Propaganda Detection Meets Generative AI: A Systematic Review of Computational Approaches on Social Media

Cristiane Melchior
Department of Social Sciences
LUT University
Lappeenranta, Finland
e-mail: cristiane.melchior@lut.fi

***Abstract*—The proliferation of Generative Artificial Intelligence (GenAI) has fundamentally transformed the propaganda landscape on Social Media Platforms (SMPs), shifting the threat from overt disinformation to subtle, AI-mediated narrative manipulation. Despite this urgency, the field lacks a unified synthesis of the computational defenses developed to counter it. This extended abstract summarizes a Systematic Literature Review (SLR) analyzing 60 peer-reviewed articles published between 2018 and 2026. We map the state-of-the-art through an original taxonomy with three subtopics and nine categories covering computational methods, propaganda conceptualizations, and task formulation. Our key findings expose a *methodological lag* (Large Language Models appear in only 13% of studies), a *proxy tension* (84% of datasets are platform-specific, but the most reused are news-based), and a profound *theoretical deficit* (28% of studies lack a formal definition and theoretical framework of propaganda). These results reveal structural vulnerabilities in current detection ecosystems with direct consequences for trust, fairness, and algorithmic resilience in AI-based media environments.**

Keywords-propaganda detection; social media platforms; disinformation; computational methods; natural language processing; large language models; algorithmic bias; trust;

I. Introduction

The digital information ecosystem has reached a critical inflection point. Social Media Platforms (SMPs) have fundamentally altered the scale, velocity, and efficacy of propaganda dissemination, creating systemic risks for democratic institutions, public health, and geopolitical stability [1][2]. Crucially, the threat has evolved: generative AI no longer floods platforms with overt falsehoods, but instead inserts subtle, high-fidelity narratives designed to slip unnoticed into everyday digital discourse, a paradigm shift in international influence operations [3].

Propaganda, broadly defined as the strategic management of perceptions to achieve a predetermined objective [4], is now AI-assisted. Yet the scientific community lacks a consolidated understanding of how computational defenses, Natural Language Processing (NLP), Machine Learning (ML), and Large Language Models (LLMs), are being deployed to counter it. Individual studies address these methods in isolation, leaving critical structural vulnerabilities uncharted: Are LLMs keeping pace with AI-generated propaganda? Are detection models valid for the platforms they claim to protect? Do these systems rest on theoretical foundations rigorous enough to support ethical deployment?

This extended abstract summarizes a full-length Systematic Literature Review (SLR) that addresses these questions by mapping the state-of-the-art across 60 empirical studies. The work makes three primary contributions: (1) an original multi-level taxonomy of propaganda research using computational methods; (2) the identification of systemic structural weaknesses in the current detection ecosystem; and (3) a strategic roadmap for the field, advocating for theoretically grounded, multimodal, and explainable algorithmic resilience.

The remainder of this work is structured as follows. In Section II, we describe the methodology used to conduct our systematic review of the literature. In Section III, we present our new taxonomy and outline the main findings of the study. In Section IV, we discuss what these findings mean for trust and algorithmic resilience, offering a roadmap for future research. Finally, in Section V, we conclude the research with a brief summary of our work.

II. Methodology

We conducted an SLR following established protocols [5], covering seven academic databases: Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Wiley Online Library, Emerald Insight, and Google Scholar. The search string combined computational terms (*"machine learning," "NLP," "LLMs"*) with the target construct (*"propaganda"*) and the platform context (*"social media"*). Searches were executed between January and March 2026.

After removing 124 duplicates from an initial pool of 263 records and applying a five-criterion exclusion protocol, filtering out non-primary sources, off-topic studies (e.g., hate speech, fake news, bots), purely qualitative work, low-quality papers, and literature reviews, the final corpus comprised **60 articles**. All included studies passed a custom Quality Assessment Criteria (QAC) instrument scored on a 12-point scale covering objective clarity, research design, method justification, algorithmic reporting, and findings rigor (threshold: $\geq$6/12; mean achieved: 10.2/12). The corpus spans 2018–2026, with 75% of studies published in 2023 or later, confirming the field's rapid growth phase.

III. Taxonomy and Key Findings

We developed a **taxonomy** organizing the field into three subtopics, as illustrated in Figure 1, each decomposed into

