Author: Meriem Mehri November 2024

AI Against Climate Misinformation

Assessing the far-reaching implications of automated detection on socio-environmental sustainability

Introduction

The rapid spread of climate misinformation on social media threatens public understanding and effective climate action, often undermining support for necessary mitigation policies. Social media platforms, optimized for engagement, have unintentionally amplified contrarian climate narratives and have deepened polarization across political and social lines. This urgent challenge calls for automated systems capable of detecting and categorizing misinformation to enable timely, effective interventions. Addressing this need, *"Hierarchical machine learning models can identify stimuli of climate change misinformation on social media"* [1]explores the Augmented CARDS model, a hierarchical machine learning framework designed to identify and classify climate misinformation on Twitter. Analyzing over 5 million climate-related tweets, the study reveals key triggers of misinformation surges – namely, natural disasters, political events, influential figures – offering insights that inform strategies to enhance automated detection and promote climate literacy.

1. Dataset & model overview: Foundations of the Augmented CARDS approach

The Augmented CARDS (Computer Assisted Recognition of Denial and Skepticism) model was selected for this analysis due to its specialized application in identifying and categorizing climate misinformation on social media, specifically on Twitter. This model, designed to combat the rapid proliferation of climate contrarian claims, aligns well with sustainable development goals by aiming to protect public climate literacy and integrity. The model leverages the Climate Change Twitter Dataset, which contains over 5 million climate-themed tweets collected over a six-month period in 2022. This dataset has been instrumental in both training and testing the model, as it captures a broad spectrum of climate-related discourse, including instances of both verified information and misinformation. By structuring the model as a two-step hierarchical classifier, the Augmented CARDS model first distinguishes between "convinced" and "contrarian" claims before classifying specific types of misinformation.

While the dataset documentation includes basic details on tweet labels and sources, it could benefit from additional information on user demographics, geographic locations, and language diversity, which would enhance reproducibility and contextual understanding across varied populations. Insights into the research team's composition could also promote transparency and inclusivity, crucial for identifying biases in data interpretation. Overall, the Augmented CARDS model and dataset serve as a solid foundation for this analysis, offering a needed solution to mitigate the spread of climate dis/misinformation.

2. Linking AI to sustainable development goals: Impact pathways & alignment

The Augmented CARDS model plays a key role in supporting Sustainable Development Goal (SDG) **13** (Climate Action) and **SDG 16** (Peace, Justice & Strong Institutions), by directly addressing targets that emphasize climate resilience and the promotion of transparent, accurate information. Specifically, SDG 13.3 highlights the importance of improving climate change education, awareness, and institutional capacity, while SDG 16.10 seeks to ensure public access to reliable information. By enabling automated detection and categorization of climate misinformation, the Augmented CARDS model aligns with these objectives by actively mitigating the spread of false narratives that could erode public trust in climate science and diminish support for environmental policies. Such calibration underscores the model's potential to protect climate literacy and reinforce societal resilience against misinformation.

[1] There are 2 versions of this paper. In this coursework, the more concise published version is considered.
Cristian Rojas et al., "Hierarchical machine learning models can identify stimuli of climate change misinformation on social media," *Communications Earth & Environment* (2024), https://doi.org/10.1038/s43247-024-01573-7 **(concise)**
Cristian Rojas et al., "Augmented CARDS: A machine learning approach to identifying triggers of climate change misinformation on Twitter," *arXiv preprint* (2024), https://arxiv.org/abs/2404.15673 **(extended)**

As an enabler, the Augmented CARDS model offers substantial support for advancing SDG targets by equipping platforms and stakeholders with a dependable tool to preemptively counter misinformation. This capability not only protects public discourse but also reinforces educational efforts by promoting fact-based narratives that encourage climate-positive behaviors. However, the model's automated nature presents risks that could inhibit SDG progress. There is a potential for inadvertently censoring legitimate discussions or misclassifying nuanced perspectives, which may introduce unintended biases in information dissemination. Furthermore, heavy reliance on automated systems could reduce human oversight and lead to over-dependence on algorithmic outputs and weakening accountability. Addressing these challenges is essential to ensure the model serves as a net positive force for SDGs 13 and 16, while fostering a balanced approach that respects both free expression and the integrity of climate information.

