The Rhetoric of Hate in Algorithms: A Critical Analysis of Nazi Propaganda Through the Lens of Artificial Intelligence

Nuno Santos, January 31, 2026

By Nuno Santos (with AI assistance in research and editing)

aipropagandanazi.com

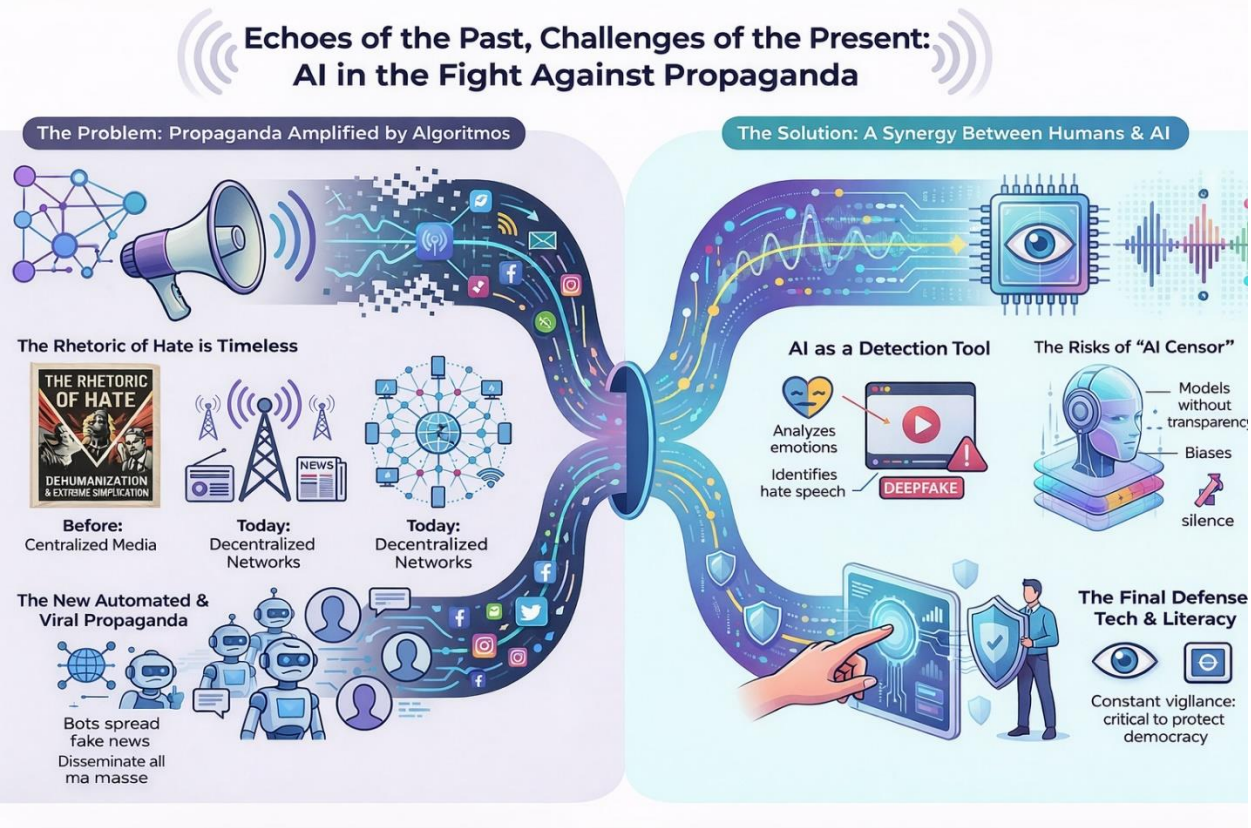
Abstract

This article proposes a critical analysis of historical propaganda, using the Nazi regime as a case study to explore the role of Artificial Intelligence (AI) in deconstructing hate narratives and detecting contemporary manipulative content. The methodology integrates classical rhetorical analysis with computational approaches in Natural Language Processing (NLP) and Machine Learning (ML). The study focuses on identifying patterns of dehumanization, modeling the algorithmic dissemination of such patterns, and evaluating the effectiveness of AI in developing ethical countermeasures. The results highlight the continuity of historical rhetorical patterns in the digital ecosystem and the urgency of a Human-AI synergy for preserving democratic integrity.

