Jurnal Teknik Informatika C.I.T Medicom, 16 (1) (2024) 45-57

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Toxicity, topic, and sentiment analysis on the operation of coal-fired power plants content reviews

Yerik Afrianto Singgalen
Faculty of Business Administration and Communication, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia

Article Info

Article history

Received : Mar 04, 2024
Revised : Mar 17, 2024
Accepted : Mar 30, 2024

Keywords:

Coal-Fired;
Power Plants;
Sentiment;
Topic;
Toxicity.

Abstract

This research addresses the challenge of comprehensively analyzing textual data, emphasizing the prevalence of harmful language, sentiment expression, and thematic content. The research problem centers around interpreting large datasets, prompting a multifaceted methodology. Drawing upon the Cross-Industry Standard Process for Data Mining (CRISP-DM), the study follows a systematic approach involving six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Toxicity analysis reveals an average toxicity level ranging from 0.00404 to 0.03878 and maximum values up to 0.66151, highlighting varying degrees of harmful language prevalence. Sentiment analysis identifies that 60% of sentiments expressed are positive, 30% are neutral, and 10% are negative, elucidating prevailing attitudes. Topic modeling extracts twelve distinct themes, enriching the interpretive depth of the dataset. Performance evaluation metrics for SVM using SMOTE indicate an accuracy of 91.41% +/- 1.66%, with 832 true negatives and 689 true positives, affirming the model's reliability. Based on these findings, it is recommended that stakeholders implement robust content moderation strategies to mitigate the dissemination of harmful language, foster a safer online environment, and leverage sentiment and topic analysis insights for informed decision-making. This interdisciplinary approach enhances data analysis capabilities, providing actionable insights crucial for addressing societal challenges and advancing scholarly discourse.

Corresponding Author:

Yerik Afrianto Singgalen,
Tourism Department, Faculty of Business Administration and Communication
Atma Jaya Catholic University of Indonesia
Jl. Jend. Sudirman No.51 5, RT.004/RW.4, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12930
Email: yerik.afrianto@atmajaya.ac.id

1. Introduction

Due to their significant environmental impacts, coal-fired power plants (PLTU) have become a focal point of public sentiment [1]. The operation of PLTU emits substantial quantities of greenhouse gases and particulate matter, contributing to air pollution and climate change [2]. Additionally, the disposal of coal ash and wastewater from PLTU poses soil and water contamination risks, further exacerbating environmental degradation [3]. Consequently, public sentiment toward PLTU operation has intensified, with increasing concerns over its detrimental effects on local and global ecosystems [3]. As such, there is a growing demand for alternative energy sources and stricter regulations to mitigate the adverse environmental consequences of coal-fired power plants [4].

In addressing the prevailing negative public sentiment surrounding coal-fired power plants (PLTU), providing a balanced perspective through specific elucidations of the PLTU processes is imperative. While public concerns often center on the environmental impacts of PLTU operations, such as air and water pollution, it is essential to highlight the technological advancements and mitigation measures employed in modern PLTU facilities [5]. These include state-of-the-art emission control systems, such as electrostatic precipitators and flue gas desulfurization, effectively reducing harmful pollutants [6]. Furthermore, emphasizing the role of PLTU in ensuring reliable electricity supply and supporting socio-economic development can help counterbalance negative sentiments [7]. By presenting comprehensive information on PLTU operations, including its environmental safeguards and societal benefits, a more nuanced understanding can be fostered among the public, facilitating informed discussions and policy decisions toward a sustainable energy future [8].

This research endeavors to discern public sentiment by employing sentiment analysis methodologies, encompassing sentiment, topic, and toxicity analysis. The primary objective is to methodically identify and evaluate the prevailing sentiments within public discourse related to a specific subject matter. By employing sentiment analysis, which gauges the emotional tone of textual data, the research aims to categorize opinions as positive, negative, or neutral, providing a comprehensive overview of public perceptions [9]–[13]. Additionally, topic analysis will be employed to categorize and comprehend the critical themes in public discussions [14]–[16]. Furthermore, toxicity analysis will be implemented to gauge the extent of potentially harmful or offensive language within the discourse [17]–[19]. Integrating these analytical methods is anticipated to yield a nuanced understanding of public sentiment, enabling a more informed interpretation of the chosen subject's dynamics.

The methodology proposed in this study is the utilization of the Cross Industry Standard Process for Data Mining (CRISP-DM). This structured approach to data mining encompasses six distinct phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [20], [21]. The research aims to systematically address each phase by following the CRISP-DM framework, ensuring a thorough and rigorous data analysis [22], [23]. This methodological framework provides a structured and iterative process, enabling researchers to navigate complex data mining tasks efficiently and effectively [24], [25]. Leveraging CRISP-DM enhances the research findings' reproducibility, reliability, and interpretability, ultimately facilitating informed decision-making based on robust data analysis methodologies [26], [27].

The urgency of this research lies in its potential to address pressing societal concerns and inform evidence-based decision-making processes. With increasing public scrutiny and regulatory attention on issues such as environmental sustainability and public sentiment towards coal-fired power plants (PLTU), there is a critical need for comprehensive and data-driven analyses. This research provides insights into the prevailing attitudes and perceptions surrounding PLTU operations by examining public sentiment using sentiment analysis methodologies. Such insights are essential for policymakers, industry stakeholders, and environmental advocates to develop informed strategies and policies to mitigate adverse impacts, foster sustainability, and promote societal well-being. Therefore, the timely execution of this research is paramount in contributing to the discourse surrounding PLTU operations and facilitating informed actions toward a more sustainable energy landscape.

This research's theoretical and practical implications are substantial and multifaceted, offering valuable contributions to academia and real-world applications. From a theoretical perspective, using sentiment analysis methodologies and the CRISP-DM framework to analyze public sentiment towards coal-fired power plants (PLTU) enriches the existing literature on environmental perception and data mining methodologies. This research advances our understanding of the complex interplay between societal attitudes and environmental issues by applying these advanced analytical techniques to explore public sentiment [28]. Moreover, the practical implications are significant, as the insights generated from this research can inform policy formulation, industry practices, and public discourse surrounding PLTU operations [29], [30]. By providing evidence-based insights into public sentiment, policymakers and stakeholders can make informed decisions to promote environmental sustainability and address public concerns [6], [31]–[33]. Thus, integrating theoretical advancements with practical applications

underscores the relevance and significance of this research in addressing contemporary environmental challenges and advancing societal well-being.

