KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 5, April 2024, Hal 2485-2494
DOI 10.30865/klik.v4i5.1828
https://djournals.com/klik

Comparative Analysis of DT and SVM Model Performance with SMOTE in Sentiment Classification

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−This research investigates the efficacy of employing the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to analyze sentiment classification models. The study focuses on evaluating the performance of Decision Trees (DT) and Support Vector Machine (SVM) models integrated with the Synthetic Minority Over-sampling Technique (SMOTE) across various performance metrics, including accuracy, precision, recall, f-measure, and Area Under the Curve (AUC). Using CRISP-DM, the research ensures a systematic data preprocessing, modeling, and evaluation approach. The findings reveal that both DT and SVM models with SMOTE achieve high accuracy rates, with DT yielding an accuracy of 98.37% +/- 0.48% and SVM achieving 98.91% +/- 0.59%. These models effectively distinguish between positive and negative sentiments, as precision, recall, and f-measure scores indicate. Additionally, the AUC scores underscore the robustness of the models in sentiment analysis tasks. These results highlight the potential of CRISP-DM as a structured methodology for sentiment classification research, providing insights into the performance of different machine learning algorithms in handling imbalanced datasets. Based on these findings, it is recommended that future studies further explore the application of CRISP-DM in sentiment analysis tasks and investigate the scalability of DT and SVM models with SMOTE in larger datasets.

Keywords: DT; Sentiment Classification; SMOTE; SVM; Model Performance

1. INTROsDUCTION

The proliferation of digital media and sharing platforms has significantly enhanced opportunities for broadening insights, particularly in comprehending societal livelihoods and settlements through digitally disseminated content by creators. This evolution has ushered in a paradigm shift in how information is accessed and shared, revolutionizing the dissemination of knowledge about diverse cultural and societal dynamics [1]–[4]. Consequently, digital content creators are pivotal in fostering a deeper understanding of various facets of human existence and community structures, fostering a more interconnected global society [5], [6]. In essence, the synergy between digital media and content creators fosters greater awareness and empathy toward the multifaceted nature of human civilization and settlement patterns.

The advancement of digital technology has not only escalated participation in the virtual realm but has also catalyzed the creativity of content creators. This transformative phenomenon has democratized content creation, empowering individuals to produce and distribute innovative digital media across various online platforms [7]–[12]. Consequently, a burgeoning array of creative content emerges, ranging from immersive storytelling to interactive multimedia experiences, enriching the digital landscape and captivating audiences worldwide [13], [14]. Thus, the symbiotic relationship between technological advancements and content creators' ingenuity underscores digital innovation's pivotal role in shaping contemporary cultural expressions and fostering a vibrant digital ecosystem.

Digital video works are classified based on entertainment, education, politics, ecology, and socio-cultural themes. This categorization serves as a framework for understanding the diverse purposes and intentions behind digital video content creation [15]–[19]. In entertainment, videos aim to captivate and engage audiences through storytelling, humor, and visual spectacle [20]–[24]. Conversely, educational videos leverage the medium to impart knowledge, skills, and insights on various subjects, catering to formal and informal learning environments [25]–[27]. Moreover, political videos utilize the power of visual storytelling to communicate messages, advocate for causes, and mobilize public opinion on socio-political issues [28]–[31]. Thus, the classification of digital videos according to thematic elements reflects the multifaceted nature of contemporary digital media and its role in shaping discourse, entertainment, and education in the digital age.

This research aims to analyze digital content based on viewer reviews, employing the CRISP-DM methodology utilizing Decision Trees (DT) and Support Vector Machines (SVM) models. By leveraging these analytical techniques, the study seeks to extract meaningful insights from large datasets of viewer feedback, enabling a comprehensive understanding of audience preferences, sentiments, and engagement patterns; by applying CRISP-DM, a structured approach to data mining, the research endeavors to uncover actionable insights inform content creators and digital platforms to optimize the offerings to cater to viewer preferences better and enhance the overall user experience [32]–[34]. Thus, utilizing DT and SVM models within the CRISP-DM framework represents a methodologically robust approach to digital content analysis, offering valuable insights for academic research and practical applications in the digital media industry.

The urgency of this research lies in its potential to address pressing challenges and capitalize on emerging opportunities within the digital landscape. By delving into the intricate dynamics of viewer feedback analysis through advanced methodologies such as CRISP-DM, Decision Trees (DT), and Support Vector Machines (SVM), this study aims to provide actionable insights to inform strategic decision-making processes across various sectors [30], [35]–[37].

Furthermore, in an era characterized by rapid technological advancements and evolving consumer preferences, deciphering and leveraging viewer sentiments is paramount for content creators, digital platforms, and other stakeholders seeking to stay ahead in a competitive marketplace [38]. Hence, the timeliness and significance of this research lie in its potential to drive innovation, enhance user engagement, and foster sustainable growth within the digital ecosystem.

This research's theoretical and practical implications are profound, offering valuable insights for academia and industry. Theoretically, the utilization of advanced methodologies such as CRISP-DM, Decision Trees (DT), and Support Vector Machines (SVM) contributes to the ongoing discourse surrounding data mining techniques and applications in analyzing digital content. By demonstrating the efficacy of these methodologies in extracting actionable insights from large datasets of viewer reviews, this research enriches the body of knowledge within the fields of data science, digital media studies, and consumer behavior analysis [39], [40]. On a practical level, the findings of this research have significant implications for content creators, digital platforms, and other stakeholders operating within the digital media landscape. By informing strategic decision-making processes and content optimization strategies, the insights gleaned from this study enhance user engagement, improve content relevance, and ultimately drive business growth [38], [41], [42]. This research's theoretical and practical implications underscore its relevance and potential impact in academic and industry contexts.

The limitation of this research lies in its reliance on a specific dataset, centered around a single video with the identifier "rL4Hb9hWEhA," which garnered 1,392,613 views and elicited 3,371 comments since February 10, 2024. While this dataset offers valuable insights into viewer engagement and feedback for the particular video under scrutiny, its narrow focus may limit the generalizability of findings to broader contexts within the digital media landscape. Consequently, caution must be exercised when extrapolating conclusions or making broader assertions based solely on the characteristics of this singular dataset. However, despite this limitation, the detailed analysis of this case provides a foundation for further explorations and comparative studies, thereby contributing to a deeper understanding of viewer behavior and content dynamics within digital platforms.

