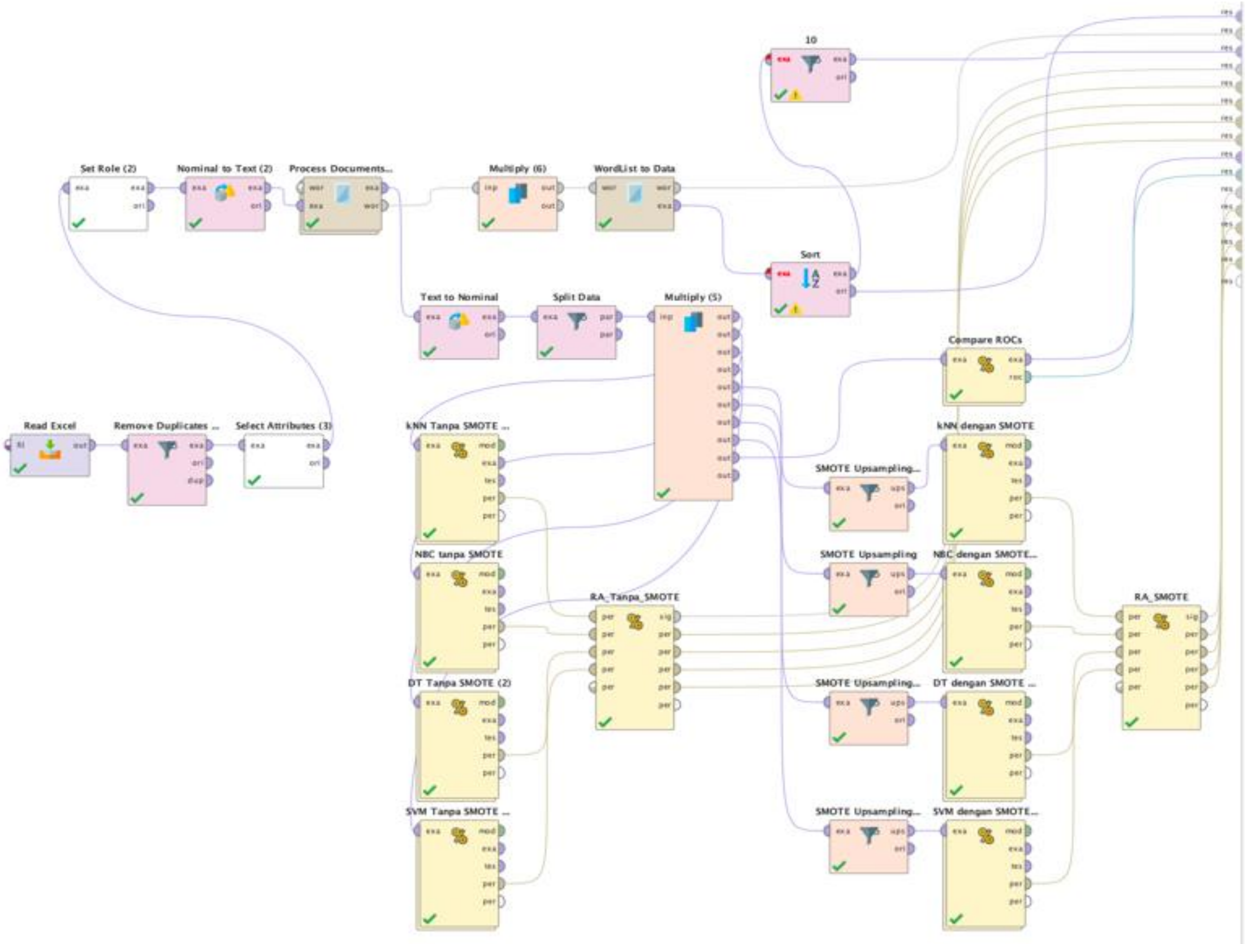

Figure 6. Implementation of SVM and DT Models in Rapidminer

Figure 6 shows the modeling process in Rapidminer. The testing phase involves partitioning the data into 70% for training and 30% for testing purposes while comparing the model performance with and without Synthetic Minority Over-sampling Technique (SMOTE). This systematic approach allows for assessing model robustness and generalization capabilities under different conditions, providing valuable insights into the impact of data imbalance on classification accuracy. By comparing SMOTE-enhanced models with those without SMOTE, researchers can discern the efficacy of employing oversampling techniques in addressing class imbalance issues and improving the overall performance of the classification model. Such rigorous testing procedures contribute to the refinement and optimization of the sentiment classification model, ensuring its reliability and applicability in real-world scenarios.