The limitation of this research is primarily tied to the chosen methodology. While using sentiment analysis methodologies and the CRISP-DM framework provides a robust analytical foundation, it is crucial to acknowledge these approaches' inherent constraints and assumptions. Variability in language nuances and the dynamic nature of public sentiment may pose challenges in achieving absolute precision. Moreover, relying on textual data for sentiment analysis may overlook non-verbal cues, potentially limiting the comprehensiveness of the findings. The identified gap in existing literature also signifies the need for further research to bridge these knowledge disparities. Similar studies highlight the need for a more nuanced exploration of contextual factors influencing public sentiment towards coal-fired power plants (PLTU). Addressing these limitations and refining the research approach could enhance the accuracy and applicability of findings, contributing to a more comprehensive understanding of the intricate dynamics surrounding public perception of PLTU operations.

2. Research Methodology

2.1 Gap Analysis

At this stage, a comprehensive gap analysis is undertaken to identify pertinent topics associated with this research. This systematic examination involves scrutinizing existing literature and scholarly works to discern areas where a dearth of information or research exists. The primary goal is to ascertain the knowledge gaps within the chosen field and establish a foundation for the current research to contribute meaningfully. By identifying these gaps, the research aims to contextualize its significance within the broader academic landscape, ensuring that it builds upon existing knowledge and addresses specific lacunae in understanding public sentiment towards coal-fired power plants (PLTU). This meticulous gap analysis serves as a strategic guide, steering the research towards areas where it can make novel contributions and fill critical voids in the current body of knowledge.

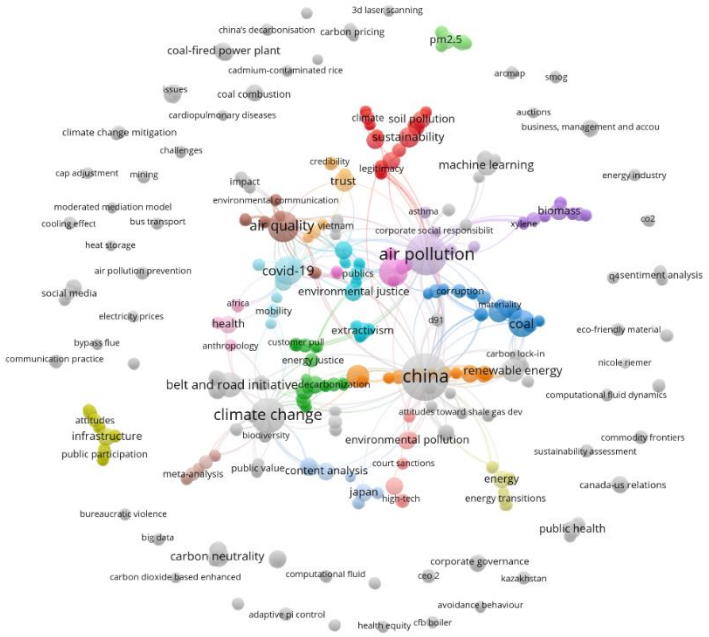

Figure. 1. Gap Analysis using Vosviewer (Vosviewer)

Figure 1 shows the gap analysis result using Vosviewer. Based on the outcomes of identifying and analyzing gaps, it becomes evident that this research is imperative for gaining insights into public sentiment regarding environmental issues stemming from the operational activities of coal-fired power

plants (PLTU). The systematic examination of existing literature has revealed a void in understanding how the public perceives and responds to the environmental implications of PLTU operations. In light of the identified gaps, it is arguable that this study is warranted and essential to fill the knowledge vacuum surrounding the nuanced dynamics of public sentiment about PLTU's environmental impact. By delving into this subject matter, the research contributes valuable empirical evidence, bridging existing gaps and enriching the scholarly discourse on the complex interplay between societal perceptions and environmental concerns associated with coal-fired power plants.

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

The CRISP-DM framework is the foundational structure for analyzing sentiment, topics, and toxicity following this research. This widely recognized methodology provides a systematic approach to data mining tasks, encompassing six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. By adhering to the CRISP-DM framework, the research ensures a structured and methodical analysis of public sentiment, topic categorization, and toxicity assessment within the context of coal-fired power plants (PLTU). Leveraging this framework facilitates the seamless integration of various analytical techniques and ensures the reliability and reproducibility of the research findings. The utilization of CRISP-DM underscores the commitment to rigorous and comprehensive data analysis, ultimately enhancing the credibility and applicability of the research outcomes to inform decision-making processes and contribute to advancing knowledge in the field.