In considering similar research and recommendations for further investigation, it is imperative to explore comparative studies that examine a broader spectrum of digital content and viewer interactions across diverse platforms. Expanding the scope beyond the confines of a single video dataset, this research elucidates commonalities and disparities in viewer engagement patterns, sentiment analysis, and content preferences across different genres, formats, and audience demographics. Furthermore, exploring the efficacy of alternative methodologies and machine learning algorithms in analyzing digital content could provide valuable insights into the robustness and generalizability of findings. Thus, by fostering interdisciplinary collaborations and leveraging diverse datasets, future research endeavors advance our understanding of digital media consumption behaviors and inform evidence-based strategies for content creators and platform administrators.

2. RESEARCH METHODOLOGY

A rigorous analysis approach is delineated in the methodology section, incorporating both the Gap Analysis method and the CRISP-DM framework, which systematically dissects each stage. This structured methodology not only facilitates a comprehensive understanding of the research landscape but also enables the identification of discrepancies between the current and desired states, thus fostering informed decision-making processes. By integrating these analytical tools, the study enhances its methodological robustness and ensures a systematic exploration of the research problem, ultimately contributing to advancing knowledge in the field.

2.1 Gap Analysis

Gap analysis is essential for identifying pertinent topics in sentiment analysis using Decision Trees (DT) and Support Vector Machines (SVM) within the context of digital content and sharing media platforms. By thoroughly examining existing literature and research, this research pinpoints areas where current knowledge falls short and additional investigation is warranted. This process enables scholars to identify gaps in understanding, such as unexplored aspects of sentiment analysis methodologies, underrepresented content genres, or overlooked audience demographics. Through meticulous gap analysis, this research delineates avenues for future inquiry that promise to enrich the field by addressing unanswered questions and advancing methodological frameworks. Consequently, the systematic identification of research gaps serves as a cornerstone for guiding the development of innovative approaches and generating valuable insights in sentiment analysis within the dynamic landscape of digital content and sharing media platforms.

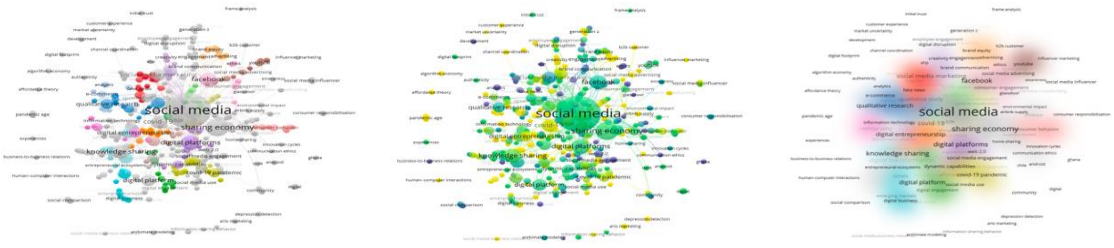


Figure 1. Gap Analysis using Vosviewer

Figure 1 shows the network, density, and overlay visualization. Based on the results of gap identification, it is evident that digital content is intrinsically linked to social media, digital platforms, and knowledge-sharing ecosystems. These interconnected domains are integral conduits for creating, disseminating, and consuming digital content across diverse audiences and contexts. Moreover, the symbiotic relationship between digital content and these platforms underscores the pivotal role in shaping contemporary communication patterns, information dissemination dynamics, and knowledge exchange mechanisms. Consequently, understanding the intricate interplay between digital content and its associated platforms is imperative for devising effective strategies to harness the full potential of digital media in facilitating meaningful interactions, fostering community engagement, and driving knowledge creation and dissemination initiatives in the digital age.

A disparity is evident among knowledge sharing, digital platforms, video content, and sentiment analysis. While these areas are crucial in shaping the contemporary digital landscape, the integration and synergy remain underexplored. Despite the growing recognition of the significance of knowledge-sharing platforms, digital platforms, and video content in facilitating information exchange and fostering community engagement, there is a notable lack of comprehensive research that examines the intersectionality between these domains and sentiment analysis methodologies. Consequently, bridging this gap through interdisciplinary investigations unlocks new insights into audience behaviors, content dynamics, and sentiment patterns within the digital realm, paving the way for more nuanced understandings and informed digital media research and practice strategies.

This research proposes the implementation of Decision Trees (DT) and Support Vector Machines (SVM) models within the CRISP-DM framework for sentiment analysis. By leveraging these advanced analytical techniques, the study aims to enhance the accuracy and efficiency of sentiment analysis processes, thereby enabling a more nuanced understanding of audience perceptions and reactions towards digital content. Through the systematic application of DT and SVM models within the CRISP-DM methodology, the research endeavors to uncover actionable insights to inform content creators, digital platforms, and other stakeholders in optimizing the strategies to better resonate with audience sentiments. Consequently, integrating these methodologies holds promise for advancing sentiment analysis methodologies and facilitating evidence-based decision-making in the digital media landscape.

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

The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework comprises five key stages: business understanding, data understanding, modeling, evaluation, and deployment. This structured approach provides a systematic and iterative methodology for conducting data mining projects, ensuring that all process aspects are carefully considered and executed. Beginning with the business understanding stage, stakeholders define project objectives and requirements, laying the groundwork for subsequent data collection and analysis. The data understanding stage involves exploring and familiarizing oneself with the dataset, identifying relevant variables, and assessing data quality. Subsequently, in the modeling stage, various data mining techniques, such as Decision Trees or Support Vector Machines, are applied to develop predictive or descriptive models. These models are then evaluated in the next stage to determine the effectiveness and accuracy in meeting the project objectives. Finally, successful models are deployed in real-world applications, with ongoing monitoring and refinement to ensure continued relevance and effectiveness. In conclusion, the CRISP-DM framework provides a comprehensive and structured approach to data mining projects, facilitating the successful implementation of data-driven solutions across various domains and industries.

Figure 2. Cross-Industry Standard Process for Data Mining (CRISP-DM) Framework

Figure 2 shows the CRISP-DM framework. The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework is highly pertinent to this research due to its comprehensive consideration of video content context and research constraints. Given the complexity of analyzing digital video content and the need to navigate specific research parameters, CRISP-DM offers a structured approach that ensures systematic exploration and utilization of available data. This research effectively addresses sentiment analysis's intricacies in video content by adhering to the stages of business understanding, data understanding, modeling, evaluation, and deployment. Moreover, CRISP-DM facilitates the incorporation of relevant contextual factors and research limitations, thereby enhancing the rigor and applicability of the study's findings. In essence, using CRISP-DM in this research underscores its relevance and effectiveness in guiding data mining projects within the dynamic landscape of digital media analysis.