3. Sustainability analysis of the model's broader impacts

a. Enhancing public trust & equity in climate discourse [Social sustainability]

The Augmented CARDS model contributes significantly to social sustainability by fostering a more informed and aware public, specifically around climate issues. By accurately identifying and categorizing climate misinformation on Twitter, this supervised ML model plays a prominent role in enhancing the quality of life for individuals and communities, both now and in the future. Ensuring accurate climate information is accessible empowers people to make well-informed choices about sustainable practices and concrete policies to support societal resilience against misinformation. However, reliance on Twitter data introduces potential biases, as its user base lacks full demographic and cultural representation, which could inadvertently suppress certain viewpoints, disproportionately affecting marginalized groups. To uphold social sustainability fully, addressing these potential biases is essential, ensuring that the model's outputs reflect a diverse spectrum of perspectives and fostering a democratic and fair approach to climate information flows.

b. Balancing technological gains & resource consumption [Environmental sustainability]

The resource-intensive nature of the augmented model raises concerns about its environmental sustainability. Training and deploying a large-scale machine learning model that processes millions of tweets requires substantial computational power, which translates to significant energy consumption. While the model's environmental benefits lie in its ability to reduce misinformation that could hinder climate action, the energy demands associated with its architecture, training and deployment, pose some questions about its carbon footprint. Environmental costs need to be evaluated carefully to ensure that the technological gains in misinformation mitigation do not come at an unsustainable environmental price. Despite these challenges, the model could yield indirect environmental benefits by fostering an informed public more likely to support climate-positive actions. Reducing misinformation can enhance accurate perceptions of climate risks, potentially inspiring patterns that lower carbon emissions. Integrating sustainable energy sources into the model's operations could mitigate its environmental impact and align deployment with broader ecological goals.

c. Supporting innovation & equitable access [Economic sustainability]

Economically, the Augmented CARDS model presents both opportunities and challenges for sustainability. By addressing climate hoaxes, it has the potential to drive AI innovation beyond social media, expanding applications across science and sustainability. As a pioneer in automated misinformation detection, CARDS can inspire further digital tools for similar challenges in order to enhance AI's role in the digital economy and generating economic activity in climate advocacy and tech sectors. However, equitable access remains a concern. If made open-source, CARDS could empower non-profits, educational institutions, and smaller organizations. Yet, well-funded entities might benefit disproportionately, risking economic disparities. Making sure deployment is balanced across the board, is essential to distribute the model's economic benefits equitably.

4. Additional sustainability considerations: Expanding the scope of impact

Additional factors provide further insight into the model's broader impact on sustainable development. Addressing these questions reveals layered effects, some beneficial and others requiring careful management.

How adaptable is the model to evolving misinformation tactics, and what are the implications for long-term climate literacy?

The model's adaptability is essential for sustainability, especially as fact distortion tactics evolve with new narratives and platforms. A dynamic model that can detect emerging skepticism and misinformation patterns will consistently support climate literacy and reinforce informed public discourse over time. Without such adaptability, the model risks becoming outdated and missing new forms of fact alteration, limiting its effectiveness in a rapidly changing information/news landscape.

Could CARDS model inadvertently stifle constructive debate, potentially leading to self-censorship or overreach in moderation?

Automated detection systems could inadvertently classify valid viewpoints as misinformation, suppress genuine debate and potentially discourage public engagement due to perceived censorship. Such unintended effects may lead to self-censorship, affecting both contrarianism and open discourse dynamics. The model's capacity to discern false claims from constructive dissent is core for maintaining a balanced and inclusive public dialogue.

How might large-scale implementation of this model influence the responsibility and accountability of social media platforms in content generation & management?

As the model scales, its use could shift content moderation responsibility toward automated systems, raising accountability questions. Over-reliance on automation may reduce human oversight, impacting transparency in moderation practices. Balancing automated and human approaches is essential for responsible content management, so that the model upholds transparency, constructive dialogue, and public trust.

