

Multimodal Sentiment and Emotion Detection in Social Media: Methods, Modalities, and Application Perspectives

Munmun Kakkar* and Hemant Patidar

Department of Electronics and Communication Engineering, Oriental University, Indore, India

*Corresponding author: munmun.kakkar@gmail.com

Abstract. The growing popularity of social media has produced large volumes of multimodal user-generated content in the form of text, images, and speech, creating new opportunities to study how users express sentiment and emotion. This paper presents a narrative review of multimodal sentiment and emotion detection methods designed for social media data. The review covers representation methods for the textual, visual, and speech modalities, together with the machine learning and deep learning models used to analyze each modality and to fuse multimodal representations. The literature is analyzed in terms of feature extraction methods, fusion methods, and application contexts such as mental health monitoring, disaster management, education, and healthcare. The review shows that several weaknesses persist: text-based models still dominate, heterogeneous modalities are poorly integrated, standardized multimodal datasets are lacking, and deep multimodal systems are difficult to interpret. To address these concerns, the paper consolidates methodological knowledge across modalities and frames a unified multimodal analysis perspective that focuses on combining complementary features and on application-specific issues. The review offers a reference framework to guide future research and development in social media sentiment and emotion analysis by consolidating approaches, identifying research gaps, and clarifying the practical requirements of emotion-aware multimodal systems.

Keywords. Multimodal sentiment; emotion detection; social media analytics; deep learning; multimodal fusion

I. INTRODUCTION

Social media platforms have become important channels of communication, where users share opinions, experiences, and emotions through text, images, videos, and audio. The large volume of such user-generated data has increased the importance of sentiment and emotion analysis for understanding public opinion and human behavior.

Traditional sentiment analysis methods focus mainly on textual information. Although these approaches can identify opinions expressed in written content, they often fail to capture the emotional cues present in images, facial expressions, and speech signals. As a result, sentiment interpretation based only on text may be incomplete in many social media scenarios.

Multimodal sentiment analysis combines information from different data sources, such as text, visual content, and audio signals, to achieve a fuller understanding of emotions. Recent developments in machine learning and deep learning have enabled effective extraction and fusion of features from different modalities, leading to improved sentiment prediction.

This narrative review summarizes the major multimodal sentiment and emotion analysis approaches used in social media applications. It discusses commonly used modalities, fusion techniques, application areas, key challenges, and future research directions in the field.

II. LITERATURE REVIEW

Multimodal sentiment and emotion analysis has progressed from text-based sentiment classification to frameworks that integrate textual, visual, and acoustic information. Morency, Mihalcea, and Doshi demonstrated that combining language, facial expressions, and vocal cues improves sentiment prediction compared with unimodal methods [1]. Poria and colleagues integrated semantic textual features with visual descriptors and reported improved classification performance [2], while Abburi and colleagues employed deep neural networks to combine audio and textual features for multimodal sentiment analysis [3].

Advanced fusion strategies then became a major research focus. Zadeh and colleagues introduced the Tensor Fusion Network (TFN), which models unimodal, bimodal, and trimodal interactions for improved sentiment prediction [4], and later proposed the Memory Fusion Network (MFN) to capture temporal dependencies in sequential data [5]. Hazarika and colleagues highlighted the importance of contextual modeling in conversational emotion recognition [6] and later proposed the modality-invariant and modality-specific representation (MISA) framework, which improves stability across domains [7].

Attention-based architectures further enhanced multimodal learning. Poria and colleagues incorporated multimodal attention mechanisms to identify emotionally relevant features [8], whereas Majumder and colleagues pro-

posed hierarchical fusion with context modeling for multimodal sentiment analysis [9]. Tsai and colleagues introduced the Multimodal Transformer, which uses cross-modal attention to learn interactions among the text, audio, and visual streams [10]. Subsequent studies improved multimodal representation learning through routing and attention mechanisms [11], modality-specific representation learning [12], and contrastive learning techniques [13].

Research has also explored modality-specific emotion recognition. Soleymani and colleagues developed a multimodal affect recognition resource that combines facial, acoustic, and physiological signals [14]. Reliable facial emotion recognition models have also been proposed, and Gupta, Kumar, and Tekchandani applied facial emotion analysis to the assessment of learner engagement [15]. Speech features are important in sentiment analysis, and Mittal and colleagues reported improved emotion recognition by fusing facial, textual, and speech cues [16].

Several studies have demonstrated practical applications of multimodal sentiment analysis. Yazdavar and colleagues applied multimodal learning to mental health assessment [17]. Jain and colleagues proposed a real-time sentiment analysis framework that uses multimedia input [18], and Bryan-Smith and colleagues demonstrated the value of real-time social media sentiment analysis for disaster monitoring [19].