The limitations of the CRoss-Industry Standard Process for Data Mining (CRISP-DM) in this research lie in its reliance on dataset size and the operationalization of Decision Trees (DT) and Support Vector Machines (SVM) models supported by the Synthetic Minority Over-sampling Technique (SMOTE) operator. While CRISP-DM provides a structured framework for conducting data mining projects, its effectiveness is hindered by constraints related to the availability and quality of datasets and the operational intricacies of machine learning algorithms. In particular, the dependency on dataset size may limit the generalizability of findings. At the same time, the operationalization of DT and SVM models supported by the SMOTE operator may introduce complexities in model implementation and interpretation.

Despite these limitations, CRISP-DM remains a valuable tool for guiding data mining endeavors, albeit necessitating careful consideration and adaptation to address specific research challenges and constraints.

2.2.1 Business Understanding

During the business understanding stage, it is essential to identify the characteristics of the video content to be analyzed. This research is confined to analyzing digital video content, leveraging statistical data from viewership, encompassing 1,392,613 views as of February 10, 2024, and comments totaling 3,371. The study aims to gain insights into viewer engagement, sentiment, and preferences regarding the analyzed video content by focusing on these metrics. Consequently, a thorough understanding of the content's attributes and audience interactions at this initial stage sets the foundation for subsequent data collection, analysis, and interpretation, ensuring the relevance and effectiveness of the research outcomes.

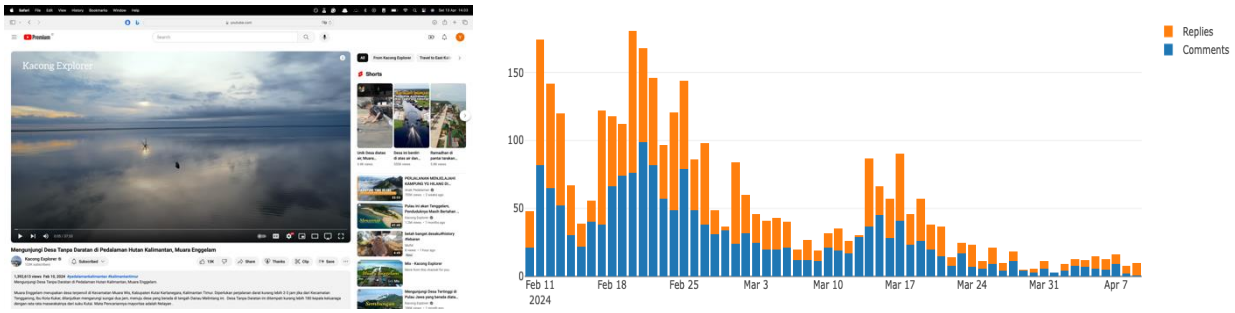


Figure 3. Content Video and Post-Per-Day Statistic

Figure 3 shows the video and post-per-day statistics. The Kacong Explorer channel provides the following description of the video content: "Muara Enggelam is a remote village in the Muara Wis District, Kutai Kartanegara Regency, East Kalimantan. It requires a land journey of approximately 2-3 hours from the Tenggarong District, the capital of Kutai Kartanegara, followed by a two-hour boat ride to reach the village in the middle of Lake Melintang. This landless village is inhabited by approximately 180 families, predominantly from the Kutai tribe. The primary livelihood of its inhabitants is fishing." This detailed depiction offers insights into the geographical and sociocultural aspects of Muara Enggelam, providing viewers with a comprehensive understanding of the village's location, demographics, and leading economic activities. Based on the post-per-day statistics provided, it is evident that there were fluctuations in the frequency of posts across different dates in February 2024. The data indicates that February 21st had the highest number of posts with 99, followed closely by February 11th and 22nd with 82 posts each.

Conversely, February 13th had the lowest number of posts, with 52, suggesting a notable decrease in activity on that particular date. These fluctuations may reflect variations in user engagement, content relevance, or external factors influencing posting behaviors within the analyzed period. Therefore, a detailed analysis of these trends offers valuable insights into audience dynamics and content consumption patterns, informing strategic decisions for content creators and digital platforms.

Based on the data of the top ten posters, it is apparent that @KacongExplorer dominates with 390 posts, indicating a significant presence and likely a leading role in disseminating content related to the discussed topic. Meanwhile, the remaining posters, such as @tengkuferdiansyah8617 and @IdaFarida-yr4dc, have contributed significantly fewer posts, suggesting a comparatively lesser influence or engagement within the community. This distribution of posting activity underscores the prominence of @KacongExplorer as a pivotal contributor to the discourse surrounding the subject matter, potentially serving as a central source of information or discussion platform for interested individuals.

Figure 4. Top Ten Poster (Communalytic)

Figure 4 shows the statistics of the top ten posters. Based on the data of the top-ten posters, it is evident that the channel owner is highly active in responding to viewer comments. With the channel owner, presumably @KacongExplorer, posting 390 times, significantly more than any other contributor, it suggests a proactive engagement with the audience. This level of responsiveness fosters community and interaction, enhancing viewer satisfaction and loyalty. Consequently, such dedicated interaction contributes to the channel's credibility and popularity, reinforcing its position as a prominent figure within the digital content landscape.

Henceforth, it is discernible that the context of the content to be analyzed is associated with exploring the Muara Enggelam location. As described in the information provided by the Kacong Explorer channel, the content revolves around the remote village of Muara Enggelam, detailing its geographical features, demographics, and socio-economic aspects. This thematic coherence suggests a focused examination of the village's dynamics, potentially encompassing cultural heritage, community livelihoods, and environmental sustainability. Consequently, the alignment between the content context and the analytical scope enhances the relevance and coherence of the research endeavor, facilitating a comprehensive exploration of Muara Enggelam's significance within the digital media landscape.

2.2.2 Data Understanding

During the data understanding stage, the process entails data cleaning and extraction before proceeding to model testing. This crucial phase involves identifying and rectifying any inconsistencies, errors, or missing values within the dataset to ensure its accuracy and reliability for subsequent analysis. Additionally, data extraction involves selecting and transforming relevant variables or features essential for model development and evaluation. This research mitigates potential biases or inaccuracies by meticulously cleaning and extracting data at this initial stage, thus laying a solid foundation for robust model testing and validation. Consequently, the diligent execution of data understanding processes enhances the integrity and validity of research outcomes, facilitating informed decision-making and actionable insights.