2.2.4 Evaluation

During the evaluation stage, the performance of the SVM and DT algorithms will be scrutinized using confusion matrices, which encompass metrics such as accuracy, precision, recall, F-measure, and Area Under the Curve (AUC). This comprehensive evaluation framework enables researchers to assess the effectiveness and reliability of the classification models in accurately distinguishing between positive and negative sentiments within the text data. By examining multiple performance metrics, researchers can gain a holistic understanding of the algorithms' strengths and limitations, facilitating informed decision-making regarding model selection and refinement. Such rigorous evaluation criteria contribute to the robustness and validity of the sentiment analysis results, ensuring their applicability and reliability in practical contexts.

2.2.5 Deployment

During the deployment stage, the performance of the best-performing model is recommended as superior for sentiment classification. Subsequently, within a contextual framework, it serves as valuable input for content creators in structuring storyboards to captivate viewer interest in line with exploratory and educational content preferences. This strategic integration of sentiment analysis findings into content creation processes enables creators to tailor their narratives and visual presentations effectively, maximizing viewer engagement and satisfaction. As a result, deploying the optimal sentiment classification model enhances the quality and relevance of content. It contributes to the overall success and impact of exploratory and educational video content on digital platforms.

3. RESULT AND DISCUSSION

Exploratory and educational content has garnered significant popularity and piqued viewer interest on the YouTube platform alongside entertainment content. This surge in demand can be attributed to the growing appetite for informative and enriching content experiences among digital audiences [49]. Additionally, the accessibility and diversity of exploratory and educational content cater to varying interests and preferences, providing viewers with valuable insights, knowledge, and entertainment [50]. As a result, creators and content producers increasingly leverage the appeal of exploratory and educational content to engage audiences, foster learning experiences, and cultivate loyal viewer communities.

Based on the analysis of response patterns in the form of emojis, it is evident that positive emojis dominate the feedback landscape. This prevalence of positive emojis signifies a favorable reception and engagement with the content, reflecting viewer satisfaction and appreciation. Such a trend underscores the effectiveness of the content in eliciting positive emotions and fostering a supportive and enthusiastic viewer community. Consequently, creators and content producers can leverage this insight to tailor their content strategies further and messaging to resonate with the prevailing positive sentiment among audiences, ultimately enhancing viewer satisfaction and loyalty.

Figure 7. Emoji and Words Cloud (Communalytic)

Figure 7 shows the emoji statistic of the dataset. Based on the statistical data of emoji usage within the dataset, it is evident that the ❤ emoji is the most frequently used, with a count of 34 occurrences, followed closely by the 👍 emoji at 31 occurrences. This distribution highlights the prevalence of positive sentiment conveyed through emojis, as evidenced by the predominance of smiling and heart emojis. The significant utilization of these emojis underscores the expression of appreciation, satisfaction, and positivity among viewers in response to the content. Such insights into emoji usage patterns provide valuable feedback for content creators, guiding them in understanding and catering to audience preferences and emotions, ultimately contributing to enhanced engagement and resonance with the target audience.

Furthermore, based on the analysis of frequently used words, it is evident that expressions of greetings and gratitude such as "(Salam)" and "(Terimaksih)" are among the most prevalent, with "(Salam)" appearing 120 times and "(Terimaksih)" 115 times in the dataset. Additionally, words associated with well-being and happiness, such as "(sehat)" and "(bahagia)," are also frequently utilized, suggesting a positive tone in the content and viewer interactions. Furthermore, terms related to rural settings and natural beauty, such as "(desa)" and "(indah)," feature prominently, reflecting an appreciation for the picturesque landscapes and simple living portrayed in the content. These findings provide valuable insights into the themes and sentiments resonating with the audience, guiding content creators in crafting engaging and relevant narratives that align with viewer preferences and emotions.