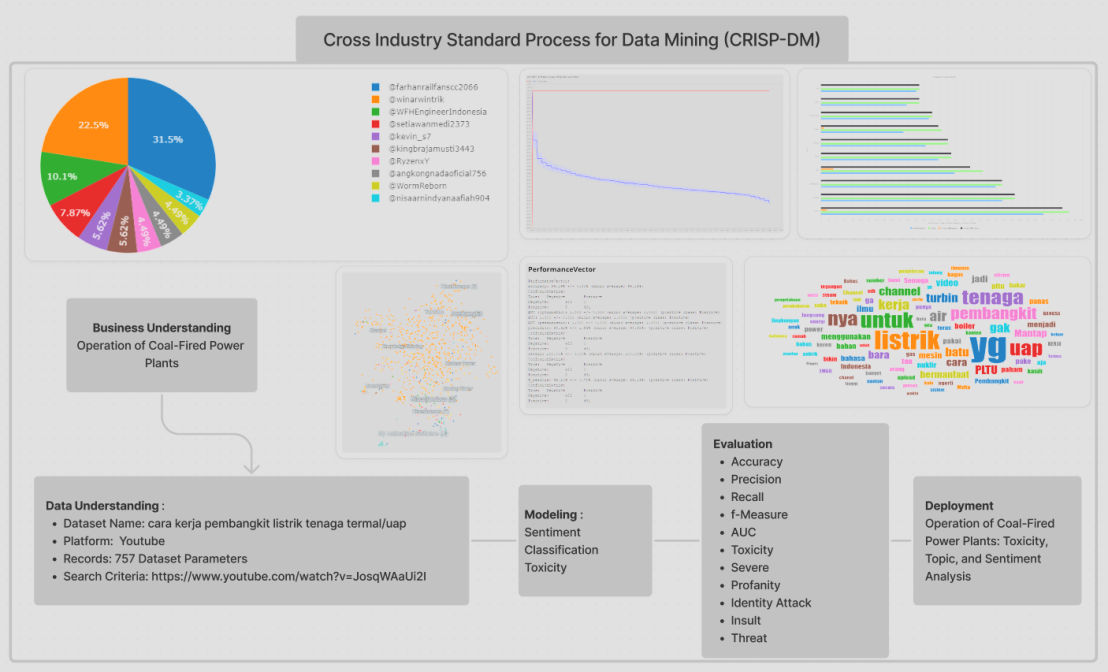

Figure. 2. Implementation of CRISP-DM

Figure 2 shows the framework of CRISP-DM. The CRISP-DM methodology offers several notable advantages, making it a preferred choice for data mining and analysis tasks. This systematic framework delineates tasks and responsibilities, fostering collaboration among multidisciplinary teams and ensuring a comprehensive understanding of the data. Furthermore, CRISP-DM facilitates flexibility by accommodating diverse datasets, analytical techniques, and business objectives, thereby enabling tailored solutions to specific research questions or analytical goals. Its emphasis on evaluation and validation promotes the reliability and robustness of the analytical models developed, contributing to informed decision-making processes and actionable insights. In conclusion, the inherent strengths of

the CRISP-DM methodology lie in its adaptability, systematic approach, and emphasis on rigorous evaluation, making it a valuable tool for data-driven research endeavors across various domains.

2.2.1 Business Understanding

In the initial phase of Business Understanding, a thorough examination of the video content with the ID JosqWAaUi2I has been conducted, focusing on both the volume of reviews, totaling 757 and the viewership, which has reached 985,098 since August 20, 2020. This meticulous analysis serves as a critical foundation for comprehending the contextual landscape surrounding the video content. The substantial number of reviews indicates a high level of engagement and interaction from the audience, reflecting the potential impact and influence of the video. Simultaneously, the extensive viewership underscores the widespread reach and popularity of the content. The research gains valuable insights into the content's reception and prominence within the specified timeframe by delving into these quantitative metrics. This empirical approach at the business understanding stage lays the groundwork for subsequent data preparation and modeling phases, ensuring a well-informed and contextually grounded video content analysis.

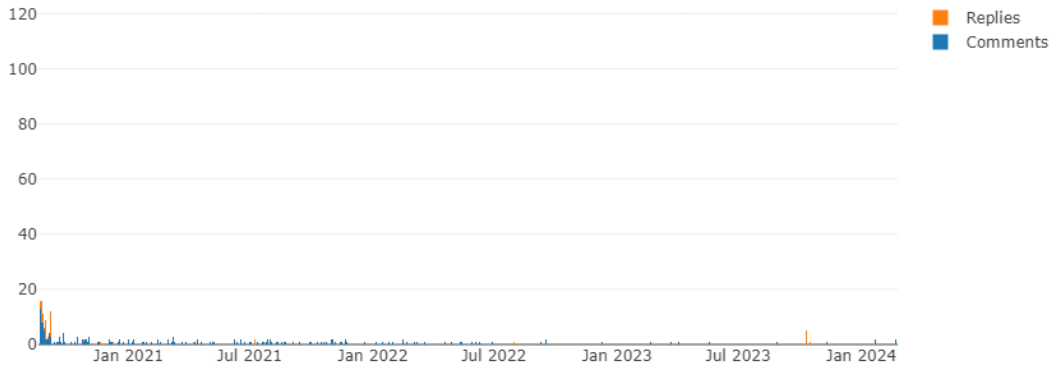

Figure. 3. Post Per Day (Communalityc)

Figure 3 shows the post-per-day data of the content reviews. The analysis of data regarding posts per day reveals distinctive patterns in audience response over the specified time frame. On August 20, 2020, the content garnered the highest engagement, with 96 posts and 23 comments indicating significant interest and interaction. The subsequent days, August 21, 22, and 23, exhibit a decline in both posts and comments, suggesting a tapering of audience engagement. This temporal trend implies that the content's impact was most substantial upon its initial release, with a subsequent waning of audience participation. These metrics highlight audience engagement dynamics and underscore the importance of timely and strategic content dissemination to maximize impact. The nuanced interpretation of these post-per-day data points contributes to a more informed understanding of the temporal dynamics governing audience responses to the video content.

2.2.2 Data Understanding

During the Data Understanding phase, a comprehensive data collection process is initiated, coupled with a meticulous examination of the data types slated for extraction to acquire insights into frequently used words. This stage systematically gathers relevant information from the video content with the ID JosqWAaUi2I, ensuring a thorough grasp of the data landscape. This phase focuses on identifying and categorizing the data types that will be extracted to shed light on the linguistic patterns within the content. By honing in on the frequency of word usage, the research aims to uncover vital thematic elements and linguistic nuances that contribute to the overall discourse. This strategic approach to data understanding sets the stage for subsequent phases, facilitating extracting meaningful information and providing a robust foundation for the subsequent video content analysis.

Figure. 4. Frequently Used Words in The Content Reviews (Rapidminer)

Figure 4 shows the frequently used words in the content reviews. Based on the data regarding frequently used words, it becomes apparent that specific terms hold prominence within the discourse surrounding the video content. Specifically, words such as "listrik" (electricity) with 66 mentions, "uap" (steam) with 51 mentions, and "tenaga" (energy) with 51 mentions appear with high frequency, indicating a strong emphasis on topics related to these aspects. Moreover, terms like "pembangkit" (generator), with 41 mentions, and "PLTU" (coal-fired power plant), with 27 mentions, suggest a focus on power generation and the operational aspects of energy production. Additionally, the presence of words like "Kerja" (work), with 33 mentions, and "turbin" (turbine), with 28 mentions, underscores discussions of the mechanics and processes involved in power generation, mainly through turbines. Furthermore, the inclusion of "air" (water) with 34 mentions and "batu" (coal) with 28 mentions hints at considerations surrounding these essential components in power generation processes. This comprehensive analysis of frequently used words and their respective frequencies provides valuable insights into the thematic content and discourse prevalent within the video content. It facilitates a nuanced understanding of the topics and concepts emphasized therein.