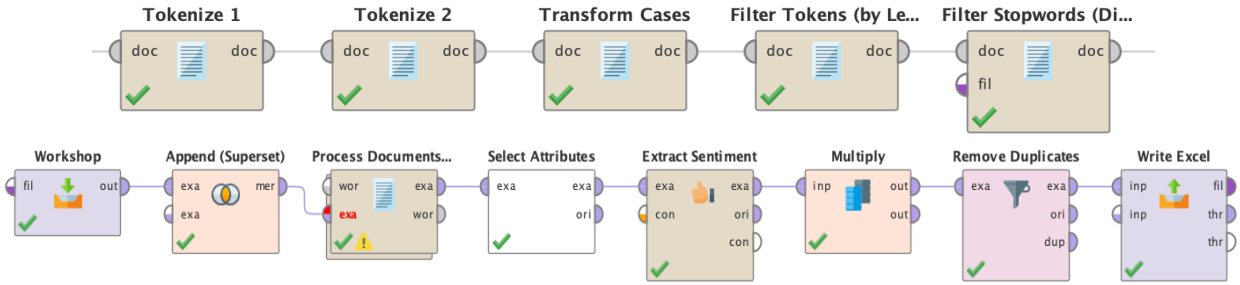


Figure 5. Data Cleaning and Extract Sentiment

Figure 5 shows the data cleaning and extract sentiment process in Rapidminer. The cleaned data, processed through operators such as tokenize, transform cases, filter tokens, and stopwords, undergoes extraction to obtain string scores. This meticulous preprocessing, involving tokenization, normalization of cases, removal of irrelevant tokens, and elimination of stopwords, enhances the quality and relevance of the extracted strings. By employing these operators, this research ensures that the extracted data aligns with the research objectives and is conducive to meaningful analysis. Consequently, the systematic application of these preprocessing techniques facilitates the extraction of string scores essential for subsequent sentiment analysis and model evaluation, thus contributing to the overall rigor and validity of the research outcomes.

Based on the string scores, the review data is classified into negative and positive classes, facilitating its progression to the modeling process. This classification enables data segmentation based on sentiment polarity, distinguishing between negative and positive sentiments expressed within the reviews. By categorizing the data into these classes, this research effectively trains and evaluates sentiment analysis models to accurately predict sentiment labels for unseen data. Consequently, using string scores for data classification is a crucial preparatory step in modeling, laying the groundwork for robust sentiment analysis and insightful interpretation of digital content.

2.2.3 Modeling

During the modeling stage, algorithm performance is evaluated using both SMOTE and non-SMOTE operators. This comparative analysis allows for assessing the effectiveness of SMOTE in addressing class imbalance and enhancing model generalization. By testing the algorithms under both conditions, this research determines whether the application of SMOTE improves the classification accuracy and robustness of the models. This systematic evaluation enables informed decision-making regarding selecting the most suitable algorithm configuration for sentiment analysis tasks, thereby ensuring the reliability and effectiveness of the modeling process.

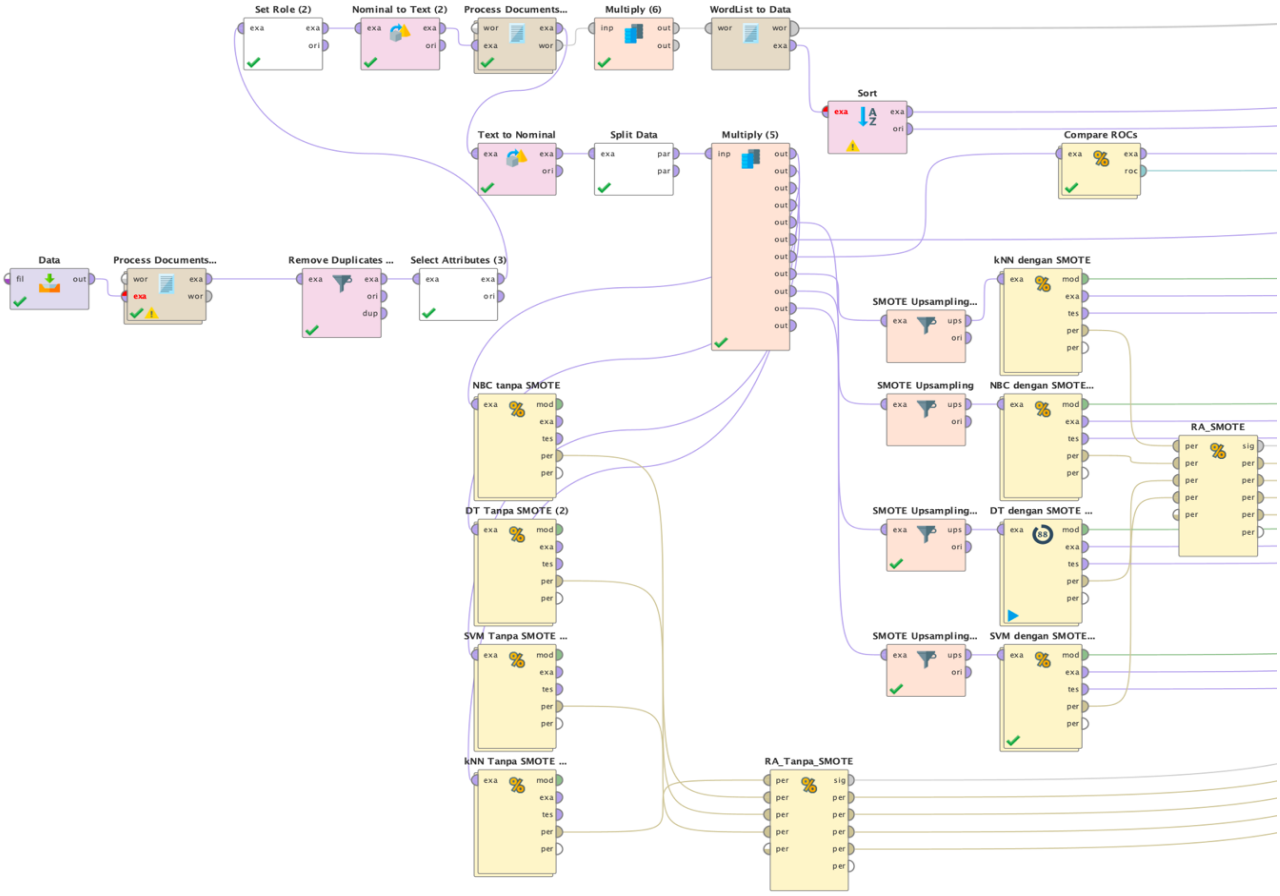


Figure 6. Implementation of DT and SVM Models in Rapidminer

Figure 6 shows the implementation of the DT and SVM models with and without SMOTE. Based on the results of model performance testing, various metrics such as accuracy, precision, recall, F-measure, and AUC are interpreted. These metrics provide quantitative measures of the model's effectiveness in correctly classifying instances, capturing the trade-offs between accurate positive and false favorable rates and overall model performance. By analyzing these metrics comprehensively, this research gains valuable insights into the strengths and limitations of the sentiment analysis models, enabling informed decisions regarding model selection and optimization strategies. Consequently, interpreting these performance metrics is crucial in evaluating the reliability and efficacy of sentiment analysis algorithms in digital content analysis.