Research Question: *How can AI detect and classify rhetorical patterns of dehumanization derived from historical propaganda in contemporary digital ecosystems?*

Scope: *This study examines computational methods for propaganda analysis without training empirical models, focusing instead on theoretical frameworks and documented case studies.*

Keywords: Nazi Propaganda; Rhetoric; Artificial Intelligence; Discourse Analysis; Hate Speech; Computational Propaganda.



Graphical Abstract

Figure 1: Conceptual framework of the dual role of AI in the dissemination and detection of computational propaganda. The diagram illustrates the transition from centralized historical media to decentralized digital networks and the necessary synergy between technology and human literacy.

1. Introduction: The Paradox of Scale

In the 21st century, the integrity of democratic systems is increasingly shaped by the ways information is produced, algorithmically distributed, and consumed. Contemporary informational manipulation no longer depends exclusively on centralized censorship apparatuses but rather on decentralized digital networks and algorithms that are fed by human behavioral patterns. In this scenario, Artificial Intelligence (AI) assumes a dual role: it is the necessary tool for monitoring data flows that humans alone cannot process, but it is also the engine that can intensify polarization.

Historical precedent indicates that manipulation is deliberate and strategic. By analyzing Nazi propaganda, we understand how constructed narratives can fabricate enemies and justify atrocities

(Welch, 2004). This article explores the development of AI models dedicated to identifying these patterns, evaluates their technical foundations and ethical risks, and examines their role as a shield for democracy.

2. Theoretical and Technical Framework

2.1. From Aristotelian Rhetoric to Computational Propaganda

Nazi propaganda subverted the pillars of classical rhetoric for destructive purposes:

- Ethos (Credibility): Built around the myth of the infallible leader (Führerprinzip).
- Logos (Logic): Replaced by simplistic and fallacious arguments, reducing complex problems to single causes.
- Pathos (Emotion): The central pillar, exploiting fear, anger, and national pride to isolate specific groups.

Today, these principles manifest in Computational Propaganda, defined as the use of algorithms, automation, and large data volumes to manipulate public opinion (Howard et al., 2023). This includes the use of bots and automation to simulate popular support and amplify emotionally intense content through recommendation algorithms.

2.2. The Technological Lens: NLP and Transformers

To quantify patterns that had previously been analyzed only qualitatively, AI uses Natural Language Processing (NLP) tools.

- Word Embeddings: Numerical representations projecting words onto a semantic map. Words with similar meanings appear close in this vector space, enabling AI to identify hierarchical and prejudicial relationships.
- Transformer Models (BERT, GPT, RoBERTa): These architectures use self-attention mechanisms to capture contextual nuances and irony, making them effective in detecting hate speech at a global scale.

2.3. Methodology: Analytical Design

This study employs a qualitative rhetorical analysis combined with a computational framing approach. The analytical focus encompasses:

- Analysis Object: Linguistic patterns of dehumanization and dissemination mechanisms in historical and contemporary propaganda.
- Approach: Integration of classical rhetorical analysis (Ethos, Pathos, Logos) with computational propaganda frameworks and NLP techniques.
- Data Sources: Historical propaganda materials, documented contemporary case studies, and existing academic literature on AI detection systems.
- Limitations: This study does not involve empirical model training or quantitative dataset analysis but focuses on theoretical frameworks and documented applications.

The methodological framework enables replication by applying the same rhetorical and computational categories to other historical or contemporary propaganda contexts.

3. Analysis Axes: Patterns and Dissemination

3.1. Dehumanization and Sentiment

AI-driven NLP systems enable scalable identification of dehumanization strategies via linguistic markers, including contamination metaphors, animalization, and existential-threat framing. These systems quantify the frequency of language that denies humanity to the "other," treating minorities as "plagues" or "existential threats." Sentiment analysis algorithms map strategic appeals to fear and moral panic, identifying content with high manipulative potential before factual verification.