Based on emoji and word cloud analysis, it is possible to discern the intentions and interests of viewers regarding the content depicted in the video. These visual representations provide valuable insights into the audience's sentiments, preferences, and focal points within the video content, offering a comprehensive understanding of their engagement and interaction patterns. By analyzing the frequency and context of emojis and words used by viewers, content creators can tailor their strategies to align more closely with audience expectations and preferences, enhancing viewer engagement and satisfaction. In essence, leveraging emoji and word cloud analysis enables a deeper understanding of viewer perceptions and interests, facilitating more effective content creation and delivery strategies.

Furthermore, the results of testing the SVM and DT algorithms using SMOTE demonstrate commendable performance. This finding underscores the efficacy of employing oversampling techniques to address class imbalance issues and enhance the classification accuracy of sentiment analysis models. By mitigating the impact of data imbalance, SMOTE facilitates the creation of more robust and reliable classifiers, thereby improving the overall effectiveness of sentiment analysis algorithms. This validation of SMOTE-enhanced models reaffirms its utility as a valuable tool in data preprocessing for sentiment analysis tasks, contributing to advancing machine learning techniques in text classification applications.

SVM with SMOTE

```
PerformanceVector:
accuracy: 99.21% +/- 0.35% (micro average: 99.21%)
ConfusionMatrix:
True:   Negative        Positive
Negative:       1130    13
Positive:       5       1122
AUC (optimistic): 1.000 +/- 0.000 (micro average: 1.000) (positive class: Positive)
AUC: 1.000 +/- 0.000 (micro average: 1.000) (positive class: Positive)
AUC (pessimistic): 1.000 +/- 0.000 (micro average: 1.000) (positive class: Positive)
precision: 99.57% +/- 0.93% (micro average: 99.56%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       1130    13
Positive:       5       1122
recall: 98.85% +/- 0.84% (micro average: 98.85%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       1130    13
Positive:       5       1122
f_measure: 99.20% +/- 0.35% (micro average: 99.20%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       1130    13
Positive:       5       1122
```

DT with SMOTE

```
PerformanceVector:
accuracy: 98.33% +/- 0.88% (micro average: 98.33%)
ConfusionMatrix:
True:   Negative        Positive
Negative:       1122    25
Positive:       13      1110
AUC (optimistic): 0.996 +/- 0.004 (micro average: 0.996) (positive class: Positive)
AUC: 0.985 +/- 0.008 (micro average: 0.985) (positive class: Positive)
AUC (pessimistic): 0.974 +/- 0.014 (micro average: 0.974) (positive class: Positive)
precision: 98.84% +/- 0.94% (micro average: 98.84%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       1122    25
Positive:       13      1110
recall: 97.79% +/- 0.96% (micro average: 97.80%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       1122    25
Positive:       13      1110
f_measure: 98.31% +/- 0.89% (micro average: 98.32%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       1122    25
Positive:       13      1110
```

Figure 8. SVM and DT Performance Using SMOTE

Figure 8 shows the SVM and DT performance using SMOTE in Rapidminer. In addition, the SVM performance using SMOTE is as follows: an accuracy of 99.21% with a micro average of 99.21%, as indicated by the confusion matrix showing the precise classification of negative and positive classes, with negligible misclassifications. The area under the curve (AUC) demonstrates optimal performance, indicating high confidence in classifying positive instances. Furthermore, precision, recall, and f-measure metrics consistently exhibit high scores, with precision at 99.57%, recall at 98.85%, and f-measure at 99.20%. These results underscore the robustness and reliability of SVM with SMOTE in sentiment analysis tasks, reaffirming its suitability for real-world applications requiring accurate sentiment classification.