In this study, the data partitioning scheme entails allocating 70% of the dataset for training and 30% for testing. The difference in the distribution of training and testing data proportions is a crucial factor influencing model performance evaluation. By systematically adjusting the ratio between training and testing data, this research assesses the model's ability to generalize to unseen data and its robustness under varying conditions. This approach enables a comprehensive examination of model performance across different data distributions, thereby enhancing the reliability and validity of the study's findings.

2.2.4 Evaluation

Evaluation is conducted by comparing the accuracy, precision, recall, F-measure, and AUC values between Decision Trees (DT) and Support Vector Machines (SVM), both with and without the application of the Synthetic Minority Over-sampling Technique (SMOTE). This comparative analysis allows for an in-depth assessment of the performance of each algorithm under different conditions, considering the impact of class imbalance mitigation techniques. By examining these metrics across various scenarios, this research identifies each algorithm's strengths and weaknesses and determines the most suitable approach for sentiment analysis tasks. Consequently, this systematic evaluation enhances the understanding of algorithmic performance and informs the selection of optimal models for digital content analysis.

Higher accuracy and AUC values indicate a model's strong performance and recommend it as a relevant classifier in sentiment classification based on the volume of review data. These metrics are reliable indicators of a model's ability to correctly classify instances and distinguish between positive and negative sentiments within the dataset. Therefore, models demonstrating elevated accuracy and AUC values are deemed more suitable for sentiment analysis tasks, as they offer greater confidence in the predictive capabilities and generalization to unseen data. Consequently, prioritizing models with high accuracy and AUC values ensures the selection of robust and effective classifiers for digital content analysis.

2.2.5 Deployment

The deployment stage of this research involves implementing and integrating the developed sentiment analysis models into practical applications or systems. During this phase, the models are deployed to real-world environments, allowing them to analyze and classify sentiment in digital content streams in real-time. This deployment enables stakeholders to leverage the insights generated by the models to make informed decisions, enhance user experiences, or automate processes related to sentiment analysis. Consequently, the deployment stage represents the culmination of the research efforts, transitioning theoretical models into actionable tools that contribute to advancing sentiment analysis practices in digital content analysis.

Based on the evaluation results of algorithm or classification model performance, it is recommended that media-sharing platform providers consider integrating real-time sentiment identification features into the applications. This addition would serve as a valuable tool for content creators, providing them with immediate feedback on the sentiment of the content. By offering insights into the audience's reactions, content creators understand the impact of the videos or content and make informed decisions to enhance quality. Consequently, implementing real-time sentiment analysis features fosters a more engaging and meaningful content creation process, ultimately benefiting content creators and platform users.

3. RESULT AND DISCUSSION

Based on the results of implementing the Decision Trees (DT) model and Synthetic Minority Over-sampling Technique (SMOTE), a comparison of performance before and after using SMOTE was conducted. This comparative analysis allows for an evaluation of the effectiveness of SMOTE in addressing class imbalance and improving the model's predictive capabilities. By examining the performance metrics such as accuracy, precision, recall, F-measure, and AUC before and after applying SMOTE, insights were gained into the impact of class rebalancing techniques on the model's ability to classify instances accurately. Consequently, this comparison facilitates a comprehensive understanding of the benefits of incorporating SMOTE into the modeling process, aiding in selecting optimal strategies for addressing class imbalance in sentiment analysis tasks.

The performance of Decision Trees (DT) without utilizing the Synthetic Minority Over-sampling Technique (SMOTE) reveals notable accuracy, precision, recall, and f-measure metrics. Specifically, the accuracy is reported at 98.73% +/- 0.32%, with a micro average of 98.73%. Similarly, the precision, recall, and f-measure exhibit high values of 98.81% +/- 0.29%, 99.91% +/- 0.19%, and 99.36% +/- 0.16%, respectively, underscoring the model's capability to classify instances accurately. Furthermore, the area under the curve (AUC) values, including optimistic and pessimistic scenarios, further validate the model's resilience in distinguishing between negative and positive classes. Hence, the collective performance metrics underscore the effectiveness of the DT model in sentiment analysis tasks, highlighting its potential for real-world applications in digital content analysis.

DT without SMOTE

```
PerformanceVector:
accuracy: 98.73% +/- 0.32% (micro average: 98.73%)
ConfusionMatrix:
True:   Negative        Positive
Negative:       14      2
Positive:       27      2239
AUC (optimistic): 0.999 +/- 0.001 (micro average: 0.999) (positive class: Positive)
AUC: 0.670 +/- 0.084 (micro average: 0.670) (positive class: Positive)
AUC (pessimistic): 0.340 +/- 0.168 (micro average: 0.340) (positive class: Positive)
precision: 98.81% +/- 0.29% (micro average: 98.81%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       14      2
Positive:       27      2239
recall: 99.91% +/- 0.19% (micro average: 99.91%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       14      2
Positive:       27      2239
f_measure: 99.36% +/- 0.16% (micro average: 99.36%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       14      2
Positive:       27      2239
```

SVM without SMOTE

```
PerformanceVector:
accuracy: 98.16% +/- 0.19% (micro average: 98.16%)
ConfusionMatrix:
True:   Negative        Positive
Negative:       0       1
Positive:       41      2240
AUC (optimistic): 0.682 +/- 0.195 (micro average: 0.682) (positive class: Positive)
AUC: 0.682 +/- 0.195 (micro average: 0.682) (positive class: Positive)
AUC (pessimistic): 0.681 +/- 0.195 (micro average: 0.681) (positive class: Positive)
precision: 98.20% +/- 0.14% (micro average: 98.20%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       0       1
Positive:       41      2240
recall: 99.96% +/- 0.14% (micro average: 99.96%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       0       1
Positive:       41      2240
f_measure: 99.07% +/- 0.09% (micro average: 99.07%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       0       1
Positive:       41      2240
```