Research has demonstrated that Nazi propaganda systematically used mental state language to deny cognitive and emotional capacities to targeted groups, a pattern reproduced in contemporary online hate speech (Landry, 2022).

3.2. Dissemination and Microtargeting

Unlike the unidirectional dissemination of Nazi radio and cinema, modern informational manipulation is micro-targeted. Digital platforms use granular data to personalize narratives, continuously refining them for specific user bubbles.

AI is also crucial for detecting visual manipulation and identifying inconsistencies in deepfakes generated by Generative Adversarial Networks (GANs), such as incoherent lighting or impossible microexpressions.

Table 1

Propaganda Mechanisms: Nazi vs. Algorithmic

Technique	**Nazi Propaganda Example**	**Modern Algorithmic Parallel**	**AI Detection Approach**
Dehumanization	Jews depicted as vermin, disease, or sub-human	Migrants are portrayed as "invaders" and "parasites."	Mental state language analysis (Landry, 2022)
Emotional Amplification	Mass rallies, emotional speeches exploiting fear	Algorithmic prioritization of outrage-inducing content	Sentiment analysis and emotion detection
Repetition & Saturation	Constant messaging via radio, posters, and cinema	Bot networks creating illusion of consensus	Bot detection algorithms, coordinated behavior analysis
Visual Manipulation	Staged photos, propaganda posters	Deepfakes, synthetic media, manipulated imagery	GAN-detection, forensic image analysis

Note. This table illustrates structural continuities between historical and contemporary propaganda mechanisms, despite technological differences in dissemination.

4. Contemporary Case Studies

Historical analysis suggests that rhetorical patterns from Nazi propaganda persist in modern crises:

- Pandemic (2020-2021): Disinformation networks used appeals to fear and the creation of "internal enemies" to discredit vaccines. Meta's Transparency Report (2021) and Twitter Safety Report (2021) documented coordinated bot networks amplifying anti-vaccination narratives on Facebook, Telegram, and YouTube, employing classic propaganda techniques of scapegoating and extreme simplification.
- Russia-Ukraine War: EU reports documented the circulation of manipulated videos and polarizing narratives disseminated by coordinated networks (EUvsDisinfo, 2024). Deepfakes of political and military leaders, old conflict footage presented as current, demonstrating the convergence of generative technologies and transnational propaganda.
- Portuguese Context: Fact-checking organizations reported adaptation of imported narratives. Specific documented cases include: (1) Videos alleging destruction of ballots during the 2021-2024 legislative and municipal elections, debunked by Polígrafo (2021, 2024); (2) Images of supposedly manipulated ballot boxes, taken from previous contexts or other countries, verified as false by Observador (2022); (3) Xenophobic statements falsely attributed to politicians, amplified on TikTok with high emotional charge. Research by OberCom (2023) and EUvsDisinfo (2024) confirms the circulation of adapted narratives targeting Portuguese audiences, employing techniques that mirror historical patterns of propaganda.

5. Ethical Challenges and AI Explainability

5.1. Data Quality and Bias

Models trained with biased datasets generate unfair classifications. A critical gap in robust corpora for European Portuguese limits the effectiveness of tools in the Portuguese context. While datasets like HateBR exist for Brazilian Portuguese, European Portuguese remains underrepresented in hate speech detection research, creating potential for cultural and linguistic drift in model performance.

5.2. The Black Box Risk

Autonomous moderation can lead to undue censorship or silencing of political opposition. Without transparency, algorithmic power becomes susceptible to systemic biases—penalizing minorities or confusing political criticism with manipulation—and authoritarian abuse.

5.3. Explainable AI (XAI) as a Solution

Explainable AI techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive exPlanations) enable auditing of model decisions, revealing which words or patterns led to a prediction, thereby ensuring transparency and fairness (Mehta & Passi, 2022). Contemporary models achieve F1 Scores between 0.75 and 0.90 (Jahan & Al-Hasan, 2023), yet explainability remains essential for democratic accountability.