Furthermore, DT utilizing SMOTE demonstrates the following performance: an accuracy of 98.33% with a micro average of 98.33%, as observed in the confusion matrix depicting accurate classification of negative and positive instances with minimal misclassifications. The area under the curve (AUC) metrics exhibit robust performance, with optimistic, realistic, and pessimistic scenarios indicating high confidence in classifying positive instances. Precision, recall, and f-measure metrics consistently exhibit high scores, with precision at 98.84%, recall at 97.79%, and f-measure at 98.31%. These results underscore the efficacy of DT with SMOTE in sentiment analysis tasks, affirming its reliability and suitability for practical applications necessitating precise sentiment classification.

The deployment of the research findings entails the integration of the developed sentiment analysis models into practical applications to enhance decision-making processes and improve user experiences. By implementing the trained algorithms within digital platforms or software systems, stakeholders leverage real-time sentiment analysis insights to inform marketing strategies, customer engagement initiatives, and content curation efforts. Furthermore, using these models in sentiment monitoring tools can enable organizations to swiftly identify emerging trends, sentiment shifts, and potential crises, allowing for proactive interventions and strategic responses. Ultimately, deploying the research findings facilitates the translation of theoretical insights into tangible benefits, driving innovation and efficiency in various domains reliant on sentiment analysis for informed decision-making and enhanced user engagement.

A limitation of this research lies in the dataset derived from a single video content, thus limiting the scope of insights and generalizability of findings. However, opportunities for future development can be pursued by exploring additional video content from the exact location to obtain a more diverse range of insights. By expanding the dataset to encompass a broader spectrum of video content, researchers can enhance the robustness and applicability of their findings, enriching the understanding of sentiment analysis within specific contexts. This approach facilitates a more comprehensive examination of sentiment dynamics across various video genres and themes, thereby contributing to the advancement of sentiment analysis methodologies in digital content analysis.

4. CONCLUSION

In conclusion, the research findings, guided by the CRISP-DM framework, underscore the robust performance of both Support Vector Machine (SVM) and Decision Tree (DT) algorithms when integrated with Synthetic Minority Over-sampling Technique (SMOTE) in sentiment analysis tasks. The utilization of SMOTE effectively addresses class imbalance issues, resulting in a notable improvement in classification accuracy, with SVM achieving an accuracy of 98.33% and DT achieving an accuracy of 98.33%. Moreover, the comprehensive evaluation of SVM and DT algorithms, incorporating metrics such as precision, recall, and f-measure, further supports their efficacy in accurately classifying sentiment within text data. These insights significantly advance our understanding of sentiment analysis methodologies and provide valuable guidance for implementing sentiment classification models in real-world applications.

ACKNOWLEDGEMENTS

I sincerely thank the Tourism Department, the Faculty of Business Administration and Communication, and the Atma Jaya Catholic University of Indonesia.

REFERENCES

[1] P. Jayaswal and B. Parida, “Past, present and future of augmented reality marketing research: a bibliometric and thematic analysis approach,” *Eur. J. Mark.*, vol. 57, no. 9, pp. 2237–2289, Jan. 2023, doi: 10.1108/EJM-05-2022-0397.

[2] G. Rasool and A. Pathania, “Reading between the lines: untwining online user-generated content using sentiment analysis,” *J. Res. Interact. Mark.*, vol. 15, no. 3, pp. 401–418, Jan. 2021, doi: 10.1108/JRIM-03-2020-0045.

[3] A. Abid, P. Harrigan, and S. K. Roy, “Online relationship marketing through content creation and curation,” *Mark. Intell. Plan.*, vol. 38, no. 6, pp. 699–712, Jan. 2020, doi: 10.1108/MIP-04-2019-0219.

[4] R. Harakawa, T. Ogawa, and M. Haseyama, “Extracting Hierarchical Structure of Web Video Groups Based on Sentiment-Aware Signed Network Analysis,” *IEEE Access*, vol. 5, pp. 16963–16973, 2017, doi: 10.1109/ACCESS.2017.2741098.

[5] X. Yan, H. Xue, S. Jiang, and Z. Liu, “Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling,” *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2021.2000688.