Figure 7. Performance of DT and SVM without SMOTE

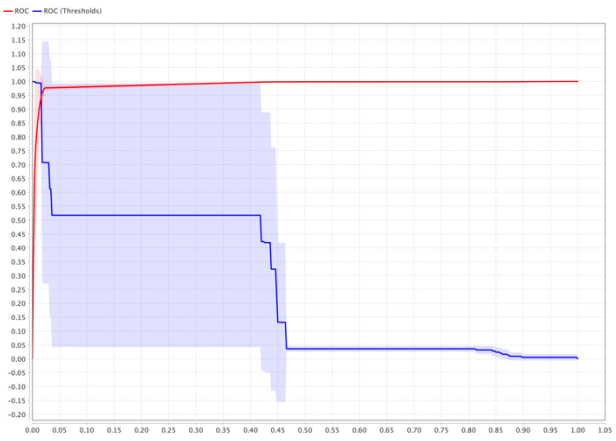
Figure 7 shows the DT and SVM without SMOTE. Moreover, the performance of the Support Vector Machine (SVM) in the absence of the Synthetic Minority Over-sampling Technique (SMOTE) demonstrates noteworthy accuracy, precision, recall, and f-measure metrics. Notably, the accuracy is recorded at 98.16% +/- 0.19%, with a micro average of 98.16%. Additionally, the precision, recall, and f-measure exhibit commendable values of 98.20% +/- 0.14%, 99.96% +/- 0.14%, and 99.07% +/- 0.09%, respectively, highlighting the model's proficiency in accurately classifying instances. The area under the curve (AUC) values further affirm the model's robustness in distinguishing between negative and positive classes. Despite encountering misclassifications, particularly within the positive class, the SVM model demonstrates vital performance metrics, implying its suitability for sentiment analysis tasks in the absence of SMOTE.

The evaluation results of the Decision Trees (DT) and Support Vector Machine (SVM) models without the Synthetic Minority Over-sampling Technique (SMOTE) indicate low values of the Area Under the Curve (AUC). This outcome suggests that the models exhibit suboptimal performance in distinguishing between positive and negative classes. Despite achieving high accuracy, precision, recall, and f-measure values, the low AUC values imply a deficiency in the

model's ability to correctly classify instances, particularly in scenarios where class imbalance exists. Consequently, addressing this limitation is imperative to enhance the models' effectiveness in sentiment analysis tasks, potentially by implementing class rebalancing techniques or other model optimization strategies.

DT with SMOTE

```
PerformanceVector:
accuracy: 98.37% +/- 0.48% (micro average: 98.37%)
ConfusionMatrix:
True:   Negative        Positive
Negative:       2222    54
Positive:       19      2187
AUC (optimistic): 0.998 +/- 0.002 (micro average: 0.998) (positive class: Positive)
AUC: 0.989 +/- 0.004 (micro average: 0.989) (positive class: Positive)
AUC (pessimistic): 0.980 +/- 0.007 (micro average: 0.980) (positive class: Positive)
precision: 99.14% +/- 0.71% (micro average: 99.14%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       2222    54
Positive:       19      2187
recall: 97.59% +/- 0.92% (micro average: 97.59%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       2222    54
Positive:       19      2187
f_measure: 98.36% +/- 0.49% (micro average: 98.36%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       2222    54
Positive:       19      2187
```



SVM with SMOTE

```
PerformanceVector:
accuracy: 98.91% +/- 0.59% (micro average: 98.91%)
ConfusionMatrix:
True:   Negative        Positive
Negative:       2241    49
Positive:       0       2192
AUC (optimistic): 0.999 +/- 0.001 (micro average: 0.999) (positive class: Positive)
AUC: 0.999 +/- 0.001 (micro average: 0.999) (positive class: Positive)
AUC (pessimistic): 0.999 +/- 0.001 (micro average: 0.999) (positive class: Positive)
precision: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       2241    49
Positive:       0       2192
recall: 97.81% +/- 1.18% (micro average: 97.81%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       2241    49
Positive:       0       2192
f_measure: 98.89% +/- 0.60% (micro average: 98.89%) (positive class: Positive)
ConfusionMatrix:
True:   Negative        Positive
Negative:       2241    49
Positive:       0       2192
```

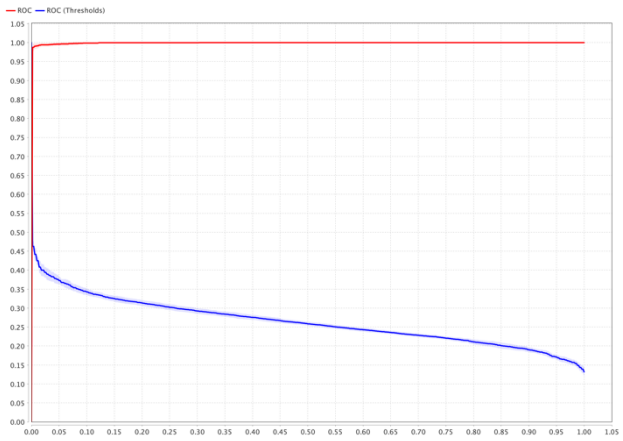


Figure 8. Performance of DT and SVM with SMOTE

Figure 8 shows the performance of DT and SVM with SMOTE. Upon evaluating the Decision Trees (DT) model utilizing the Synthetic Minority Over-sampling Technique (SMOTE), it becomes apparent that it achieves commendable performance metrics, encompassing accuracy, precision, recall, and f-measure. With an accuracy rating of 98.37% +/- 0.48%, the model adeptly categorizes instances into positive and negative classes. The Area Under the Curve (AUC) values further validate the model's resilience, with optimistic, regular, and pessimistic AUC scores at 0.998, 0.989, and 0.980, respectively. Additionally, the precision, recall, and f-measure metrics, indicative of the model's proficiency in identifying positive instances while minimizing false positives and negatives, emphasize its trustworthiness in sentiment analysis tasks. Consequently, the DT model with SMOTE emerges as a promising avenue for sentiment classification, showcasing its potential utility in practical settings.

Similarly, by evaluating the Support Vector Machine (SVM) model with the Synthetic Minority Over-sampling Technique (SMOTE), it is evident that the model attains remarkably high accuracy, precision, recall, and f-measure scores. Achieving an accuracy rate of 98.91% +/- 0.59%, the model effectively segregates instances into positive and negative categories. The AUC values further validate the model's robustness, with optimistic, regular, and pessimistic AUC scores reaching 0.999, signifying its significant predictive capability. Furthermore, the precision, recall, and f-measure metrics underscore the model's capacity to discern positive instances while minimizing erroneous classifications accurately, reinforcing its reliability in sentiment analysis endeavors. Hence, the SVM model with SMOTE emerges as a promising avenue for sentiment classification, hinting at its potential applicability in practical scenarios.