Big Tech's business model exacerbates this scenario, as profit from engagement often collides with the need for rigorous moderation. Cases such as Cambridge Analytica, the Rohingya genocide in Myanmar facilitated by Facebook, and the January 6 Capitol attack amplified by platform algorithms expose serious failures in content moderation (Bolsover & Howard, 2017).

5.4. Limitations

This study acknowledges several limitations:

- Lack of Empirical Model Training: Analysis relies on existing literature and documented applications rather than original computational experiments.
- Dataset Limitations: Scarcity of European Portuguese hate speech corpora limits the generalizability of findings to the Portuguese context.
- Cultural and Linguistic Drift: Rhetorical patterns and hate speech markers vary significantly across cultures, languages, and temporal contexts.
- Risk of Over-Moderation: Automated systems may suppress legitimate political discourse or minority voices while pursuing harmful content detection.

6. Conclusion and Action Roadmap

Informational manipulation has become more sophisticated and distributed than in any historical period. AI emerges as an indispensable defense tool, but its success depends on ethical and political balance.

Proposed Action Roadmap:

- Governments: Implement independent audits of moderation systems, establish international transparency standards, provide graduated sanctions for platforms that

conceal data or refuse audits, and fund creation of open data resources in European Portuguese.

- Platforms: Guarantee algorithmic transparency through API access for research, publish periodic reports on metrics, biases, and error rates, integrate explanation tools for algorithmic decisions, and create swift appeal mechanisms for users unfairly penalized.
- Academia: Develop explainable AI methodologies for content moderation, create datasets for European Portuguese and international comparative studies, and investigate historical continuities between historical and contemporary propaganda.
- Citizens: Promote media literacy from basic education, stimulate fact-checking practices before emotional content sharing, and develop critical awareness about emotionally intense content and algorithmic recommendation systems.

Just as Nazi propaganda reminds us of the dangers of institutionalized manipulation, the present demands constant vigilance to ensure that the systems we create reinforce—rather than undermine—truth, freedom, and human dignity. Media literacy and algorithmic transparency constitute the most effective defense against the rhetoric of hate, whether conveyed through a historical poster or a modern algorithm.

APPENDICES

Appendix A – Technical Foundations

A.1. Word Embeddings and Semantic Representations

Word embeddings are numerical representations projecting words into a multidimensional space. Words with similar meanings tend to occur closer together, with stable linguistic relationships (e.g., gender, hierarchy) emerging as consistent directional vectors. For example, the vector from "king" to "queen" parallels the vector from "man" to "woman," revealing learned gender relationships.

A.2. Transformer Architecture

Models such as BERT, RoBERTa, and GPT employ self-attention mechanisms to capture contextual relationships, long-range dependencies, and rhetorical nuances. Their ability to analyze language at multiple levels simultaneously makes them well-suited to hate speech detection, propaganda analysis, and the identification of manipulation.

A.3. Key Datasets and Evaluation Metrics

Research depends on manually annotated corpora:

- Jigsaw Toxic Comments Dataset – Comments labeled for toxicity, threats, and insults
- HateXplain – Includes human explanations, facilitating XAI research
- OLID/OffensEval – Multilingual corpus for offensive discourse
- HateBR – Brazilian Portuguese corpus (European Portuguese gap remains critical)

Models are evaluated using Precision (proportion of correct optimistic predictions), Recall (proportion of positive cases identified), and F1-score (harmonic mean of precision and recall). Contemporary models achieve F1 scores of 0.75-0.90 depending on task complexity and linguistic variation (Jahan & Al-Hasan, 2023).

Appendix B – Generative AI: Dual-Use Technology

Generative models enable the automated production of fabricated articles, fictitious testimonies, and highly persuasive political narratives tailored to user profiles (microtargeting). Tools such as Midjourney, Stable Diffusion, and advanced voice-cloning models generate hyper-realistic

synthetic content: portraits of non-existent events, false speeches with indistinguishable voices, and manipulated videos with high visual fidelity.

Simultaneously, specialized detection models identify visual artifacts, linguistic inconsistencies, and statistical patterns typical of synthetic content. These forensic tools become essential for investigating disinformation campaigns. However, the economic model of digital platforms—where algorithms prioritize emotionally intense content to increase engagement and advertising revenue—creates structural incentives for the circulation of disinformation.