[6] M. S. Viñán-Ludeña and L. M. de Campos, “Analyzing tourist data on Twitter: a case study in the province of Granada at Spain,” *J. Hosp. Tour. Insights*, vol. 5, no. 2, pp. 435–464, Jan. 2022, doi: 10.1108/JHTI-11-2020-0209.

[7] N. Amat-Lefort, F. Barravecchia, and L. Mastrogiacomo, “Quality 4.0: big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews,” *Int. J. Qual. Reliab. Manag.*, vol. 40, no. 4, pp. 990–1008, Jan. 2023, doi: 10.1108/IJQRM-01-2022-0024.

[8] I. Awajan, M. Mohamad, and A. Al-Quran, “Sentiment Analysis Technique and Neutrosophic Set Theory for Mining and Ranking Big Data from Online Reviews,” *IEEE Access*, vol. 9, pp. 47338–47353, 2021, doi: 10.1109/ACCESS.2021.3067844.

[9] R. C. Ho, M. S. Withanage, and K. W. Khong, “Sentiment drivers of hotel customers: a hybrid approach using unstructured data from online reviews,” *Asia-Pacific J. Bus. Adm.*, vol. 12, no. 3/4, pp. 237–250, Jan. 2020, doi: 10.1108/APJBA-09-2019-0192.

[10] R. Bringula, S. A. I. D. A. Ulfa, J. P. P. Miranda, and F. A. L. Atienza, “Text mining analysis on students’ expectations and anxieties towards data analytics course,” *Cogent Eng.*, vol. 9, no. 1, 2022, doi: 10.1080/23311916.2022.2127469.

[11] P. Sarin, A. K. Kar, and V. P. Ilavarasan, “Exploring engagement among mobile app developers – Insights from mining big data in user generated content,” *J. Adv. Manag. Res.*, vol. 18, no. 4, pp. 585–608, Jan. 2021, doi: 10.1108/JAMR-06-2020-0128.

[12] S. Song, S. “Brian” Park, and K. Park, “Thematic analysis of destination images for social media engagement marketing,” *Ind. Manag. Data Syst.*, vol. 121, no. 6, pp. 1375–1397, Jan. 2021, doi: 10.1108/IMDS-12-2019-0667.

[13] A. Singh, M. Jenamani, and J. Thakkar, “Do online consumer reviews help to evaluate the performance of automobile manufacturers?,” *J. Enterp. Inf. Manag.*, vol. 33, no. 5, pp. 1153–1198, Jan. 2020, doi: 10.1108/JEIM-09-2019-0292.

[14] Y. K. Oh and J. Yi, “Asymmetric effect of feature level sentiment on product rating: an application of bigram natural language processing (NLP) analysis,” *Internet Res.*, vol. 32, no. 3, pp. 1023–1040, Jan. 2022, doi: 10.1108/INTR-11-2020-0649.

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

[16] K. Suresh Kumar *et al.*, “Sentiment Analysis of Short Texts Using SVMs and VSMs-Based Multiclass Semantic Classification,” *Appl. Artif. Intell.*, vol. 38, no. 1, 2024, doi: 10.1080/08839514.2024.2321555.

[17] M. S. Hossain, H. Begum, M. A. Rouf, and M. M. I. Sabuj, “Investigation and prediction of users’ sentiment toward food delivery apps applying machine learning approaches,” *J. Contemp. Mark. Sci.*, vol. 6, no. 2, pp. 109–127, Jan. 2023, doi: 10.1108/JCMARS-12-2022-0030.

[18] I. Gurrib and F. Kamalov, “Predicting bitcoin price movements using sentiment analysis: a machine learning approach,” *Stud. Econ. Financ.*, vol. 39, no. 3, pp. 347–364, Jan. 2022, doi: 10.1108/SEF-07-2021-0293.

[19] H. He, G. Zhou, and S. Zhao, “Exploring E-Commerce Product Experience Based on Fusion Sentiment Analysis Method,” *IEEE Access*, vol. 10, no. October, pp. 110248–110260, 2022, doi: 10.1109/ACCESS.2022.3214752.