The limitation of this research lies in the scope of the video content and the number of comments, in addition to being constrained by the framework employed, namely CRISP-DM with DT and SVM models. While the study provides valuable insights into sentiment classification, the restricted dataset size and content variety may limit the generalizability of the findings to broader contexts. Moreover, the exclusive focus on DT and SVM models overlooks the potential contributions of alternative machine learning algorithms, warranting further exploration to enhance the comprehensiveness of sentiment analysis methodologies.

4. CONCLUSION

In conclusion, following the comprehensive evaluation based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, the Decision Trees (DT) model integrated with the Synthetic Minority Over-sampling

Technique (SMOTE) exhibits noteworthy performance metrics across various parameters, including accuracy, precision, recall, f-measure, and Area Under the Curve (AUC). With an accuracy of 98.37% +/- 0.48%, the DT model demonstrates robust classification capabilities, distinguishing between positive and negative instances. The AUC scores, encompassing optimistic, regular, and pessimistic scenarios, further attest to the model's resilience in sentiment analysis tasks. Moreover, the precision, recall, and f-measure metrics highlight the model's proficiency in identifying positive instances while minimizing false classifications, underscoring its reliability. Consequently, the DT model with SMOTE emerges as a promising avenue for sentiment classification within the CRISP-DM framework, suggesting its potential applicability in real-world scenarios.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] W. Van Zoonen, J. W. Treem, and A. Sivunen, "An analysis of fear factors predicting enterprise social media use in an era of communication visibility," *Internet Res.*, vol. 32, no. 7, pp. 354–375, Jan. 2022, doi: 10.1108/INTR-05-2021-0341.

[2] K. Abhari, M. Zarei, M. Parsons, and P. Estell, "Open innovation starts from home: the potentials of enterprise social media (ESM) in nurturing employee innovation," *Internet Res.*, vol. 33, no. 3, pp. 945–973, Jan. 2023, doi: 10.1108/INTR-08-2021-0556.

[3] H. Li, Z. Yang, C. Jin, and J. Wang, "How an industrial internet platform empowers the digital transformation of SMEs: theoretical mechanism and business model," *J. Knowl. Manag.*, vol. 27, no. 1, pp. 105–120, Jan. 2023, doi: 10.1108/JKM-09-2022-0757.

[4] M. T. Bui and D. J. F. Jeng, "Capture coproduction behavior in networking alumni communities: Progress from platform belongingness, knowledge sharing, and citizenship behavior," *J. Enterprising Communities*, vol. 16, no. 1, pp. 46–73, Jan. 2022, doi: 10.1108/JEC-08-2021-0112.

[5] E. Jütte and E. L. Olson, "A brand hegemony rejection explanation for digital piracy," *Eur. J. Mark.*, vol. 56, no. 5, pp. 1512–1531, Jan. 2022, doi: 10.1108/EJM-04-2020-0303.

[6] M. Törhönen, M. Sjöblom, L. Hassan, and J. Hamari, "Fame and fortune, or just fun? A study on why people create content on video platforms," *Internet Res.*, vol. 30, no. 1, pp. 165–190, Jan. 2020, doi: 10.1108/INTR-06-2018-0270.

[7] R. Casidy, C. Leckie, M. W. Nyadzayo, and L. W. Johnson, "Customer brand engagement and co-production: an examination of key boundary conditions in the sharing economy," *Eur. J. Mark.*, vol. 56, no. 10, pp. 2594–2621, Jan. 2022, doi: 10.1108/EJM-10-2021-0803.

[8] Y. Hong, S. Sawang, and H. P. (Sophie) Yang, "How is entrepreneurial marketing shaped by E-commerce technology: a case study of Chinese pure-play e-retailers," *Int. J. Entrep. Behav. Res.*, vol. 30, no. 2–3, pp. 609–631, Jan. 2024, doi: 10.1108/IJEBR-10-2022-0951.

[9] A. K. Olsson and I. Bernhard, "Keeping up the pace of digitalization in small businesses–Women entrepreneurs' knowledge and use of social media," *Int. J. Entrep. Behav. Res.*, vol. 27, no. 2, pp. 378–396, Jan. 2021, doi: 10.1108/IJEBR-10-2019-0615.

[10] B. Mastromartino and M. L. Naraine, "(Dis)Innovative digital strategy in professional sport: examining sponsor leveraging through social media," *Int. J. Sport. Mark. Spons.*, vol. 23, no. 5, pp. 934–949, Jan. 2022, doi: 10.1108/IJSMS-02-2021-0032.

[11] N. Gryllakis and M. Matsiola, "Digital audiovisual content in marketing and distributing cultural products during the COVID-19 pandemic in Greece," *Arts Mark.*, vol. 13, no. 1, pp. 4–19, Jan. 2023, doi: 10.1108/AAM-09-2021-0053.

[12] P. Tiwasing, Y. R. Kim, and S. Sawang, "The interplay between digital social capital and family-owned SME performance: a study of social media business networks," *J. Fam. Bus. Manag.*, vol. 13, no. 4, pp. 1026–1048, Jan. 2023, doi: 10.1108/JFBM-07-2022-0103.

[13] A. Boukis, "Exploring the implications of blockchain technology for brand–consumer relationships: a future research agenda," *J. Prod. Brand Manag.*, vol. 29, no. 3, pp. 307–320, Jan. 2020, doi: 10.1108/JPBM-03-2018-1780.

[14] K. K. Coker, R. L. Flight, and D. M. Baima, "Video storytelling ads vs argumentative ads: how hooking viewers enhances consumer engagement," *J. Res. Interact. Mark.*, vol. 15, no. 4, pp. 607–622, Jan. 2021, doi: 10.1108/JRIM-05-2020-0115.

[15] J. Ho, C. Pang, and C. Choy, "Content marketing capability building: a conceptual framework," *J. Res. Interact. Mark.*, vol. 14, no. 1, pp. 133–151, Jan. 2020, doi: 10.1108/JRIM-06-2018-0082.

[16] M. L. Cheung, W. K. S. Leung, M. X. Yang, K. Y. Koay, and M. K. Chang, "Exploring the nexus of social media influencers and consumer brand engagement," *Asia Pacific J. Mark. Logist.*, vol. 34, no. 10, pp. 2370–2385, Jan. 2022, doi: 10.1108/APJML-07-2021-0522.

[17] S. Fready, P. Vel, and M. W. Nyadzayo, "Business customer virtual interaction: enhancing value creation in B2B markets in the post-COVID-19 era – an SME perspective," *J. Bus. Ind. Mark.*, vol. 37, no. 10, pp. 2075–2094, Jan. 2022, doi: 10.1108/JBIM-01-2021-0074.