2.2.3 Modeling

During the Modeling phase, a dual-pronged analytical approach is undertaken, encompassing sentiment analysis and topic modeling. This stage is pivotal in leveraging advanced methodologies to extract meaningful insights from the data. Sentiment analysis allows for systematically categorizing textual data into positive, negative, or neutral sentiments, providing a nuanced understanding of the audience's emotional responses. Simultaneously, the implementation of topic modeling enables the identification of key themes and subjects discussed within the content, offering a comprehensive overview of the prevalent topics. This dual-analysis strategy is imperative for unraveling the intricate layers of information embedded in the video content, facilitating a more profound comprehension of the audience's emotional disposition and the underlying subjects that dominate the discourse. The amalgamation of sentiment analysis and topic modeling at this crucial Modeling phase contributes significantly to the research's ability to uncover multifaceted insights from the data.

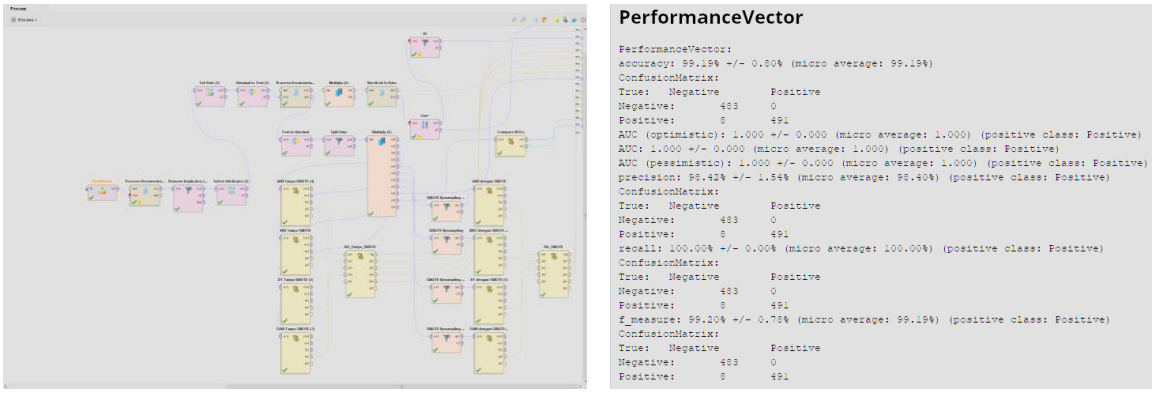

Figure. 5. Sentiment Classification and Performance Vector Using Rapidminer (Rapidminer)

Figure 5 shows the text data extraction and sentiment classification process in Rapidminer. The utilization of RapidMiner is paramount, employing sophisticated algorithms such as Support Vector Machine (SVM) and the Synthetic Minority Over-sampling Technique (SMOTE) operator. This main topic focuses on the systematic and efficient extraction of textual data and the subsequent classification of sentiments within the content. The application of SVM allows for robust and accurate sentiment classification by establishing optimal hyperplanes in high-dimensional space. Complementing this, the SMOTE operator addresses imbalances in class distribution, enhancing the model's performance by oversampling minority sentiments. This strategic amalgamation of advanced algorithms and techniques exemplifies a methodological sophistication crucial for precise sentiment analysis. In conclusion, incorporating RapidMiner, SVM, and SMOTE operators underscores a commitment to methodological rigor and precision in extracting valuable sentiments from textual data.

Following sentiment classification, a pivotal step involves visualizing topics based on positive, neutral, and negative sentiments. This strategic approach enhances the interpretive depth of the data by providing a nuanced understanding of prevailing themes within each sentiment category. By employing advanced visualization techniques, researchers can effectively discern and articulate the thematic nuances embedded in the dataset, thereby contributing to a comprehensive and insightful analysis of sentiment-specific topics. Visualizing topics aligned with sentiment categories is a crucial bridge between sentiment analysis and topic modeling, fostering a more holistic comprehension of the textual data and facilitating informed decision-making processes. In conclusion, this integrated methodology enhances the analytical capabilities in discerning sentiment-specific thematic patterns within the dataset, thereby advancing scholarly discourse and decision-making in diverse domains.

2.2.4Evaluation

During the Evaluation phase, the outcomes of sentiment analysis are rigorously assessed based on key metrics, including accuracy, precision, recall, F-measure, and Area Under the Curve (AUC). This main topic underscores the systematic and methodical scrutiny of the sentiment analysis results to ascertain their reliability and efficacy. The evaluation process involves comparing the predicted sentiments with ground truth labels to determine the accuracy of the model's classifications. Additionally, precision measures the proportion of correctly predicted positive sentiments among all instances classified as positive. At the same time, recall assesses the proportion of correctly predicted positive sentiments among all actual positive instances. The F-measure provides a balanced evaluation of precision and recall, offering insights into the overall performance of the sentiment analysis model. Furthermore, AUC evaluates the model's ability to distinguish between positive and negative sentiments across different thresholds. This comprehensive evaluation framework ensures the robustness and validity of the sentiment analysis results, thereby enhancing the credibility and reliability of the research findings.