[20] M. A. Qureshi *et al.*, “Sentiment Analysis of Reviews in Natural Language: Roman Urdu as a Case Study,” *IEEE Access*, vol. 10, no. January, pp. 24945–24954, 2022, doi: 10.1109/ACCESS.2022.3150172.

[21] M. S. Hossain, M. F. Rahman, M. K. Uddin, and M. K. Hossain, “Customer sentiment analysis and prediction of halal restaurants using machine learning approaches,” *J. Islam. Mark.*, vol. 14, no. 7, pp. 1859–1889, Jan. 2023, doi: 10.1108/JIMA-04-2021-

0125.
[22] Y. A. Singgalen, "Penerapan Metode CRISP-DM dalam Klasifikasi Data Ulasan Pengunjung Destinasi Danau Toba Menggunakan Algoritma Naïve Bayes Classifier (NBC) dan Decision Tree (DT)," *J. Media Inform. Budidarma*, vol. 7, no. 3, pp. 1551–1562, 2023, doi: 10.30865/mib.v7i3.6461.
[23] Y. A. Singgalen, "Analisis Performa Algoritma NBC , DT , SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1634–1646, 2022, doi: 10.47065/bits.v4i3.2766.
[24] Y. A. Singgalen, "Culture and heritage tourism sentiment classification through cross-industry standard process for data mining," *Int. J. Basic Appl. Sci.*, vol. 12, no. 3, pp. 110–120, 2023.
[25] Y. A. Singgalen, "Sentiment Classification of Over-Tourism Issues in Responsible Tourism Content using Naïve Bayes Classifier," *J. Comput. Syst. Informatics*, vol. 5, no. 2, pp. 275–285, 2024, doi: 10.47065/josyc.v5i2.4904.
[26] Y. A. Singgalen, "Sentiment Classification of Robot Hotel Content using NBC and SVM Algorithm," *J. Comput. Syst. Informatics*, vol. 5, no. 2, pp. 442–453, 2024, doi: 10.47065/josyc.v5i2.4924.
[27] Y. A. Singgalen, "Sentiment Classification of Climate Change and Tourism Content Using Support Vector Machine," *J. Comput. Syst. Informatics*, vol. 5, no. 2, pp. 357–367, 2024, doi: 10.47065/josyc.v5i2.4908.
[28] 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.
[29] 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.
[30] Y. A. Singgalen, "Performance Evaluation of SVM Algorithm in Sentiment Classification : A Visual Journey of Wonderful Indonesia Content," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 4, pp. 2078–2087, 2024, doi: 10.30865/klik.v4i4.1709.
[31] Y. A. Singgalen, "Toxicity , topic , and sentiment analysis on the operation of coal-fired power plants content reviews," *J. Tek. Inform. C.I.T Medicom*, vol. 16, no. 1, pp. 45–57, 2024.
[32] Y. A. Singgalen, "Toxicity Analysis and Sentiment Classification of Wonderland Indonesia by Alffy Rev using Support Vector Machine," *J. Sist. Komput. dan Inform.*, vol. 5, no. 3, pp. 538–548, 2024, doi: 10.30865/json.v5i3.7563.
[33] H. Zhang, S. Sun, Y. Hu, J. Liu, and Y. Guo, "Sentiment Classification for Chinese Text Based on Interactive Multitask Learning," *IEEE Access*, vol. 8, pp. 129626–129635, 2020, doi: 10.1109/ACCESS.2020.3007889.
[34] X. Cao, J. Yu, and Y. Zhuang, "Injecting User Identity Into Pretrained Language Models for Document-Level Sentiment Classification," *IEEE Access*, vol. 10, pp. 30157–30167, 2022, doi: 10.1109/ACCESS.2022.3158975.
[35] Y. Chen, L. Kong, Y. Wang, and D. Kong, "Multi-Grained Attention Representation with ALBERT for Aspect-Level Sentiment Classification," *IEEE Access*, vol. 9, pp. 106703–106713, 2021, doi: 10.1109/ACCESS.2021.3100299.