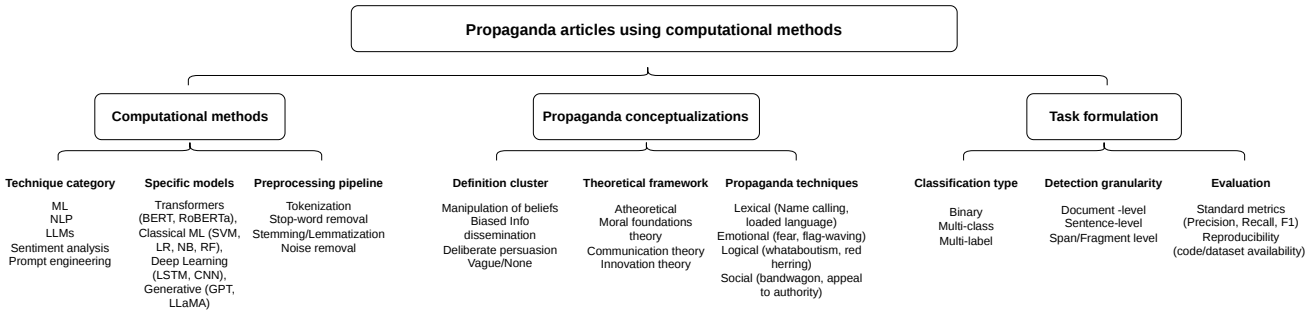

Figure 1. Taxonomy of the propaganda articles using computational methods

three categories (nine total): (i) *Computational Methods*: technique categories (ML, NLP, LLMs), specific model families, and pre-processing pipelines; (ii) *Propaganda Conceptualizations*: definitional clusters, theoretical frameworks, and typologies of propaganda techniques; (iii) *Task Formulation*: classification type (binary, multi-class, multi-label), detection granularity (document, sentence, span), and evaluation metrics.

Our analysis highlights four interconnected findings with direct relevance to trust and ethics in AI-based media systems:

Finding I: Methodological lag. Despite generative AI's centrality to the propagandistic threat, LLMs appear in only 13% of the corpus ($n = 8$), all from 2023 onward. The field remains dominated by Bidirectional Encoder Representations from Transformers(BERT)-family transformers ($n = 23$) and classical classifiers (Logistic Regression $n = 14$; SVM $n = 12$). This lag is dangerous: detection architectures risk obsolescence precisely as the threat modernizes.

Finding II: Proxy tension. Of 51 identified datasets, 84% were used by a single study only. The two most reused datasets, PTC-SemEval20 Task 11 and Qprop (used by 6 and 4 studies, respectively), are derived from *news articles*, not social media posts. Models trained on journalistic data are misapplied to the conversational, fragmented, and multilingual nature of SMP content, compromising both external validity and deployment fairness.

Finding III: Theoretical deficit. 28% of studies ($n = 17$) operationalize propaganda without any formal definitional grounding; only 5% ($n = 3$) anchor detection frameworks in established communication or persuasion theory. Black-box classifiers built on undefined targets cannot be ethically or reliably deployed in content moderation or regulatory contexts.

Finding IV: Platform and geographic bias. X (formerly Twitter) accounts for 57% of all studies ($n = 34$), while TikTok, arguably the most influential contemporary SMP, appears in a single study. Research is dominated by English-language data focused on Russia, the USA, and Ukraine. The Global South and non-Latin scenario remain understudied, introducing geographic and linguistic bias into propaganda detection tools.

IV. Implications for Trust and Algorithmic Resilience

Taken together, our findings expose a structural mismatch between the sophistication of AI-mediated propaganda and the computational defenses designed to counter it. The consequences are not merely academic: detection systems underpinned by mismatched training data, absent theoretical definitions, and outdated architectures constitute unreliable, and potentially unfair, tools for the platform governance and public policy.

The *proxy tension* raises algorithmic fairness concerns that align directly with emerging AI regulatory frameworks (e.g., the EU AI Act), which requires demonstrable validity and non-discrimination in high-risk AI applications. The *theoretical deficit* similarly undermines the transparency and explainability requirements increasingly mandated for AI applications deployed in public interest contexts.

For practitioners and researchers, our findings argue for: (1) urgent adoption of LLMs and multimodal analysis for detecting image-text propaganda formats (e.g., memes and deepfakes); (2) platform-native dataset construction, moving beyond news data; and (3) interdisciplinary collaboration integrating political science, communication theory, and computational methods to supply the missing conceptual grounding.

V. Conclusion

Our SLR provides, to the authors' knowledge, the first comprehensive synthesis of propaganda articles using computational approaches to analyze propaganda on SMPs. By exposing a *methodological lag*, a *proxy tension*, and a *theoretical deficit*, it demonstrates that technically sophisticated detection coexists with profound contextual and rhetorical underdevelopment. The full paper delivers a strategic roadmap, grounded in the nine-category taxonomy, for transitioning from black-box classification to theoretically informed, multimodal, explainable, and platform-appropriate systems. As generative AI continues to erode the boundaries between authentic and artificial content, this roadmap is not a future aspiration but an operational necessity for trustworthy, fair, and resilient media environments.

REFERENCES

[1] P. N. Ahmad et al., "Hierarchical graph-based integration network for propaganda detection in textual news articles on social media," *Scientific Reports*, vol. 15, no. 1, p. 1827, 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-024-74126-9

[2] C. Melchior, T. Warin, and M. Oliveira, "An investigation of the COVID-19-related fake news sharing on Facebook using a mixed methods approach," *Technological Forecasting and Social Change*, vol. 213, p. 123 969, 2025.

[3] B. J. Goldstein and B. V. Benson, "The Era of A.I. Propaganda Has Arrived, and America Must Act," *The New York Times, Opinion*, 2025, ISSN: 0362-4331.

[4] D. G. S. Jowett and V. O'Donnell, *Propaganda And Persuasion*. Thousand Oaks, Calif: Sage Publications, Inc, 2006, 448 pp., ISBN: 978-1-4129-0898-6.

[5] A. Khan, S. Krishnan, and A. Dhir, "Electronic government and corruption: Systematic literature review, framework, and agenda for future research," *Technological Forecasting and Social Change*, vol. 167, p. 120 737, 2021, ISSN: 0040-1625. DOI: 10.1016/j.techfore.2021.120737