Surveys by Baltrušaitis, Ahuja, and Morency [20], Lai and colleagues [21], and Alsiaity and Orji [22] summarized advances in multimodal machine learning and sentiment analysis and highlighted challenges related to data availability, model complexity, explainability, and fairness. The literature shows a clear shift from feature-concatenation methods to transformer-based and representation-learning approaches. Despite notable improvements in sentiment prediction, challenges such as computational cost, limited multimodal datasets, modality alignment, and interpretability continue to drive research in this field.

III. MULTIMODAL SENTIMENT ANALYSIS FRAMEWORK

Based on the reviewed literature, a generalized multimodal sentiment analysis framework consists of five major stages: data acquisition, preprocessing, feature extraction, multimodal fusion, and sentiment classification. The framework integrates textual, visual, and acoustic information to build a combined representation of user sentiment and emotional state. Recent studies indicate that combining multiple modalities improves sentiment prediction accuracy by capturing complementary emotional cues that single-modality analysis often misses.

The first stage collects multimodal data from social media platforms, including textual posts, images, videos, and speech signals. Because social media content is highly diverse and unstructured, preprocessing is applied to improve data quality and consistency. Text preprocessing generally includes tokenization, stop-word removal, and normalization, whereas image and audio preprocessing involve noise reduction, resizing, segmentation, and feature standardization. Effective preprocessing is essential for removing irrelevant information and improving feature quality.

After preprocessing, modality-specific features are extracted. Textual features capture semantic and contextual information through embedding and language representation techniques. Visual features represent facial expressions, object characteristics, and scene-level information, while acoustic features describe emotional properties such as pitch, energy, intensity, and speech dynamics.

The core component of the framework is multimodal fusion, in which heterogeneous features from different modalities are integrated into a unified representation. The objective of fusion is to preserve complementary information while modeling cross-modal relationships. The fused representation is then processed by classification models to predict sentiment polarity or emotion categories such as positive, negative, neutral, happiness, sadness, anger, and fear.

The final output of the framework is a sentiment or emotion prediction that can support applications such as mental health monitoring, healthcare systems, disaster response, social media analytics, and human-computer interaction. The overall framework, shown in Fig. 1, illustrates how complementary information from multiple modalities can improve sentiment understanding, contextual awareness, and emotion recognition.

IV. APPLICATIONS OF MULTIMODAL SENTIMENT AND EMOTION ANALYSIS

The integration of textual, visual, and acoustic information has extended sentiment and emotion analysis to many domains. By using complementary emotional cues from different modalities, multimodal systems provide more reliable and context-aware sentiment interpretation than conventional text-based approaches.

Mental health monitoring is one of the most widely studied application areas. Yazdavar and colleagues demonstrated that combining textual and visual information from social media posts improves the identification of indicators of depression, anxiety, and emotional distress [17]. Multimodal analysis enables the detection of both explicit and implicit emotional expressions, which supports early psychological assessment.

In healthcare, multimodal sentiment analysis has been applied to pain assessment, patient monitoring, and emotion recognition. Integrating facial expressions and speech characteristics can improve the identification of patient discomfort and emotional state. Such systems can help healthcare professionals understand patient responses better.

Disaster and crisis management is another important application area. Social media users frequently share textual updates, images, and videos during emergencies. Multimodal sentiment analysis helps identify public reactions, emotional responses, and the severity of the situation, thereby supporting emergency response and decision-making [19].

Human-computer interaction has also benefited from multimodal emotion recognition. Emotion-aware systems can analyze user sentiment through speech, facial expressions, and textual interactions to provide adaptive responses. These capabilities improve the user experience in virtual assistants, intelligent tutoring systems, and interactive digital platforms.

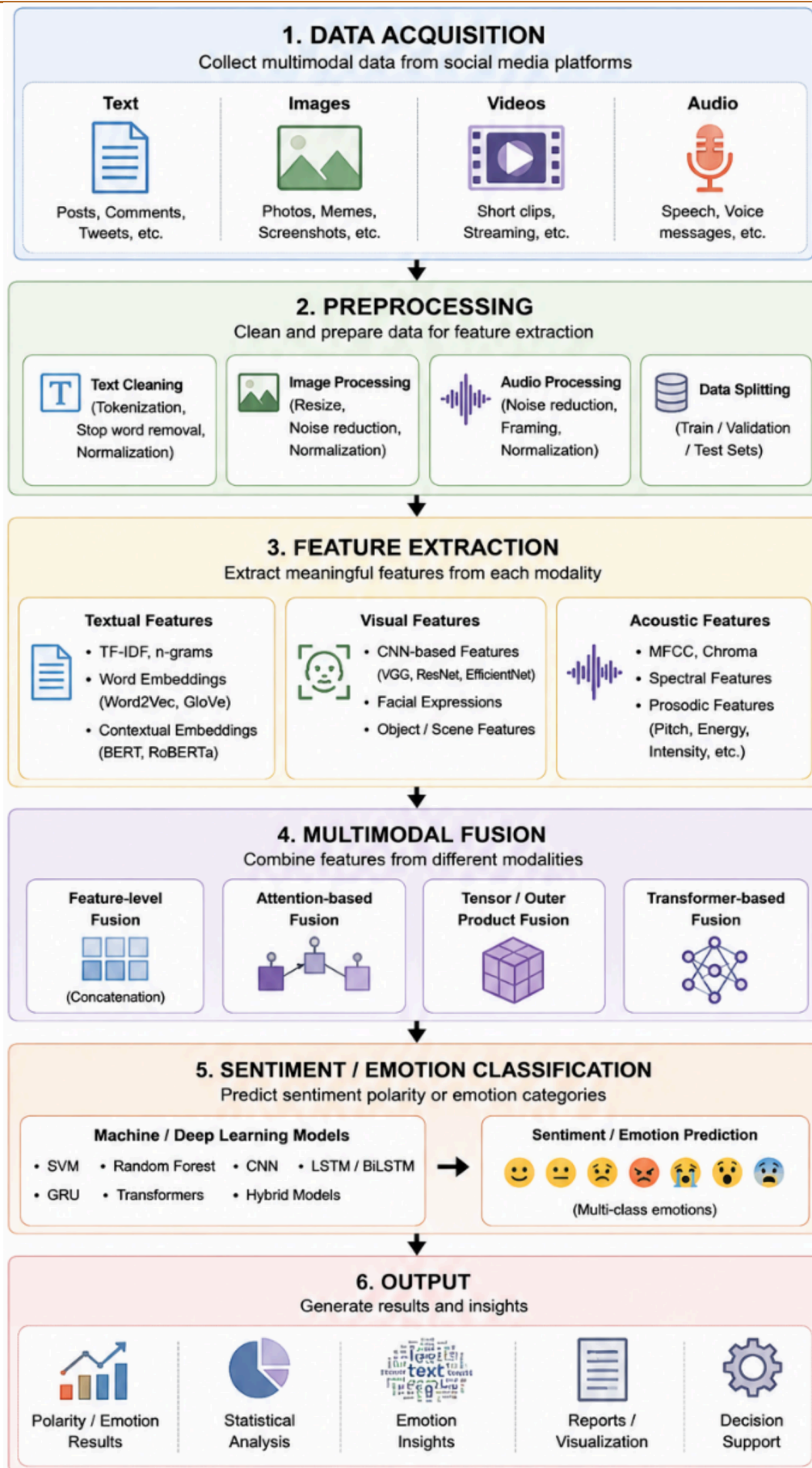


Fig. 1. Generalized framework for multimodal sentiment and emotion analysis.

The reviewed studies indicate that multimodal sentiment analysis offers significant advantages in real-world environments where emotional information is distributed across multiple communication channels. Its ability to capture contextual and affective information makes it a valuable component of next-generation intelligent systems.

V. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Despite significant progress in multimodal sentiment and emotion analysis, several challenges continue to limit its practical deployment in real-world social media environments. One of the primary challenges is the heterogeneity of multimodal data. Text, images, and audio have different structures and characteristics, which makes feature extraction, representation learning, and modality alignment complex.

Another major limitation is the scarcity of large-scale multimodal datasets with reliable sentiment and emotion annotations. Most existing datasets are domain-specific and lack linguistic, cultural, and demographic diversity, which limits model generalization across social media platforms and user communities.

The increasing complexity of multimodal deep learning architectures also raises practical concerns. Advanced fusion techniques, such as transformer-based and attention-based models, often require substantial computational resources, which limits their deployment in real-time applications. In addition, the interpretability of multimodal models remains a challenge, because it is often difficult to identify the contribution of each modality to the final prediction.

Privacy and ethical concerns have also gained importance because of the use of personal images, speech recordings, and social media content. Responsible data collection, user consent, and transparent model development are essential for the ethical use of multimodal sentiment analysis systems.

Future research should focus on lightweight and computationally efficient multimodal architectures, large-scale multilingual datasets, explainable artificial intelligence techniques, and adaptive fusion mechanisms that can handle missing or incomplete modalities. These advances are expected to improve the reliability, scalability, and practical applicability of multimodal sentiment analysis systems.

VI. CONCLUSION

Multimodal sentiment and emotion analysis has emerged as an effective approach for understanding human emotions in social media environments by integrating textual, visual, and acoustic information. The literature reviewed in this paper shows that multimodal approaches consistently achieve better sentiment interpretation and emotion recognition than conventional text-based methods. Recent advances in deep learning, attention mechanisms, and transformer-based architectures have significantly improved multimodal feature representation and fusion.

