AIMedia 2026
Nice, France · July 2026

EXTENDED ABSTRACT · SYSTEMATIC LITERATURE REVIEW

# Propaganda Detection Meets Generative AI

## A Systematic Review of Computational Approaches on Social Media

**Cristiane Melchior**
Department of Social Sciences · LUT University · Lappeenranta, Finland
cristiane.melchior@lut.fi

# Why This Matters

- Social media platforms have changed the **scale, speed, and reach** of propaganda, creating systemic risks for democracy, public health, and geopolitical stability.
- Generative AI has **shifted the threat** — from overt false content to subtle, high-fidelity narratives that blend into everyday online discourse.
- Computational defenses (NLP, ML, LLMs) are studied **in isolation**; the field lacks a unified synthesis.

## RESEARCH QUESTIONS

1. Are LLMs keeping pace with AI-generated propaganda?
2. Are detection models valid for the platforms they aim to protect?
3. Do these systems rest on sound theoretical foundations?

# Our Contributions

## 01

### An original taxonomy

A multi-level taxonomy of propaganda research using computational methods.

## 02

### Structural weaknesses

Identification of systemic vulnerabilities in the current detection ecosystem.

## 03

### A strategic roadmap

Toward theoretically grounded, multimodal, and explainable detection.

# Methodology — Systematic Literature Review



**263**
records identified

→

**−124**
duplicates removed

→

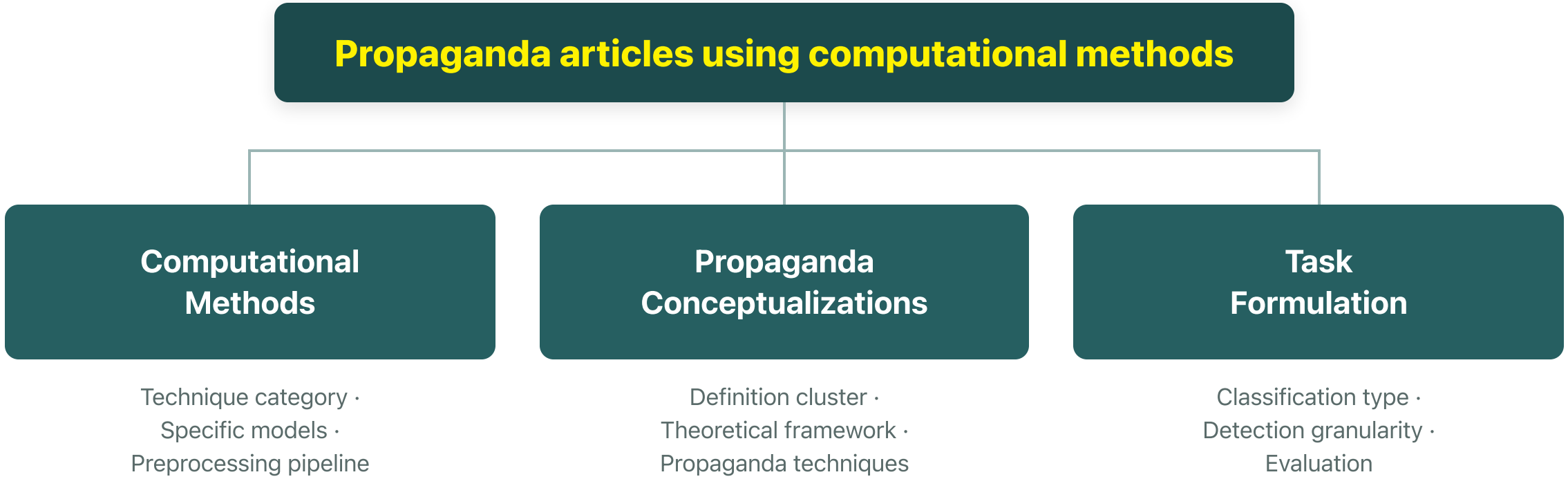
**5-criterion exclusion**
screening

→

**60**
final corpus

- **7 databases:** Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Wiley, Emerald, Google Scholar.
- **Search string:** computational terms × "propaganda" × "social media"; run Jan–Mar 2026.
- **Quality assessment:** 12-point scale (threshold ≥6; mean achieved 10.2).
- **Corpus span:** 2018–2026 — 75% of studies published in 2023 or later.