[36] F. Z. Ruskanda, M. R. Abiwardani, R. Mulyawan, I. Syafalni, and H. T. Larasati, "Quantum-Enhanced Support Vector Machine for Sentiment Classification," *IEEE Access*, vol. 11, no. July, pp. 87520–87532, 2023, doi: 10.1109/ACCESS.2023.3304990.
[37] K. Xu, H. Zhao, and T. Liu, "Aspect-Specific Heterogeneous Graph Convolutional Network for Aspect-Based Sentiment Classification," *IEEE Access*, vol. 8, pp. 139346–139355, 2020, doi: 10.1109/ACCESS.2020.3012637.
[38] L. Xiaoyan and R. C. Raga, "BiLSTM Model With Attention Mechanism for Sentiment Classification on Chinese Mixed Text Comments," *IEEE Access*, vol. 11, no. February, pp. 26199–26210, 2023, doi: 10.1109/ACCESS.2023.3255990.
[39] K. Jahanbin and M. A. Z. Chahooki, "Aspect-Based Sentiment Analysis of Twitter Influencers to Predict the Trend of Cryptocurrencies Based on Hybrid Deep Transfer Learning Models," *IEEE Access*, vol. 11, no. November, pp. 121656–121670, 2023, doi: 10.1109/ACCESS.2023.3327060.
[40] E. Saltman, F. Kooti, and K. Vockery, "New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis," *Stud. Confl. Terror.*, vol. 46, no. 9, pp. 1547–1574, 2023, doi: 10.1080/1057610X.2021.1888404.
[41] N. S. Mohd Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," *IEEE Access*, vol. 9, no. Ml, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.
[42] U. Sehar, S. Kanwal, K. Dashtipur, U. Mir, U. Abbasi, and F. Khan, "Urdu Sentiment Analysis via Multimodal Data Mining Based on Deep Learning Algorithms," *IEEE Access*, vol. 9, pp. 153072–153082, 2021, doi: 10.1109/ACCESS.2021.3122025.
[43] P. Thiengburanathum and P. Charoenkwan, "SETAR: Stacking Ensemble Learning for Thai Sentiment Analysis Using RoBERTa and Hybrid Feature Representation," *IEEE Access*, vol. 11, no. July, pp. 92822–92837, 2023, doi: 10.1109/ACCESS.2023.3308951.
[44] Y. Zheng, Y. Long, and H. Fan, "Identifying Labor Market Competitors with Machine Learning Based on Maimai Platform," *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2064047.
[45] S. Mukhopadhyay, T. Jain, S. Modgil, and R. K. Singh, "Social media analytics in tourism: a review and agenda for future research," *Benchmarking An Int. J.*, vol. 30, no. 9, pp. 3725–3750, Jan. 2023, doi: 10.1108/BIJ-05-2022-0309.
[46] S. Sekar, S. Edakkat Subhakaran, and D. Chattopadhyay, "Unlocking the voice of employee perspectives: exploring the relevance of online platform reviews on organizational perceptions," *Manag. Decis.*, vol. 61, no. 11, pp. 3408–3429, Jan. 2023, doi: 10.1108/MD-11-2022-1509.
[47] B.-H. Leem and S.-W. Eum, "Using text mining to measure mobile banking service quality," *Ind. Manag. Data Syst.*, vol. 121, no. 5, pp. 993–1007, Jan. 2021, doi: 10.1108/IMDS-09-2020-0545.
[48] V. Gupta, S. Singh, and S. S. Yadav, "The impact of media sentiments on IPO underpricing," *J. Asia Bus. Stud.*, vol. 16, no. 5, pp. 786–801, Jan. 2022, doi: 10.1108/JABS-10-2020-0404.
[49] L. Ryan Bengtsson and J. Edlom, "Commodifying participation through choreographed engagement: the Taylor Swift case," *Arts Mark.*, vol. 13, no. 2, pp. 65–79, Jan. 2023, doi: 10.1108/AAM-07-2022-0034.
[50] P. Rohde and G. Mau, "'It's selling like hotcakes': deconstructing social media influencer marketing in long-form video content on youtube via social influence heuristics," *Eur. J. Mark.*, vol. 55, no. 10, pp. 2700–2734, Jan. 2021, doi: 10.1108/EJM-06-2019-0530.