Appendix C – Extended Case Analysis: Electoral Interference

Elections in the USA (2016, 2020), Brazil (2018, 2022), India, and several European countries have registered coordinated manipulation campaigns featuring manipulated videos, false statements attributed to candidates, coordinated hashtags to manipulate trends, and automated accounts for content amplification. These technical mechanisms integrate classic rhetorical strategies: scapegoating, radical emotional appeal, and intensive repetition—precisely the techniques systematically employed by Nazi propaganda (Welch, 2004; Thompson, 2017).

The convergence of historical rhetorical patterns with contemporary algorithmic amplification constitutes a qualitatively new threat to democratic processes, requiring equally innovative countermeasures that combine human oversight with AI detection capabilities.

Glossary

Annotation – Manual data labeling process for AI training (e.g., hate speech, irony, propaganda).

Bot – Automated account simulating human activity to amplify content or manipulate conversations.

Deepfake – Synthetically generated audiovisual content with a strong appearance of authenticity.

Embedding – Numerical representation of a word's meaning in a vector space.

F1-score – Metric combining precision and recall to evaluate model performance.

Microtargeting – Targeting political or commercial messages based on granular individual data.

Transformer – AI architecture capable of analyzing linguistic sequences with attention mechanisms.

XAI (Explainable AI) – Techniques making AI models transparent and interpretable, enabling decision auditing.

References

Bolsover, G., & Howard, P. N. (2017). Computational propaganda and political big data: Moving toward a more critical research agenda. Big Data, 5(4), 271–276. https://doi.org/10.1089/big.2017.29024.cpr

EUvsDisinfo. (2024). Disinformation trends in Portugal 2022-2024. European External Action Service. Retrieved from https://euvsdisinfo.eu

Howard, P., Lin, F., & Tuzov, V. (2023). Computational propaganda: Concepts, methods, and challenges. Communication and the Public, 8(3), 209–226. https://doi.org/10.1177/20570473231185996

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Landry, A. P. (2022). Dehumanization and mass violence: A study of mental state language in Nazi propaganda. PLoS ONE, 17(9), e0274957. https://doi.org/10.1371/journal.pone.0274957

Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). Algorithms, 15(8), 291. https://doi.org/10.3390/a15080291

Meta. (2021). Meta transparency report: COVID-19 misinformation. Meta Transparency Center. Retrieved from https://transparency.fb.com

OberCom. (2023). Desinformação e literacia mediática em Portugal: Relatório 2023. Observatório da Comunicação. Retrieved from https://obercom.pt

Observador. (2022). Fact check: Imagens de urnas manipuladas são falsas. Observador Fact Check. Retrieved from https://observador.pt

Polígrafo. (2021, 2024). Verificação de factos: Eleições legislativas e autárquicas. Polígrafo - Fact-Checking. Retrieved from https://poligrafo.sapo.pt

Steizinger, J. (2018). The Significance of Dehumanization: Nazi Ideology and Its Rhetoric. The Journal of Holocaust Research, 32(1), 1–17. https://doi.org/10.1080/21567689.2018.1425144

Strubytskyi, R., & Shakhovska, N. (2023). Method and models for sentiment analysis and hidden propaganda finding. Computers in Human Behavior Reports, 8, 100253. https://doi.org/10.1016/j.chbr.2023.100328

Thompson, G. (2017). Parallels in propaganda? A comparative historical analysis of the Islamic State and the Nazi Party. Journal of Public Relations Research, 29(1), 1–16. https://doi.org/10.1080/1062726X.2017.1281136

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics, 126(1), 351–373. https://doi.org/10.1007/s11192-020-03737-6

X Corp. (formerly Twitter). (2021). *Twitter Safety Report 2021: COVID-19 Misinformation*. X Transparency. Retrieved from https://transparency.x.com

Welch, D. (2004). Nazi Propaganda and the Volksgemeinschaft. Journal of Contemporary History, 39(2), 213–238. https://doi.org/10.1177/0022009404042129