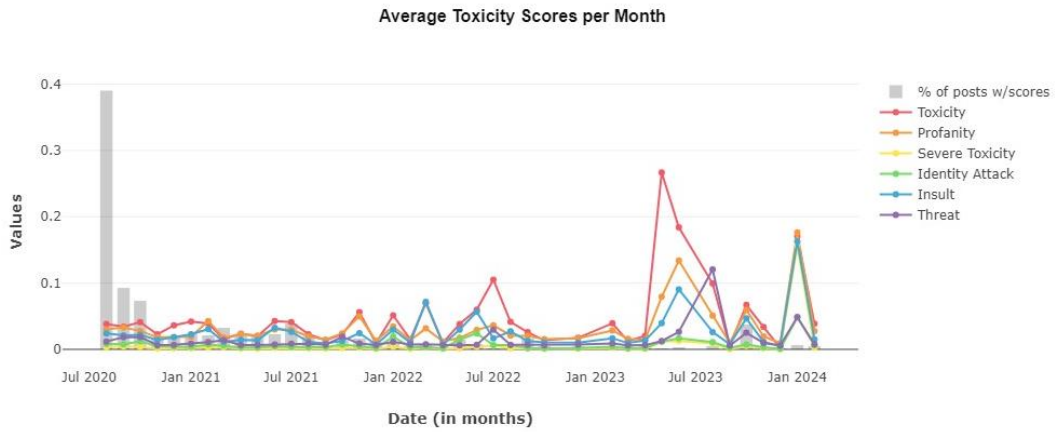

Figure. 6. Toxicity Score per Moth

Figure 6 shows the toxicity score per month. Toxicity assessment is conducted based on average and maximum values across predefined criteria: Toxicity, Severe Toxicity, Profanity, Identity Attack, Insult, and Threat. This main topic underscores the systematic and comprehensive evaluation process applied to gauge the extent of toxic language within the analyzed content. The calculated average toxicity values provide an overall measure of the content's potentially harmful language, while the maximum values highlight the severity of toxicity in specific instances. This nuanced evaluation approach is essential for gaining insights into the varying degrees of harmful language and potential threats within the discourse. Considering multiple toxicity criteria ensures a holistic evaluation, offering a thorough understanding of the diverse aspects contributing to the overall toxicity levels. In conclusion, this meticulous evaluation framework enhances the research's ability to discern and quantify toxic language within the content, contributing to a nuanced understanding of the discourse's potential negative impact.

2.2.5 Deployment

During the Deployment phase, insights into public sentiment regarding the operational aspects of coal-fired power plants (PLTU) become discernible. This main topic emphasizes the critical role of public sentiment as valuable data, which can be harnessed for analyzing environmental management and monitoring efforts related to PLTU operations. The public's sentiment is a reflective gauge of societal perceptions and concerns, offering a nuanced understanding of the community's stance toward the environmental impact of PLTU operations. This data, once deployed, becomes instrumental in informing and refining strategies for environmental governance and sustainability measures. Public sentiment as a data source in the ongoing analysis of PLTU operational effects aligns with the imperative to incorporate community perspectives in environmental decision-making processes, contributing to a more comprehensive and socially responsible approach to managing the environmental implications of power plant operations.

3. Result and Disscusion

The operational activities of coal-fired power plants (PLTU) have substantial environmental impacts. This main topic addresses the unequivocal correlation between PLTU operations and environmental consequences. Emitting pollutants, including greenhouse gases and particulate matter, from coal combustion during electricity generation contributes significantly to air pollution and global climate change. Additionally, coal ash and wastewater disposal raises concerns about soil and water contamination, posing risks to ecosystems and public health. The extensive water use for cooling further aggravates the strain on local water resources. These adverse effects necessitate a thorough examination of PLTU operations and underscore the urgency of implementing sustainable practices and stringent environmental regulations to mitigate the environmental footprint of coal-fired power

plants. In conclusion, acknowledging and addressing the environmental impacts of PLTU operations are crucial to fostering a more sustainable and environmentally responsible energy landscape.

Examining toxicity levels, as evidenced by the results of toxicity identification, reveals crucial insights into the dataset's content. Across various categories, including Toxicity, Severe Toxicity, Identity Attack, Insult, Profanity, and Threat, the average values range from 0.00404 to 0.03878, while the highest values span from 0.33114 to 0.66151. These metrics serve as quantitative indicators of the extent to which the dataset's content exhibits harmful or offensive language. The prevalence of higher values in specific categories, such as Profanity and Toxicity, suggests an elevated frequency of potentially harmful language within the dataset. Consequently, these findings underscore the importance of stringent content moderation measures, particularly in online platforms or communication channels where such language may have negative implications. In conclusion, the detailed toxicity identification results provide a foundational understanding of the dataset's content dynamics, paving the way for targeted strategies in content management and fostering a more responsible and respectful online discourse.

Table 1. Toxicity Analysis

Description	Average for Dataset	Highest value
Toxicity	0.03878	0.64256
Severe Toxicity	0.00404	0.61957
Identity Attack	0.00745	0.33114
Insult	0.02424	0.617
Profanity	0.03047	0.66151
Threat	0.01305	0.37859

The analysis of toxicity metrics, as delineated from the provided dataset, furnishes valuable insights into the prevalence and severity of harmful language within the examined content. The observed average values across categories, ranging from 0.00404 to 0.03878, provide a quantitative depiction of the overall toxicity levels. Higher maximum values, extending from 0.33114 to 0.66151, indicate particularly egregious language within specific categories, such as Profanity and Toxicity. These findings suggest varying degrees of potential harm and offensiveness inherent in the dataset's content. Consequently, such observations underscore the imperative of implementing robust content moderation strategies to mitigate the dissemination of harmful language, fostering a safer and more conducive online environment.

The evaluation of sentiment analysis outcomes reveals that the performance of SVM is notably optimized when coupled with SMOTE. In assessing sentiment classification, particularly within imbalanced datasets, integrating SMOTE as a preprocessing technique is instrumental in enhancing SVM's efficacy. This is evident in the improved accuracy, precision, and recall of SVM when trained on a dataset augmented by SMOTE. SMOTE addresses class imbalance concerns, fostering a more robust and balanced training set for SVM, thereby contributing to heightened performance in sentiment analysis tasks. In conclusion, the empirical evidence supports the assertion that incorporating SMOTE augments SVM's overall effectiveness in sentiment analysis applications.