# A Taxonomy: Three Subtopics, Nine Categories



# Taxonomy (1 of 3) — Computational Methods

PROPAGANDA ARTICLES USING COMPUTATIONAL METHODS

## Computational Methods

### Technique category

- ML
- NLP
- LLMs
- Sentiment analysis
- Prompt engineering

### Specific models

- Transformers (BERT, RoBERTa)
- Classical ML (SVM, LR, NB, RF)
- Deep Learning (LSTM, CNN)
- Generative (GPT, LLaMA)

### Preprocessing pipeline

- Tokenization
- Stop-word removal
- Stemming / Lemmatization
- Noise removal

# Taxonomy (2 of 3) — Propaganda Conceptualizations

PROPAGANDA ARTICLES USING COMPUTATIONAL METHODS

## Propaganda Conceptualizations

### Definition cluster

- Manipulation of beliefs
- Biased info dissemination
- Deliberate persuasion
- Vague / None

### Theoretical framework

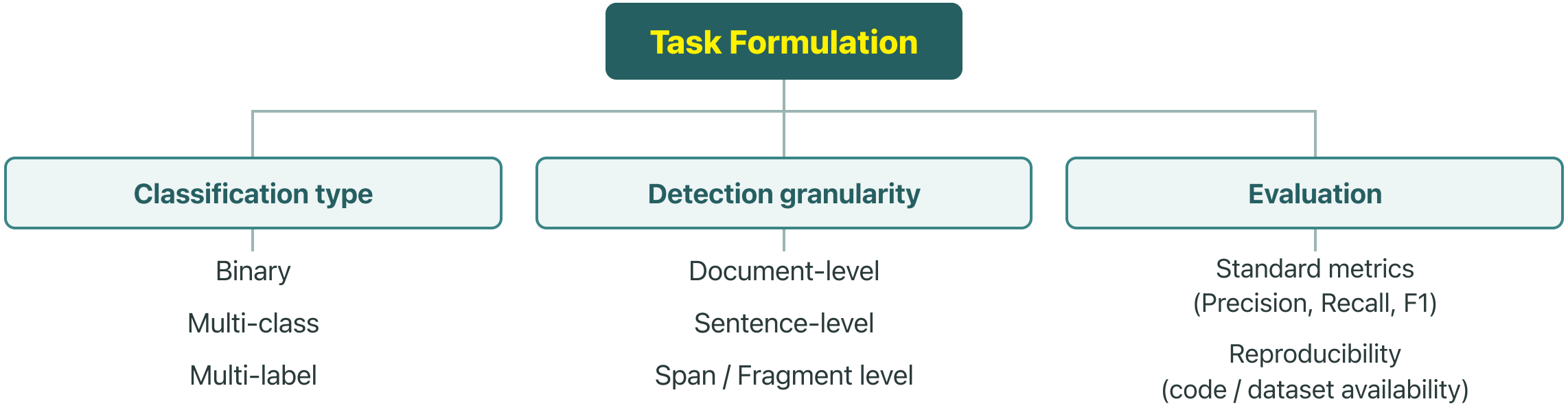
- Atheoretical
- Moral foundations theory
- Communication theory
- Innovation theory

### Propaganda techniques

- Lexical (name calling, loaded language)
- Emotional (fear, flag-waving)
- Logical (whataboutism, red herring)
- Social (bandwagon, appeal to authority)

# Taxonomy (3 of 3) — Task Formulation

PROPAGANDA ARTICLES USING COMPUTATIONAL METHODS



# Key Findings (1 of 2)



**13%**

use LLMs (n=8)

## Finding I — Methodological lag

LLMs appear in only 13% of the corpus, all from 2023 on. The field is still dominated by BERT-family transformers (n=23) and classical models (LR n=14, SVM n=12). Detection risks **obsolescence** as the threat modernizes.

**84%**

datasets used once

## Finding II — Proxy tension

Of 51 datasets, 84% were used by a single study. The two most reused (PTC-SemEval20, Qprop) are built from **news articles, not social-media posts** — undermining external validity and deployment fairness.

**28%**

no definition (n=17)

## Finding III — Theoretical deficit

28% of studies operationalize propaganda with **no formal definition**; only 5% (n=3) anchor detection in established communication or persuasion theory. Black-box classifiers on undefined targets cannot be deployed responsibly.

**57%**

study X / Twitter (n=34)

## Finding IV — Platform & geographic bias

X (Twitter) accounts for 57% of studies; TikTok appears only once. Data is mostly English, centered on Russia, the USA, and Ukraine — the **Global South remains understudied**.

# Implications for Trust & Algorithmic Resilience



## WHAT IT MEANS

- A **structural mismatch** between sophisticated AI propaganda and the defenses built to counter it.
- The proxy tension raises fairness concerns aligned with the **EU AI Act** (validity, non-discrimination).
- The theoretical deficit undermines **transparency and explainability** requirements.

## RECOMMENDATIONS

1. Adopt LLMs and multimodal analysis for image-text formats (memes, deepfakes).
2. Build platform-native datasets, moving beyond news data.
3. Integrate political science, communication theory, and computational methods.

# Conclusion



- The first comprehensive synthesis of computational propaganda detection on social media.
- It reveals a **methodological lag**, a **proxy tension**, and a **theoretical deficit**.
- Technically advanced detection coexists with **contextual and rhetorical underdevelopment**.

The roadmap — from black-box classification to **theory-informed, multimodal, explainable, platform-appropriate** systems — is not a future aspiration but an **operational necessity** for trustworthy and resilient media.




# Thank You

Questions & discussion are warmly welcome.

**Cristiane Melchior** · cristiane.melchior@lut.fi · LUT University, Lappeenranta, Finland