5. SDG interaction & cascading effects: Indirect impacts of automated misinformation detection

The Augmented CARDS model's focus on climate misinformation detection not only aligns directly with SDG 13 and SDG 16 but also creates cascading effects on **other SDGs**. By influencing public perceptions, policy support & social norms, the model can indirectly impact several additional goals, both positively and negatively.

SDG 4 (Quality Education): The model's role in reducing climate misinformation promotes informed public discourse, which in turn supports SDG 4, specifically target 4.7: ensuring learners acquire knowledge and skills needed to promote sustainable development. By reducing misleading climate information, the model indirectly supports SDG 4.7, which promotes sustainable knowledge. Limiting ambiguous narratives helps educational systems and societal structures, deliver accurate climate science, spreading through an informed and responsible population – which ultimately creates a positive cycle where improved climate education increases demand for accurate information and reinforces the model's conclusive positive outcomes.

SDG 10 (Reduced Inequalities): The model also intersects with SDG 10, which focuses on reducing inequalities. By providing an automated mechanism to combat 'untruth', the model helps reduce the impact of biased narratives that disproportionately target underserved or vulnerable communities, who are often more susceptible to misinformation due to limited access to alternative credible sources. However, a potential downside is that the model's reliance on English-language and Twitter-specific data may reinforce existing digital divides. Communities less represented on the platform or those communicating in other languages could be underserved, which further amplifies informational inequalities and information reach disparities.

SDG 12 (Responsible Consumption & Production): Tackling climate misinformation can indirectly encourage sustainable consumption patterns, aligning with SDG 12. As a matter of fact, the model's ability to curb misinformation allows individuals to make informed choices and support sustainable consumption and responsible production. By countering false narratives about climate-friendly practices (such as the benefits of renewable energy or sustainable agriculture), the model encourages consumers and businesses alike to engage in behaviors that reduce environmental impact. However, the model's resource-intensive nature poses a potential challenge to SDG 12, particularly if the computational demands are not managed sustainably.

6. Speculative solutions: Mitigating negative ramifications on society, economy & environment

To maximize the positive impact of the Augmented CARDS model on sustainable development while addressing identified challenges, several targeted adjustments and speculative solutions can be proposed. The following suggested changes encompass enhancements to the dataset, model design, governance practices, and an idealized future scenario for dataset & model functionality.

Enhancing dataset diversity & accessibility: Increasing dataset diversity is essential to mitigate bias and improve representational equity. Currently focused on English-language tweets, the dataset could be expanded to include multiple languages, regional variations, and diverse demographics, capturing a wider range of climate perspectives. This broader representation would reduce biases affecting marginalized communities. Additionally, making the dataset accessible to researchers and advocacy groups would democratize climate information, supporting collaboration and transparency.

Technical improvements to the proposed ML model: Adjusting the model architecture to handle evolving misinformation tactics would significantly improve its resilience and relevance. Adaptive learning features would allow the model to recognize new misinformation patterns, maintaining effectiveness over time. Implementing multi-lingual and cross-platform capabilities would further extend the model's utility across diverse social media channels and populations. Enhanced interpretability features would ensure accountability by providing users with clear insights into the model's decision-making process.

Strengthening governance policies & oversight: Governance interventions, including regular audits by climate, journalism & linguistics experts, would help prevent unintended censorship or misclassification. Transparent decision-making guidelines would enable users to understand the model's criteria for misinformation detection, enhancing accountability. Collaborations with educational and climate-focused organizations would align the model's deployment with ethically sound practices that emphasize public benefit.

Speculating on an ideal dataset & possible modeling scenario: In an ideal scenario, the dataset for the model would be universally representative and encompass real-time data from diverse social media platforms and public forums globally, with balanced demographic and linguistic coverage. A 'utopian' dataset would include enriched metadata, such as geographical, cultural, and contextual details, which would enable the model to interpret information with nuanced insights into regional climates and concerns. The model would incorporate self-monitoring to address biases continuously to maintain agility. It would also clarify inaccuracy or propaganda patterns, such as fallacies or manipulative tactics, while allowing collective and individual users to identify misinformation independently and discern fact from fiction autonomously.