The review highlights the growing application of multimodal sentiment analysis in areas such as mental health monitoring, healthcare, disaster management, and human-computer interaction. However, challenges related to multimodal data availability, computational complexity, model

interpretability, and privacy remain important research concerns.

Future work should focus on efficient fusion techniques, explainable multimodal models, and large-scale multilingual datasets to improve the reliability and practical deployment of sentiment analysis systems. With continued advances in multimodal learning, sentiment-aware intelligent systems are expected to play an increasingly important role in real-world decision-making and emotion-aware computing applications.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and preparation of this work and approved the final version of the manuscript.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DECLARATION OF GENERATIVE AI USE

The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. The authors take full responsibility for the content of the publication.

DATA AVAILABILITY

No new data were generated or analyzed in this study; this narrative review draws only on previously published literature.

ETHICS STATEMENT

This study is a review of published literature and did not involve human participants or animals; ethical approval was not required.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces (ICMI)*, 2011, pp. 169-176, doi: 10.1145/2070481.2070509.
- [2] Soujanya Poria, Erik Cambria, Amir Hussain, and Guang-Bin Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Netw.*, vol. 63, pp. 104-116, 2015, doi: 10.1016/j.neunet.2014.10.005.
- [3] Harika Abburi, Rajendra Prasath, Manish Shrivastava, and Suryakanth V. Gangashetty, "Multimodal sentiment analysis using deep neural networks," in *Mining Intelligence and Knowledge Exploration (MIKE 2016)*, Lecture Notes in Computer Science, vol. 10089, 2017, pp. 58-65, doi: 10.1007/978-3-319-58130-9_6.
- [4] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2017, pp. 1103-1114, doi: 10.18653/v1/D17-1115.
- [5] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 5634-5641, doi: 10.1609/aaai.v32i1.12021.
- [6] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. 2018 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, 2018, pp. 2122-2132, doi: 10.18653/v1/N18-1193.

- [7] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122-1131, doi: 10.1145/3394171.3413678.
- [8] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. 2017 IEEE Int. Conf. Data Mining (ICDM)*, 2017, pp. 1033-1038, doi: 10.1109/ICDM.2017.134.
- [9] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124-133, 2018, doi: 10.1016/j.knosys.2018.07.041.
- [10] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2019, pp. 6558-6569, doi: 10.18653/v1/P19-1656.
- [11] Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1823-1833, doi: 10.18653/v1/2020.emnlp-main.143.
- [12] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, 2021, pp. 10790-10797, doi: 10.1609/aaai.v35i12.17289.
- [13] Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Factorized contrastive learning: Going beyond multi-view redundancy," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 32971-32998. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/6818dcc65fdf3cbd4b05770fb957803e-Abstract-Conference.html
- [14] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42-55, 2012, doi: 10.1109/T-AFFC.2011.25.
- [15] Swadha Gupta, Parteek Kumar, and Raj Kumar Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimed. Tools Appl.*, vol. 82, no. 8, pp. 11365-11394, 2023, doi: 10.1007/s11042-022-13558-9.
- [16] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 2, 2020, pp. 1359-1367, doi: 10.1609/aaai.v34i02.5492.
- [17] Amir Hossein Yazdavar, Mohammad Saeid Mahdavejad, Goonmeet Bajaj, William Romine, Amit Sheth, Amir Hassan Monadjemi, Krishnaprasad Thirunarayan, John M. Meddar, Annie Myers, Jyotishman Pathak, and Pascal Hitzler, "Multimodal mental health analysis in social media," *PLoS One*, vol. 15, no. 4, Art. no. e0226248, 2020, doi: 10.1371/journal.pone.0226248.
- [18] Rishit Jain, Revant Singh Rai, Sajal Jain, Ruchir Ahluwalia, and Jyoti Gupta, "Real time sentiment analysis of natural language using multimedia input," *Multimed. Tools Appl.*, vol. 82, no. 26, pp. 41021-41036, 2023, doi: 10.1007/s11042-023-15213-3.
- [19] Lydia Bryan-Smith, Jake Godsall, Franky George, Kelly Egoode, Nina Dethlefs, and Dan Parsons, "Real-time social media sentiment analysis for rapid impact assessment of floods," *Comput. Geosci.*, vol. 178, Art. no. 105405, 2023, doi: 10.1016/j.cageo.2023.105405.
- [20] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423-443, 2019, doi: 10.1109/TPAMI.2018.2798607.
- [21] Songning Lai, Xifeng Hu, Haoxuan Xu, Zhaoxia Ren, and Zhi Liu, "Multimodal sentiment analysis: A survey," *Displays*, vol. 80, Art. no. 102563, 2023, doi: 10.1016/j.displa.2023.102563.
- [22] Alaa Alsaity and Rita Orji, "Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions," *Behav. Inf. Technol.*, vol. 43, no. 1, pp. 139-164, 2024, doi: 10.1080/0144929X.2022.2156387.