[18] Y. Wang, M. Zhang, and Y. Ming, "What contributes to online communities' prosperity? Understanding value co-creation in product-experience-shared communities (PESCs) from the view of resource integration," *Inf. Technol. People*, vol. 35, no. 7, pp. 2241–2262, Jan. 2022, doi: 10.1108/ITP-12-2020-0869.

[19] S. L. Alam, "Many hands make light work: towards a framework of digital co-production to co-creation on social platforms," *Inf. Technol. People*, vol. 34, no. 3, pp. 1087–1118, Jan. 2020, doi: 10.1108/ITP-05-2019-0231.

[20] G. Rejikumar, A. Jose, S. Mathew, D. P. Chacko, and A. Asokan-Ajitha, "Towards a theory of well-being in digital sports viewing behavior," *J. Serv. Mark.*, vol. 36, no. 2, pp. 245–263, Jan. 2022, doi: 10.1108/JSM-06-2020-0247.

[21] R. V. Kozinets, "Algorithmic branding through platform assemblages: core conceptions and research directions for a new era of marketing and service management," *J. Serv. Manag.*, vol. 33, no. 3, pp. 437–452, Jan. 2022, doi: 10.1108/JOSM-07-2021-0263.

[22] B. Senanu, T. Anning-Dorson, and N. N. Tackie, "Social media insights for non-luxury fashion SMEs in emerging markets: evidence from young consumers," *J. Fash. Mark. Manag.*, vol. 27, no. 6, pp. 965–987, Jan. 2023, doi: 10.1108/JFMM-02-2022-0026.
[23] A. Garrido-Moreno, V. García-Morales, S. King, and N. Lockett, "Social Media use and value creation in the digital landscape: a dynamic-capabilities perspective," *J. Serv. Manag.*, vol. 31, no. 3, pp. 313–343, Jan. 2020, doi: 10.1108/JOSM-09-2018-0286.
[24] E. E. Vazquez, "Effects of enduring involvement and perceived content vividness on digital engagement," *J. Res. Interact. Mark.*, vol. 14, no. 1, pp. 1–16, Jan. 2020, doi: 10.1108/JRIM-05-2018-0071.
[25] G. Oakley, "Developing pre-service teachers' technological, pedagogical and content knowledge through the creation of digital storybooks for use in early years classrooms," *Technol. Pedagog. Educ.*, vol. 29, no. 2, pp. 163–175, 2020, doi: 10.1080/1475939X.2020.1729234.
[26] E. Mora, N. Vila, and I. Küster, "Qualitative social media content analysis as teaching-learning method in higher education," *Interact. Learn. Environ.*, pp. 1–15, 2022, doi: 10.1080/10494820.2022.2150222.
[27] N. Al Said, L. Vorona-Slivinskaya, and E. Gorozhanina, "Data mining in education: managing digital content with social media analytics in medical education," *Interact. Learn. Environ.*, pp. 1–13, 2023, doi: 10.1080/10494820.2023.2194330.
[28] M. Lindfors and A. D. Olofsson, "The search for professional digital competence in Swedish teacher education policy—A content analysis of the prerequisites for teacher educators' dual didactic task," *Cogent Educ.*, vol. 10, no. 2, 2023, doi: 10.1080/2331186X.2023.2272994.
[29] Y. Zhou, B. J. Calder, E. C. Malthouse, and Y. K. Hessary, "Not all clicks are equal: detecting engagement with digital content," *J. Media Bus. Stud.*, vol. 19, no. 2, pp. 90–107, 2022, doi: 10.1080/16522354.2021.1924558.
[30] H. Liang, U. Ganeshbabu, and T. Thorne, "A Dynamic Bayesian Network Approach for Analysing Topic-Sentiment Evolution," *IEEE Access*, vol. 8, pp. 54164–54174, 2020, doi: 10.1109/ACCESS.2020.2979012.
[31] M. Sohi, M. Pitesky, and J. Gendreau, "Analyzing public sentiment toward GMOs via social media between 2019-2021," *GM Crop. Food*, vol. 14, no. 1, pp. 1–9, 2023, doi: 10.1080/21645698.2023.2190294.
[32] R. Thomas and J. R. Jeba, "A novel framework for an intelligent deep learning based product recommendation system using sentiment analysis (SA)," *Automatika*, vol. 65, no. 2, pp. 410–424, 2024, doi: 10.1080/00051144.2023.2295148.
[33] S. Sommariva, J. Beckstead, M. Khaliq, E. Daley, and D. Martinez Tyson, "An approach to targeted promotion of HPV vaccination based on parental preferences for social media content," *J. Soc. Mark.*, vol. 13, no. 3, pp. 341–360, Jan. 2023, doi: 10.1108/JSOCM-08-2022-0164.
[34] R. Odoom, "Digital content marketing and consumer brand engagement on social media- do influencers' brand content moderate the relationship?," *J. Mark. Commun.*, vol. 00, no. 00, pp. 1–24, 2023, doi: 10.1080/13527266.2023.2249013.
[35] M. Arevalillo-Herraez, P. Arnau-Gonzalez, and N. Ramzan, "On Adapting the DIET Architecture and the Rasa Conversational Toolkit for the Sentiment Analysis Task," *IEEE Access*, vol. 10, no. September, pp. 107477–107487, 2022, doi: 10.1109/ACCESS.2022.3213061.
[36] U. Naqvi, A. Majid, and S. A. Abbas, "UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods," *IEEE Access*, vol. 9, pp. 114085–114094, 2021, doi: 10.1109/ACCESS.2021.3104308.
[37] A. P. Rodrigues, N. N. Chiplunkar, and R. Fernandes, "Aspect-based classification of product reviews using Hadoop framework," *Cogent Eng.*, vol. 7, no. 1, 2020, doi: 10.1080/23311916.2020.1810862.
[38] 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.
[39] 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.
[40] 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.
[41] J. L. Arroyo Barrigüete, L. Barcos, C. Bellón, and T. Corzo, "One year of European premiers leadership and empathy in times of global pandemic: a Twitter sentiment analysis," *Cogent Soc. Sci.*, vol. 8, no. 1, 2022, doi: 10.1080/23311886.2022.2115693.
[42] W. Zhao, X. Yang, and N. Sun, "Do Digital City Policies Promote Corporate ESG Performance ? Evidence from Research on Textual Analysis of China Do Digital City Policies Promote Corporate ESG Performance ? Evidence from Research on Textual Analysis of China," *Emerg. Mark. Financ. Trade*, vol. 00, no. 00, pp. 1–20, 2024, doi: 10.1080/1540496X.2024.2331013.