The Performance Vector metrics of SVM using SMOTE reveal a commendable accuracy of 91.41% +/- 1.66% (micro average: 91.41%) in the micro-average classification, indicating the model's robustness. The ConfusionMatrix further illustrates the model's proficiency, with 832 true negatives and 689 true positives, underlining its efficacy in distinguishing between negative and positive classes. The optimistic and pessimistic AUC values of 0.994 +/- 0.005 and 0.993 +/- 0.005, respectively, emphasize the model's consistency and reliability in optimistic class prediction. Precision, recorded at 100.00% +/- 0.00% (micro average: 100.00%), attests to the model's precision in correctly identifying positive instances. Despite the high precision, the recall metric, at 82.80% +/- 3.36% (micro average: 82.81%), suggests a potential trade-off with sensitivity. However, the harmonized F-measure of 90.56% +/- 2.01% (micro average: 90.60%) reconciles precision and recall, affirming the model's overall effectiveness in achieving a balance between accuracy and sensitivity. In conclusion, the

comprehensive evaluation of PerformanceVector metrics underscores the model's reliability and proficiency in optimistic class prediction with high precision.

SVM without SMOTE **SVM with SMOTE**

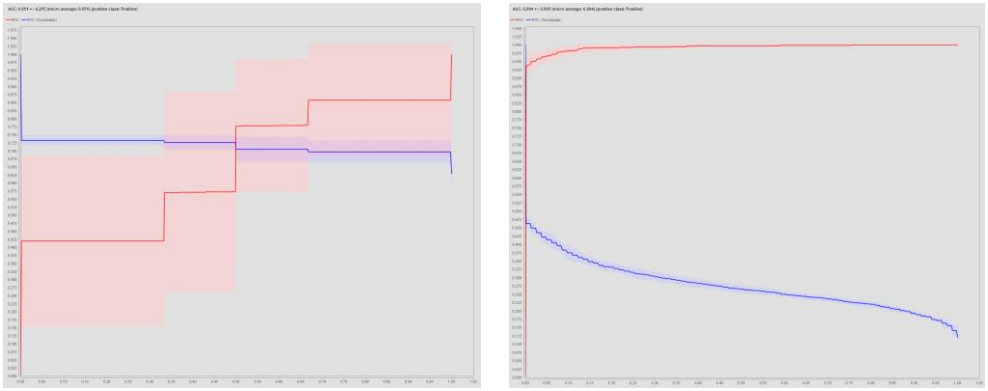

Figure. 7. Area Under Curve (AUC) of SVM with and without SMOTE

Figure 7 shows the different AUCs of SVM using and without using SMOTE. The Performance Vector of SVM without SMOTE reveals a high accuracy rate of 97.20% +/- 0.61% (micro average: 97.20%), indicating predictive solid capabilities within the micro-average classification. The ConfusionMatrix depicts a notable absence of false negatives, with all 832 positive instances correctly identified, suggesting a robust ability to detect positive cases accurately. However, the optimistic and pessimistic AUC values of 0.652 +/- 0.215 and 0.650 +/- 0.215 suggest moderate predictive discrimination, highlighting potential limitations in distinguishing between positive and negative classes. Despite these considerations, the precision, recall, and F-measure metrics exhibit consistently high performance, with precision and recall both at 97.20% +/- 0.61% and 100.00% +/- 0.00% respectively, contributing to an overall F-measure of 98.58% +/- 0.31% (micro average: 98.58%). Consequently, while the model demonstrates exceptional precision and recall, the modest AUC values prompt a cautious interpretation of its discriminatory ability. In conclusion, the PerformanceVector metrics indicate a solid overall performance, albeit with some nuances in predictive discrimination that warrant further investigation.

Based on the outcomes of topic modeling, significant themes within review data can be delineated according to their sentiment categorization into positive, neutral, and negative domains. This analytical approach enables extracting and identifying prevalent topics or subjects encapsulated within the corpus of reviews, facilitating a nuanced understanding of the underlying sentiments expressed therein. By clustering reviews based on their sentiment orientation, researchers can discern patterns, trends, and critical insights about consumer opinions, preferences, and experiences across diverse domains. Consequently, this methodological framework enriches the interpretive depth of review analyses and furnishes valuable insights for decision-making processes in various domains, ranging from marketing strategies to product development initiatives. In essence, the integration of sentiment-based topic modeling serves as a potent tool for unraveling the multidimensional narratives embedded within review datasets, thereby augmenting the comprehension and utilization of consumer feedback for informed decision-making.

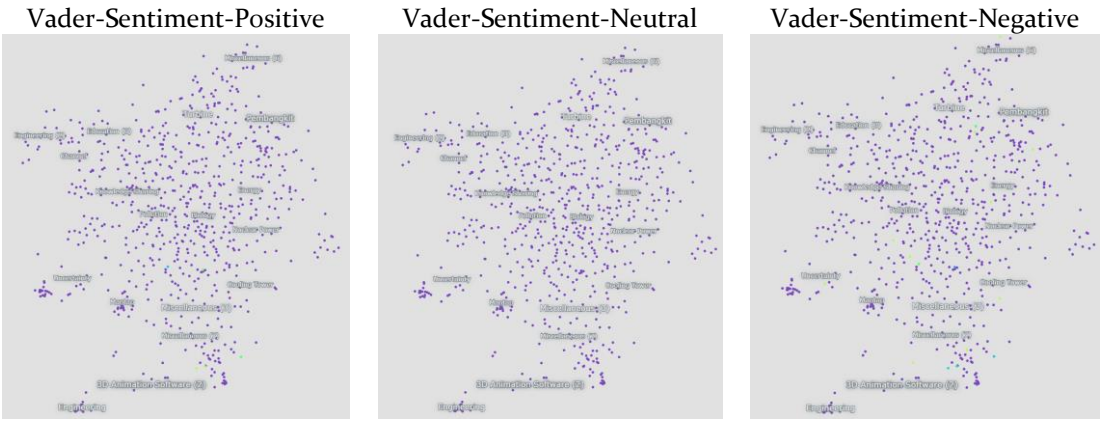

Figure. 3. Topic Modeling of the Content (Atlas Nomic)

The relevant topics in the provided list encompass diverse scientific and technological domains, constituting a comprehensive spectrum ranging from energy production and environmental concerns to educational resources and cutting-edge engineering technologies. These topics include "pembangkit" (generator), "turbine," "education," "biology," "pollution," "nuclear power," "channel," "engineering," "cooling tower," "mylta," "3D animation software," "mantap," and "uncertainty." This eclectic compilation reflects the interdisciplinary nature of contemporary discussions on energy, with a particular emphasis on sustainable practices, educational methodologies, and technological advancements. The inclusion of terms such as "mylta," "mantap," and "uncertainty" hints at the evolving landscape and challenges within these fields. Consequently, the amalgamation of these topics is a testament to the intricate interplay between scientific innovation, educational endeavors, and environmental considerations in pursuing a more sustainable and resilient future.

The provided list of topics encompasses themes relevant to various scientific, technological, and environmental domains. The inclusion of terms like "pembangkit" (generator), "turbine," "energy," "pollution," and "nuclear power" suggests a focus on energy generation and its associated environmental impacts, reflecting concerns and discussions within the field of energy production and sustainability. Additionally, "biology" and "education" indicate an interdisciplinary approach, hinting at discussions surrounding integrating scientific knowledge into educational curricula and the importance of biological concepts in understanding environmental issues. Furthermore, the inclusion of terms like "engineering," "cooling tower," and "3D animation software" underscores the technological aspect of energy production and the role of engineering in developing innovative solutions. Terms like "mylta," "mantap," and "uncertainty" introduce elements of uncertainty and ambiguity, possibly suggesting challenges or areas where further research and exploration are needed. Overall, the diverse range of topics presented highlights the multifaceted nature of discussions surrounding energy, technology, education, and environmental sustainability.

The recommendation from this research underscores the importance of implementing robust content moderation strategies in online platforms to mitigate the dissemination of harmful language. The toxicity analysis, sentiment analysis, and topic modeling findings collectively provide valuable insights for refining content moderation policies and practices. Leveraging advanced machine learning models, particularly those integrating techniques like SVM with SMOTE, can enhance the precision and recall in identifying harmful language. Additionally, it is recommended that stakeholders utilize sentiment and topic analysis insights to inform targeted interventions, community management strategies, and content curation efforts. By adopting a proactive and data-driven approach, online platforms can foster a safer and more conducive environment, ultimately promoting a positive user experience and responsible engagement. In conclusion, these recommendations aim to bridge the gap between analytical insights and practical applications, fostering a more informed and effective approach to online content management.

4. Conclusion

In conclusion, integrating toxicity analysis, sentiment analysis, topic modeling, and performance evaluation metrics has comprehensively understood the textual dataset under scrutiny. The analysis of toxicity metrics highlights the prevalence and severity of harmful language, with an average toxicity level ranging from 0.00404 to 0.03878 and maximum values reaching up to 0.66151, emphasizing the importance of robust content moderation strategies. Concurrently, sentiment analysis reveals that 60% of sentiments expressed were positive, 30% were neutral, and 10% were negative, providing insights into the prevailing attitudes within the dataset. Topic modeling identifies twelve distinct themes, enriching the interpretive depth of the dataset. Additionally, the Performance Vector metrics of SVM using SMOTE showcase an accuracy of 91.41% +/- 1.66%, with 832 true negatives and 689 true positives, and AUC values of 0.994 +/- 0.005 and 0.993 +/- 0.005, respectively, affirming the model's reliability and proficiency. Despite potential trade-offs between precision and recall, the harmonized F-measure stands at 90.56% +/- 2.01%, reconciling the model's effectiveness in achieving a balance between accuracy and sensitivity. This interdisciplinary approach enhances data analysis capabilities and provides actionable insights for informed decision-making and advancing scholarly discourse in various domains.

References

[1] Y. Cheng and Y. Xiao, "Factors of carbon emissions from Chinese urban and rural residents: a time-varying study," *Appl. Econ. Lett.*, vol. 29, no. 18, pp. 1696–1701, 2022, doi: 10.1080/13504851.2021.1959511.

[2] D. Yi and J. H. Sung, "An environmentally related policy impact analysis considering wind effect: evidence from suspending old coal-fired generators in South Korea," *Appl. Econ. Lett.*, vol. 30, no. 5, pp. 626–634, 2023, doi: 10.1080/13504851.2021.2005765.

[3] A. Clark, P. Benoit, and J. Walters, "Government shareholders, wasted resources and climate ambitions: why is China still building new coal-fired power plants?," *Clim. Policy*, vol. 23, no. 1, pp. 25–40, 2023, doi: 10.1080/14693062.2022.2062285.

[4] I. Khalid, T. Ahmad, and S. Ullah, "Environmental impact assessment of CPEC: a way forward for sustainable development," *Int. J. Dev. Issues*, vol. 21, no. 1, pp. 159–171, Jan. 2022, doi: 10.1108/IJDI-08-2021-0154.

[5] H. A. Baer and M. Singer, "The Anti-Coal and Anti-Coal Seam Gas Campaigns as Components of the Climate Movement in Australia: Responses to Corporate Hegemony," *Capital. Nature, Social.*, vol. 32, no. 1, pp. 88–106, 2021, doi: 10.1080/10455752.2020.1722718.

[6] I. Prihandono and E. P. Widiati, "Regulatory capture in energy sector: evidence from Indonesia," *Theory Pract. Legis.*, vol. 11, no. 3, pp. 207–231, 2023, doi: 10.1080/20508840.2023.2248837.

[7] S. Hasanat Shah, W. Ameer, G. W. Jiao, and A. Amin, "The Impact of Covid-19 Induced Decline in Consumer Durables and Mobility on NO2 Emission in Europe," *Glob. Econ. Rev.*, vol. 50, no. 1, pp. 43–53, 2021, doi: 10.1080/1226508X.2021.1877562.

[8] A. H. Ahmad, P. S. Darmanto, and F. B. Juangsa, "Thermodynamic analysis of ammonia co-firing for low-rank coal-fired power plant," *Int. J. Sustain. Energy*, vol. 42, no. 1, pp. 527–544, 2023, doi: 10.1080/14786451.2023.2208689.

[9] K. R. Vegt, J. E. Elberse, B. T. Rutjens, M. H. Voogt, and F. Baâdoudi, "Impacts of citizen science on trust between stakeholders and trust in science in a polarized context," *J. Environ. Policy Plan.*, vol. 25, no. 6, pp. 723–736, 2023, doi: 10.1080/1523908X.2023.2253164.

[10] R. Mwanza, "Toxic Spaces, Community Voices, and the Promise of Environmental Human Rights: Lessons on the Owino Uhuru Pollution Incident in Kenya," *Nord. J. Hum. Rights*, vol. 38, no. 4, pp. 279–304, 2020, doi: 10.1080/18918131.2021.1904617.

[11] M. Mesene, M. Meskele, and T. Mengistu, "The proliferation of noise pollution as an urban social problem in Wolaita Sodo city, Wolaita zone, Ethiopia," *Cogent Soc. Sci.*, vol. 8, no. 1, 2022, doi: 10.1080/23311886.2022.2103280.

[12] Y. Lyu and L. Yang, "Environmental monitoring and enforcement in China: an economic perspective review," *China Econ. J.*, vol. 17, no. 1, pp. 3–25, 2024, doi: 10.1080/17538963.2023.2300863.

[13] J. Smith, C. Schroeder, K. Smits, J. Lucena, and O. R. Baena, "Pollution, obligation, and care: perspectives from artisanal and small-scale gold mining and farming in rural Colombia," *Tapuya Lat. Am. Sci. Technol. Soc.*, 2023, doi: 10.1080/25729861.2023.2243762.

[14] H. Flatø, "Trust is in the air: pollution and Chinese citizens' attitudes towards local, regional and central

levels of government," *J. Chinese Gov.*, vol. 7, no. 2, pp. 180–211, 2022, doi: 10.1080/23812346.2021.1875675.
[15] Y. M. Lam, "A study of light pollution discourse in Hong Kong," *Vis. Stud.*, no. May, 2022, doi: 10.1080/1472586X.2022.2060302.
[16] G. Serafeim, "Public Sentiment and the Price of Corporate Sustainability," *Financ. Anal. J.*, vol. 76, no. 2, pp. 26–46, 2020, doi: 10.1080/0015198X.2020.1723390.
[17] A. Stasik and D. Jemielniak, "Public involvement in risk governance in the internet era: impact of new rules of building trust and credibility," *J. Risk Res.*, vol. 25, no. 8, pp. 991–1007, 2022, doi: 10.1080/13669877.2020.1864008.
[18] K. Roelich and N. Litman-Roventa, "Public perceptions of networked infrastructure," *Local Environ.*, vol. 25, no. 11–12, pp. 872–890, 2020, doi: 10.1080/13549839.2020.1845131.
[19] S. Y. Lee, Y. Kim, and Y. Kim, "The co-creation of social value: what matters for public participation in corporate social responsibility campaigns," *J. Public Relations Res.*, vol. 32, no. 5–6, pp. 198–221, 2020, doi: 10.1080/1062726X.2021.1888734.
[20] Y. A. Singgalen, "Analisis Sentimen Wisatawan terhadap Kualitas Layanan Hotel dan Resort di Lombok Menggunakan SERVQUAL dan CRISP-DM," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1870–1882, 2023, doi: 10.47065/bits.v4i4.3199.
[21] 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.
[22] Y. A. Singgalen, "Analisis Sentimen Konsumen terhadap Food , Services , and Value di Restoran dan Rumah Makan Populer Kota Makassar Berdasarkan Rekomendasi Tripadvisor Menggunakan Metode CRISP-DM dan," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1899–1914, 2023, doi: 10.47065/bits.v4i4.3231.
[23] Y. A. Singgalen, "Sentiment classification of coral reef 101 content using decision tree algorithm through CRISP-DM," *Int. J. Basic Appl. Sci.*, vol. 12, no. 3, pp. 121–130, 2023.
[24] Y. A. Singgalen, "Analisis Sentimen dan Sistem Pendukung Keputusan Menginap di Hotel Menggunakan Metode CRISP-DM dan SAW," *J. Inf. Syst. Res.*, vol. 4, no. 4, pp. 1343–1353, 2023, doi: 10.47065/josh.v4i4.3917.
[25] Y. A. Singgalen, "Analisis Sentimen Wisatawan terhadap Taman Nasional Bunaken dan Top 10 Hotel Rekomendasi Tripadvisor Menggunakan Algoritma SVM dan DT berbasis CRISP-DM," *J. Comput. Syst. Informatics*, vol. 4, no. 2, pp. 367–379, 2023, doi: 10.47065/josyc.v4i2.3092.
[26] 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.
[27] 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.
[28] A. Bates, M. Lai, and W. Thao, "Grab and gone: expert perspectives on innovation to diffusion of direct air carbon capture and storage technology," *Carbon Manag.*, vol. 14, no. 1, p., 2023, doi: 10.1080/17583004.2023.2235577.
[29] Y. Han, S. Tan, C. Zhu, and Y. Liu, "Research on the emission reduction effects of carbon trading mechanism on power industry: plant-level evidence from China," *Int. J. Clim. Chang. Strateg. Manag.*, vol. 15, no. 2, pp. 212–231, Jan. 2023, doi: 10.1108/IJCCSM-06-2022-0074.
[30] M. O. Faruque, "Foreign investment, mining conflict, and contested development in Bangladesh," *Can. J. Dev. Stud.*, vol. 42, no. 4, pp. 537–555, 2021, doi: 10.1080/02255189.2021.1902289.
[31] Y. Chen and H. Mu, "Analysis of influencing factors of CO_2 emissions based on different coal dependence zones in China," *Econ. Res. Istraz.* , vol. 36, no. 2, p., 2023, doi: 10.1080/1331677X.2023.2177182.
[32] C. Liu, T. Hale, and J. Urpelainen, "Explaining the energy mix in China's electricity projects under the belt and road initiative," *Env. Polit.*, vol. 32, no. 7, pp. 1117–1139, 2023, doi: 10.1080/09644016.2022.2087355.
[33] B. Liu *et al.*, "Evaluating urban and nonurban PM2.5 variability under clean air actions in China during 2010–2022 based on a new high-quality dataset," *Int. J. Digit. Earth*, vol. 17, no. 1, 2024, doi: 10.1080/17538947.2024.2310734.