References

- Ranney, M. A., & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science, 8*(1), 49–75.
- Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges, 1*(2), 1600008.
- Geiger, N., & Swim, J. K. (2016). Climate of silence: Pluralistic ignorance as a barrier to climate change discussion. *Journal of Environmental Psychology, 47*, 79–90.
- Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE, 12*(5), e0175799.
- Ross Arguedas, A. A., Badrinathan, S., Mont'Alverne, C., Toff, B., Fletcher, R., & Nielsen, R. K. (2022). "It's a battle you are never going to win": Perspectives from journalists in four countries on how digital media platforms undermine trust in news. *Journalism Studies, 23*(14), 1821–1840.
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236.
- Martens, B., Aguiar, L., Gomez-Herrera, E., & Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news. *European Commission Joint Research Centre.*
- Tsfati, Y., Boomgaarden, H. G., Strömbäck, J., Vliegenthart, R., Damstra, A., & Lindgren, E. (2020). Causes and consequences of mainstream media dissemination of fake news: Literature review and synthesis. *Annals of the International Communication Association, 44*(2), 157–173.
- Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., & Quattrociocchi, W. (2022). Growing polarization around climate change on social media. *Nature Climate Change, 12*(12), 1114–1121.
- Jang, S. M., & Hart, P. S. (2015). Polarized frames on "climate change" and "global warming" across countries and states: Evidence from Twitter big data. *Global Environmental Change, 32*, 11–17.
- Pearce, W., Holmberg, K., Hellsten, I., & Nerlich, B. (2014). Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC Working Group 1 report. *PLOS ONE, 9*(4), e94785.
- Anderson, A. A., & Huntington, H. E. (2017). Social media, science, and attack discourse: How Twitter discussions of climate change use sarcasm and incivility. *Science Communication, 39*(5), 598–620.
- Floridi, L., & Cowls, J. (2019). *A unified framework of five principles for AI in society. Harvard Data Science Review, 1(*1)
- Pérez-Ortiz, M., & Adeel, A. (2022). Planet-centered artificial intelligence: Responsible technology for the Anthropocene. *AI and Society, 37*(1), 1–12.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: *Mapping the debate. Big Data & Society, 3*(2), 1–21.
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22–36.
- Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., & Yu, C. (2015). The quest to automate fact-checking. In *Proceedings of the 2015 Computation+ Journalism Symposium*.
- Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics, 10*, 178–206.
- Boussalis, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change, 36*, 89–100.
- Farrell, J. (2016). Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences, 113*(1), 92–97.
- Stecula, D. A., & Merkley, E. (2019). Framing climate change: Economics, ideology, and uncertainty in American news media content from 1988 to 2014. *Frontiers in Communication, 4*, 6.
- Alhindi, T., Chakrabarty, T., & Musi, E. (2023). Multitask instruction-based prompting for fallacy recognition.
- Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., & Mihalcea, R. (2022). Logical fallacy detection.
- Zanartu, F., Cook, J., Wagner, M., & Gallego, J. G. (2023). Automatic detection of fallacies in climate change misinformation.
- Coan, T., Boussalis, C., Cook, J., & Nanko, M. (2021). Computer-assisted detection and classification of misinformation about climate change.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171–4186.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692.*
- He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654.*
- Pradhan, P., Costa, L., Rybski, D., Lucht, W., & Kropp, J. P. (2017). A systematic study of Sustainable Development Goal (SDG) interactions. *Earth's Future, 5*(11), 1169–1179.
- Cook, J. (2020). Deconstructing climate science denial. In *Research Handbook on Communicating Climate Change*, 62–78.
- Climate Action Against Disinformation, Friends of the Earth, & Greenpeace. (2023). Climate of misinformation: Ranking big tech.
- Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition, 6*(4), 353–369.
- Chinn, S., Hart, P. S., & Soroka, S. (2020). Politicization and polarization in climate change news content, 1985–2017. *Science Communication, 42*(1), 112–129.
- Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences, 116(*7), 2521–2526.
- Twitter Climate Change Sentiment Dataset. Retrieved from [https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset].