LREC 2026

Information Disorder Workshop (InDor) @ LREC 2026

Workshop Proceedings

Editors
Simona Frenda, Marco Antonio Stranisci, Shaina Ashraf, Ada Ren, Ioannis Konstas, Usman Naseem

12 May 2026

ISBN 978-2-493814-87-6

Preface

Online disinformation poses a critical challenge to modern societies due to its influence on public opinion, behaviour, and democratic processes. The Natural Language Processing (NLP) community has increasingly addressed this issue through the development of resources and automated tools for detecting disinformation online. However, the lack of a shared theoretical and interdisciplinary framework leads to fragmented research approaches and leaves open challenges related to cultural and contextual interpretations of disinformation.

The Information Disorder (InDor) workshop aims to foster interdisciplinary and intersectoral dialogue to advance NLP research on disinformation. InDor starts from the Information Disorder framework (Wardle & Derakshan, 2017), which classifies false information into misinformation, disinformation, and malinformation and emphasises the importance of contextual and communicative factors (production, intentions, audience) in its spread. In this context, InDor seeks to establish a common and interdisciplinary theoretical foundation for disinformation research, examine cultural influences on susceptibility to disinformation, encourage interdisciplinary collaboration in dataset and model development, and explore the impact of real-world applications designed to counter disinformation.

This first edition of InDor features as its keynote speaker one of the creators of the Information Disorder framework: Professor Claire Wardle. Her talk *Revisiting the Information Disorder Framework: Reflecting on the use and relevance of definitions in our contemporary ecosystems* offers a reflection on this framework, created in 2017, analyzing the new challenges of the information process. This event wants to start an interdisciplinary discussion about disinformation providing remote and poster presentations of both archival and non-archival contributions and a round table with experts from NLP, phylosophy and communication. We will also award the positional paper titled *Combating Disinformation: Is There No Alternative?* as the InDor best paper. Authors propose to interpret information disorders not only as informational distortion but as a “syndrome with complex causes embedded in contexts of economic precarity, institutional distrust, and informational inequality” (Bassi et al., 2026).

InDor received 25 submissions, of which 19 were accepted: 6 non-archival and 13 archival. The fields covered range from linguistics, psychology, phylosophy, NLP, and machine learning. To celebrate the success of this first edition, we would like to thank all the members of the Program Committee who helped us promote the event and review the contributions, and thank Google DeepMind for sponsoring the InDor workshop!

Keynote Talk

Revisiting the Information Disorder Framework: Reflecting on the use and relevance of definitions in our contemporary ecosystems.

Claire Wardle
Cornell University
12 May 2026 - Room **5**

Abstract: In 2017, Claire Wardle and Hossein Derakshan wrote a report for the Council of Europe, entitled 'Information Disorder'. The frameworks laid out in the report has been used in different contexts globally, but in this talk, Wardle reflects on their usefulness. She considers whether they still have utility or whether they have simplified and limited the ways in which we understand the current challenges posed by contemporary information ecosystems.

Bio: Claire Wardle is an Associate Professor in the Department of Communication at Cornell University. She is considered a leader in the field of misinformation, verification and user generated content. In 2015, Claire co-founded the non-profit First Draft, a pioneer in innovation, research and practice in the field of misinformation. She went on to co-found the Information Futures Lab at Brown University's School of Public Health. Over the past decade she has developed an organization-wide training program for the BBC on eyewitness media, verification and misinformation, led social media policy at UNHCR, been a Fellow at the Shorenstein Center for Media, Politics and Public Policy at Harvard's Kennedy School, and been the Research Director at the Tow Center for Digital Journalism at Columbia University's Graduate School of Journalism. She has authored a number of articles and reports, including Information Disorder: An interdisciplinary Framework for Research and Policy for the Council of Europe and A Conceptual Analysis of the Overlaps and Differences between Hate Speech, Misinformation and Disinformation for the United Nation's Department of Peace Operations. She holds a Ph.D. in Communication from the University of Pennsylvania.

Organizing Committee

Workshop Organizers

- Simona Frenda, Heriot-Watt University
- Marco Antonio Stranisci, University of Turin
- Shaina Ashraf, University of Bonn
- Ada Ren, Macquarie University
- Ioannis Konstas, Heriot-Watt University
- Usman Naseem, Macquarie University

Program Committee

- Ada Ren, Macquarie University
- Aiqi Jiang, Heriot-Watt University
- Chiara Zanchi, University of Pavia
- Daniel Russo, University of Trento and Fondazione Bruno Kessler
- Erik Marino, Universidade de Évora
- Gavin Abercrombie, Heriot-Watt University
- Ioannis Konstas, Heriot-Watt University
- Ion Androutsopoulos, Athens University of Economics and Business
- Jonathan Tonglet, TU Darmstadt
- Juraj Vladika, Technical University of Munich
- Katarina Laken, Fondazione Bruno Kessler and Universidade de Santiago de Compostela
- Manuela Sanguinetti, University of Cagliari
- Mara Floris, Università Vita-Salute San Raffaele
- Mirko Lai, Università degli Studi del Piemonte Orientale
- Nancie Gunson, Heriot-Watt University
- Nikolaos Vitsakis, IT University of Copenhagen and Copenhagen University
- Paloma Piot, Universidade da Coruña
- Sara Gemelli, University of Pavia and University of Bergamo
- Shaina Ashraf, University of Bonn
- Simona Frenda, Heriot-Watt University
- Usman Naseem, Macquarie University
- Weronika Sieińska, Heriot-Watt University

Table of Contents

Workshop Program

Tuesday, May 12, 2026

14:00–14:10 **Introduction**

14:10–15:00 **Keynote speech: Claire Wardle**

15:00–15:15 **Best paper award**

Combating Disinformation: Is There No Alternative?
Davide Bassi, Søren Kirkegaard Fomsgaard, Erik Bran Marino and Katarina Laken

15:15–16:00 **Remote presentations**

Unraveling Deceptive Narratives: A Study of Conceptual Frameworks
Elie Alhajjar

High Accuracy, Low Generalization: Structural Homogeneity and Cross-Dataset Evaluation in Fake-News Benchmarks
Hiram Calvo and Mayte H. Laureano

Culturally Adaptive Explainable LLM Assessment for Multilingual Information Disorder: A Human-in-the-Loop Approach
Maziar Kianimoghadam Jouneghani

Benchmarking Check-Worthiness Models on LLM Generated Claims
Charlie George Roadhouse, Matthew Shardlow and Ashley Williams

Media Bias within Information Disorder: Bridging Two Research Communities through a Systematic Review
Francisco-Javier Rodrigo-Ginés and Jorge Chamorro-Padial

Towards an Intelligent Assistive System for Contextualized Ad-Hoc Information Verification and Discernment
Yunqian Bao and ChengXiang Zhai

Tuesday, May 12, 2026 (continued)

16:00–16:30 **Coffe break**

16:00–17:00 **Poster presentation**

Reliable News or Propagandist News? A Neurosymbolic Model Using Genre, Topic, and Persuasion Techniques to Improve Robustness in Classification
Géraud Faye, Benjamin Icard, Morgane Casanova, Guillaume Gadek, Guillaume Gravier, Wassila Ouerdane, Celine Hudelot, Sylvain Gatepaille and Paul Égré

Propaganda across Text Types in Russian Wartime Narratives
Anastasiia Vestel and Stefania Degaetano-Ortlieb

Grounding Information Disorder in NLP: A Theoretical and Operational Framework
Wajdi Zaghouani

Emotion and Information Disorder in NLP: A Systematic Mapping and Benchmark Blueprint
Renatha Vieira and Alvaro Figueira

Efficient and Explainable Hate Speech Detection through Distillation, Reasoning, and Agentic AI
Paloma Piot and Javier Parapar

Beyond Information Threats: Identifying the Psychosocial Indicators of Conspiracy-driven Victimisation
Sundara Kashyap Vadapalli, Mioara Cristea and Katerina Strani

A Multilingual Linguistic Analysis of Human vs LLM-Generated News in a Disinformation Context
Silvia Gargova, Alba Perez-Montero, Elena Lloret Pastor and Paloma Moreda Pozo

Disinformation between Knowledge and Ignorance. an Epistemological Comparison
Antonio Lizzadri

Population Replacement Conspiracy Theories Detection on Telegram and News Headlines: Benchmarking LLMs and BERT Models in Portuguese and Italian
Erik Bran Marino and Renata Vieira

Tuesday, May 12, 2026 (continued)

16:00–17:00 **Poster presentation**

A Cognitively-Grounded Bayesian Framework for Misinformation Susceptibility
Pranava Madhyastha

Mapping Discourse Reframing: A Multi-Layer Network Approach to Italian HPV Vaccine Discourse on X (2010-2024)
Lorella Viola

17:00–17:55 **Round table discussion**

17:55–18:00 **Closing remarks**

Combating Disinformation: Is There No Alternative?

Davide Bassi[1*], **Søren Kirkegaard Fomsgaard**[2*], **Erik Bran Marino**[3*], **Katarina Laken**[1,4]
[1]CiTIUS - Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[2]GREYC - University of Caen, Caen, France
[3]CIDEHUS - University of Évora, Évora, Portugal
[4]FBK - Fondazione Bruno Kessler, Trento, Italy
davide.bassi@usc.es, soren.fomsgaard@unicaen.fr, erik.marino@uevora.pt, alaken@fbk.eu

Abstract

This position paper critiques the dominance of detection-centered approaches in misinformation research. We argue that the prevailing paradigm treats information disorders as a content-level anomaly to be identified and suppressed, thereby obscuring the structural conditions under which different forms of information disorders emerge and resonate. Drawing on critical anthropology, we propose an alternative "clinical" model: information disorders should be understood not only as informational distortion, but as a syndrome with complex causes embedded in contexts of economic precarity, institutional distrust, and informational inequality. Treating detection as the ends rather than the means of intervention risks misguiding our efforts. Rather than positioning NLP primarily as a tool for boundary enforcement, we outline a reorientation toward structural diagnosis: diversifying data beyond WEIRD contexts, extracting socioeconomic and trust-related signals from discourse, and integrating computational outputs within interdisciplinary causal frameworks. Under this model, detection becomes a means for an epidemiology of discourse, subordinated to the broader objective of cultivating long-term epistemic resilience in our online environments.

Keywords: Disinformation, Natural Language Processing, Research Paradigm

1. NLP Research in the Age of Infodemics

Taken as a codification of medical ethics, the Hippocratic Oath famously frames the moral role of the physician in the practice of their craft. Until the turn of the 20th century, physicians under this oath treated their patients with a beneficent paternalistic attitude, trying to diagnose the causes of their ailments (Will, 2011). Much of the work in NLP based on what we will term the «Detection Paradigm» (cf. Section 2), takes a similar stance with respect to tackling information disorders (Wardle and Derakhshan, 2017). However, instead of focusing on diagnosing and treating causes, we argue that current work hyperfocuses on treating information disorders via detecting and eliminating, from a position of authority.

This paper argues for an uprooting and a reframing both of the paradigm of intervention in information disorders and of the role of the researcher within this paradigm. Instead of focusing exclusively on understanding mis- and disinformation in terms of content (and its production, dissemination and consumption[1]), with the aim of intervening on it through detection and moderation alone, we argue for an etiological re-orientation of the epistemics, practice of NLP research on information disorders.

One of the great epidemics of our time is arguably the epistemic crisis brought on by the combination of more or less regulated mass social media and the undermining of epistemic trust caused by the prevalence and political weaponization of synthetic content created using generative artificial intelligence. This so-called "infodemic"[2] is a multifaceted problem that must be tackled from multiple angles, for example as outlined in the five pillars of the High-Level Expert Group on Fake News and Online Disinformation (European Commission. Directorate General for Communications Networks, Content and Technology., 2018; European Parliament. Directorate General for Parliamentary Research Services., 2019). Our position is that a failure to correctly frame the nature of the problem of online mis- and disinformation as a syndrome of a broader disease, leads to a failure of treatment. Rather than focusing on providing palliative care, which is what relying on detection and moderation entails, we should pivot towards treating the root causes of the disease. Doing this requires understanding the underlying structural causes that lead to the production, dissemination and consumption lifecycle of information disorders. In this context, we argue that we should think of NLP researchers and practitioners as being akin to physicians or nurses. The role of researchers should in the first place be to diagnose and understand the disease based on the syndrome exhibited by the patient. The rest of the paper proceeds as follows. We first

[*]Equal contribution.
[1]cf. the C5 model (Kruijver et al., 2025).
[2]WHO.

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 1–15
12 May 2026.

characterize the Detection Paradigm in Section 2 and criticize examples of it when put into practice in Section 3. Section 4 looks at medical practice for a holistic understanding of disease treatment as a candidate model. Section 5 discusses sociological causes of disinformation. In Section 6, we turn our attention to pointing out possible avenues for reorienting future NLP research away from its focus on detection towards understanding the causes of information disorders, by drawing inspiration from clinical practices.

2. Defining the Detection Paradigm

What we refer to here as the Detection Paradigm (DP) is the predominant trend in NLP and computational social science research to focus on combating adverse social phenomena by *reducing* them to algorithmic detection problems. This ties in with a broader discussion of the 'taskification' of NLP, in which complex problems are reduced to tasks with clear benchmarks (Schlangen, 2021; Litschko et al., 2023). The Detection Paradigm is driven by reductionist epistemics, and it tends to favor solutionist interventions in practice.

Reductionist Epistemics. The DP is based on the tacit assumption that adverse complex phenomena such as misinformation can be captured by proxy of digital representations of language and media. The DP understands misinformation through a combination of two reductive epistemic components: propositionalism[3] and the reductive conceptual framework of the Conduit Metaphor (Reddy, 1979). The first reifies the problem of misinformation and construes it to be reducible to assertions about states of affairs, whose contents are verifiably false. Under this view, misinformation can be understood purely in terms of the truth value of its contents, without considering social, cultural, psychological or technological contexts within which information is produced, distributed and consumed. The second, originally inspired by information theory, frames communication as the unidirectional process of transmitting some propositional, informational content by a sender through a channel to a receiver. Miscommunication, and by extension, misinformation, is the result of imperfect transmission under this conceptual framework. Together, these two reductive components constrain both our understanding of the complex interrelated causes and effects of misinformation and our capacity to intervene on it.

[3]While propositionalism is relevant for information disorders such as mis- and disinformation, this aspect of DP does not necessarily extend to other harmful social phenomena typically modeled in NLP such as hate speech or propaganda.

Solutionist Interventionism. At the level of praxis, the DP is characterized by solutionist interventionism: the tendency to construe the social and political problem posed by misinformation as a technical problem that requires technical (engineering) solutions. Researching misinformation thus becomes the problem of accurately detecting it so that it can be eliminated, as opposed to, for instance, understanding what causes it or how and why it is spread.

The literature in NLP applied to adverse social online phenomena, broadly speaking, tends to reflect and favor the DP. For example, work has been done on detecting hate speech (Warner and Hirschberg, 2012; Djuric et al., 2015; Waseem and Hovy, 2016; Davidson et al., 2017), cyberbullying (Dinakar et al., 2011), trolling (Mihaylov and Nakov, 2016), offensive (Razavi et al., 2010) and abusive language (Nobata et al., 2016), propaganda (Amazon Alexa AI et al., 2021; Yoosuf and Yang, 2019), fake news (Ahmed et al., 2017; Ghosh and Shah, 2019; Pérez-Rosas et al., 2018), rumors (Derczynski and Zubiaga, 2020), conspiracy theories (Gupta, 2025), harassment (Bretschneider et al., 2014), political bias (Gangula et al., 2019), polarizing content (Graumas et al., 2019), to name some popular research topics of the last decade.

3. Critique of the Detection Paradigm

3.1. The Logic of the Instrument

The solutionist character of the DP, as defined above, is not merely a policy preference but a structural consequence of the available technical inventory. As Maslow's maxim states: «It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail» (Maslow, 1966). Classification models enforce a reduction of output into discrete decision boundaries by their very design. Content is ultimately classified as safe or unsafe, true or false, spam or legit, to trigger automated removal or labeling. This reconfiguration is based on an epistemic reduction: the efficacy of the system relies on the stability of truth values, a condition that holds for only a subset of communicative acts. It is necessary to distinguish between two modes of epistemic discourse, meaning the way we frame knowledge through discourse:

Mode A (Empirical Verification): Discourse regarding verifiable facts, such as election dates, physical events, or geospatial data. These claims possess a binary truth value verifiable against a shared objective reality.

Mode B (Political Hermeneutics): Discourse regarding values, legal interpretations, historical narratives, or future predictions. These are subjective, normative, and dependent on the observer's

political framework.

The detection paradigm applies operational logic designed for Mode A to discourse belonging to Mode B. This constitutes a «category mistake» that attempts to purify political hybridity into binary facts (Latour, 2012). Automated Fact-Checking (AFC) systems exemplify this category mistake by treating interpretive discourse (Mode B) as empirically verifiable (Mode A). Because AFC architectures demand single veracity labels, they force underspecified claims, such as 'It is illegal in Illinois to record a conversation', into reductive binary classifications (e.g., True/False), completely ignoring unstated contextual variables (like whether it is done secretly or not) (Glockner et al., 2024). Furthermore, this strict true/false dichotomy proves epistemically inadequate even for scientific information. Given the inherently probabilistic and evolving nature of scientific knowledge, assessing gradients of accuracy is far more meaningful than enforcing absolute binary classifications. And, as La Barbera et al. (2025) argue, these binary labels do more than oversimplify reality; they act as authoritative judgments that delegitimize valid alternative interpretations and justify unwarranted interventions.

Crucially, this algorithmic flaw is the crystallization of human epistemic confusion. For instance, following U.S. House Resolution 894, which moved to equate anti-Zionism with antisemitism, Meta updated its technical moderation policy on the term «Zionist» (U.S. House of Representatives, 2023; Meta Transparency Center, 2024) to reflect this shift in political (mode B) discourse. When such political definitions are operationalized into moderation systems the infrastructure of safety automates the suppression of counter-hegemonic discourse (Gramsci, 1929/2020), censoring political debate under the designation of «misinformation correction».

The persistence of this category mistake serves a functional role in the preservation of the status quo. The exclusive focus on content moderation acts as a palliative measure, lowering the «temperature» of the public arena by suppressing the symptoms (e.g., toxic polarization and informational noise) without addressing the underlying infection.

3.2. The Detection Paradigm in Practice

The pervasiveness of such an approach is also evident from the institutional responses to the problem. Considering the European context, while the EU recognizes how the information crisis intersect with other crises, identifying it as one among multiple elements of the polycrises we are facing (F et al., 2025), when it comes to defining the actions to be put in place to address it, adopts a purely information based approach. This is particularly evident when considering the "EU Code of Conduct on Misinformation" (European Commission, 2025). This document outlines 6 macro strategies through which we should tackle misinformation:

- *Demonetization of disinformation*: cutting financial incentives by preventing ads and monetization next to disinformation content.
- *Political advertising transparency*: making political and issue ads clearly identifiable, traceable, and publicly accessible.
- *Integrity of services*: reducing manipulative behaviors (bots, fake accounts, coordinated inauthentic behavior, deepfakes).
- *User empowerment*: enable users to recognize, contextualize, and challenge disinformation (with labels and digital tools or fostering media literacy).
- *Empowering researchers & fact-checkers*: providing such roles with data access and structural support for independent scrutiny
- *Monitoring, transparency & accountability*: ensuring measurable commitments, reporting, audits, and public oversight of implementation.

Although articulated as distinct interventions, these measures share a conceptualization of information disorders as a detectable anomaly to be contained and eradicated from the information environment. In this sense, they constitute variations of the same logic, converging on the same instrumental objective: the detection and suppression of problematic content. Even media literacy initiatives aimed at empowering users can be understood as operating within this logic: instead of automating detection externally, they internalize it within users, reconfiguring them as self-monitoring nodes of the information environment. However, once examined beyond its immediate operational aims, the apparent coherence of this approach reveals a set of critical repercussions that unfold at multiple and interrelated levels.

Effectiveness: given the low base rate of misinformation consumption in the "Global North" (approximately 0.15–6% of total media consumption), intervention focused on its detection and suppression are bound to yield only marginal improvements in the general quality of the informational environment (Acerbi et al., 2022); thereby exposing a structural limit of detection-centered approaches.

Feasibility: detection-based governance presupposes sustained investment by private platforms in trust and safety infrastructures (e.g. technical systems, human moderators, and audits). Despite regulatory initiatives such as the EU Digital Services Act (DSA), which formally mandates systemic risk mitigation and transparency obligations,

recent developments showed how the viability of such paradigm rests on corporate incentives, which, in turn, are neither stable nor inherently aligned with long-term public epistemic resilience. In 2023, Meta laid off approximately 21,000 employees, with reported consequences for its Trust and Safety operations (Field and Vanian, 2023). Similarly, following Elon Musk's takeover, X dismissed nearly 80% of engineers dedicated to trust and safety functions (Brewster, 2023). In the same period, Google reduced the staff of *Jigsaw*, its internal unit focused on countering online toxicity and hate speech by roughly one third (Field and Vanian, 2023). These developments reveal a structural fragility: the effectiveness of detection-centered interventions depends on the continuous allocation of resources by profit-driven corporations. When economic priorities shift, the operational capacity of the paradigm contracts accordingly, regardless of formal regulatory frameworks. The scenario is then further complicated by initiatives like the portal `freedom.gov` hosted by the US government, intended to allow citizens in Europe and elsewhere to access material that European laws have sought to restrict — including alleged hate speech and content linked to terrorism — by bypassing local content restrictions. Although framed by U.S. officials as an effort to promote digital freedom, critics argue that the portal could undercut the regulatory aims of the Digital Services Act and analogous frameworks, and that it embodies a fundamentally different normative stance toward harmful content (Down, 2026).

Public Legitimacy: as shown by Rich et al. (2020), the support for social media fact checking is highly uneven and strongly structured by partisanship. While most of the people endorse fact-checking "in the abstract", support declines when interventions target politically salient actors or in-group figures, with fact-checks frequently perceived as partisan or punitive rather than corrective. As a result, fact-checking is effective mainly among those already inclined to trust the institutions performing it, thereby limiting its broader legitimacy in polarized information environments.

Explanatory scope: detection-centered approaches systematically under-specify the determinants of mis- and disinformation by privileging content-level properties over both individual and structural drivers. At the individual level, they largely ignore the role of attitudinal factors—such as partisanship, identity management, and motivated reasoning—that shape how users actively select, interpret, and engage with information (Altay et al., 2023; Robertson et al., 2023; Bakshy et al., 2015). At the structural level, they overlook the broader socio-economic and institutional conditions under which misinformative and conspiratorial narratives flourish, including corruption, economic insecurity, inequality, and deficits of institutional trust (Hornsey and Pearson, 2022; Adamus et al., 2024; Amazeen et al., 2024). By doing so, the DP obscures how engagement with misinformation functions as a response to unresolved societal needs, such as equal access to information, political representation and agency, institutional credibility and shared spaces for negotiating identity and social belonging (we elaborate more on this in Sec. 5). In this sense, the detection paradigm exemplifies what Morozov (2013) terms *technological solutionism*: the displacement of political and structural questions by ostensibly neutral technical fixes.

A physician that treats the symptom itself as the disease, however, cannot prescribe an adequate cure.

4. The Ideal Clinical Model

It thus becomes evident that the DP fosters a reductive view of mis- and disinformation, that remains blind to the underlying societal causes of the issue, by modeling it as a problem of the verification of context-free propositions. Staying with our medical metaphor, our critique for this "blind obsession with the symptom", (i.e. misinformation) closely parallels long-standing critiques advanced by critical medical anthropology. In that context, the reduction of illness to isolated biological dysfunctions, while neglecting their sociological, contextual, and phenomenological dimensions, has been shown to result in the medicalization of fundamentally social problems (Singer, 2009; Farmer, 2004).

An 'ideal clinical model' would instead treat misinformation as a diagnostic signal, not as a pathological object to be excised from the information environment. Just as critical medical approaches situate illness within broader configurations of social inequality, institutional power and lived experience, a structural approach to misinformation must situate communicative practices within the political–economic conditions that make certain narratives resonate. Engagement with misinformative content should therefore be analyzed as an action occurring within an arena structured by historically sedimented distrust, economic precarity, identity fragmentation, and transformations in the architecture of the public sphere. In this sense, mis- and disinformation become intelligible as relational phenomena emerging at the intersection between individual and group agencies and structural constraints. They are not reducible to epistemic deficit; rather, they reflect the dynamic interplay between the social construction of interpretative categories and the material conditions that shape everyday life. From this perspective, the task is not "merely" to correct claims, but to understand the processes through which meanings are produced,

circulated, and appropriated. Finally, reframing mis- and disinformation as a capability-depriving condition embedded in the political economy of the public sphere, the most effective intervention is the one that aims to restore the epistemic and social capabilities required for meaningful participation in collective sense-making, rather than solely at suppressing inaccurate content (Nussbaum, 2011).

5. Diagnosing the Sociological Causes of Disinformation

To operationalize the 'ideal clinical model' described in Sec. 4, this section aims at outlining (some of) the structural and contextual conditions under which misinformative narratives emerge, circulate, and acquire relevance. Disinformation is here understood as embedded in transformations of the public sphere (Jungherr and Schroeder, 2021): susceptibility and engagement are systematically patterned by social position and institutional credibility rather than by simple deficits in reasoning (Altay and Mercier, 2026; Schirmer et al., 2025).

Unequal Access to High-Quality Information: as professional journalism increasingly relies on subscription-based models, high-quality news becomes partially excludable, while sensationalist or low-quality content remains usually freely accessible. In high-income countries, only 17% of individuals pay for online news and subscribers are disproportionately affluent and highly educated (Newman et al., 2024). Informational quality thus becomes stratified along socioeconomic lines. Crucially, much information disorder research presupposes equal access to credible information, thereby overlooking how structural inequalities condition exposure opportunities in the first place (Schirmer et al., 2025). When everyday life is organized around financial instability and short-term survival concerns, cognitive resources are disproportionately allocated to immediate problem-solving, leaving diminished capacity for effortful tasks such as cross-checking sources or evaluating evidentiary quality (De Bruijn and Antonides, 2022; Burlacu et al., 2022). This makes information verification not simply a matter of motivation or competence, but a resource-intensive practice. Under these conditions, vulnerability to low-quality information reflects infrastructural asymmetry and structurally constrained cognitive bandwidth, rather than merely individual epistemic failure (Ronzani, 2025).

Structural Transformations and Gatekeeping: The digital public arena has eroded the influence of traditional gatekeepers, such as journalists and political parties, who previously filtered information flows (Jungherr and Schroeder, 2021). While this has lowered barriers to entry for marginalized topics and perspectives previously neglected by mainstream media, it has also produced an environment in which information visibility depends less on professional verification and more on algorithmic amplification and user engagement (Marino et al., 2024). Most major social media platforms operate on engagement-based business models in which clicks, shares and watch time drive advertising revenue (Napoli, 2019): where early internet content discovery was largely user-led, algorithmic recommendation feeds shifted this paradigm toward system-driven distribution, in which emotionally arousing and polarizing content is systematically amplified regardless of its truth value (Vosoughi et al., 2018). This represents an externalization of epistemic responsibility: the burden of filtering, once held by institutional gatekeepers, has been reallocated as a substantial cognitive load onto the individual. Such pressure is reflected in the finding that 39% of the global population actively avoids the news not because they consider it false, but because they find it overwhelming and emotionally draining (Newman et al., 2024). The resulting vulnerability is not primarily to the persuasion of false claims, but to the fragmentation of the shared frames through which collective reality is constructed.

Economic Uncertainty and Loss of Agency: socioeconomic dynamics can also shift how people make sense of information by undermining perceived control and institutional legitimacy. Higher levels of objective and perceived economic inequality increase openness to alternative explanatory frames and endorsement of conspiracy beliefs (Altay and Mercier, 2026). This effect is fully mediated by *anomie* (i.e. the perception that society and its institutions are breaking down) thereby motivating a search for meaning and control that conspiratorial narratives promise to restore (Casara et al., 2022). In this view, conspiracy beliefs are not merely incorrect beliefs, but compensatory explanations that become attractive under existential threat: when perceived agency is low, conspiratorial accounts offer coherent causal stories, identifiable culprits, and a sense of epistemic and moral leverage over events (Douglas et al., 2019). Crucially, this also means that counter-disinformation strategies focused exclusively on fact-checking fail to address the demand for alternative narratives, which frequently stems from a grain of truth generated by material inequality — what Benford and Snow (2000) call *frame resonance*: a narrative succeeds not simply because it is available, but because it resonates with the lived experience of those who feel abandoned by institutions (Waisbord, 2018).

Historical Roots of Systemic Distrust: distrust in official information is also historically structured. Communities that have experienced repeated neglect, discrimination, or harm at the hands of trusted

authorities develop durable patterns of distrust that constitute rational responses to structural conditions (Smith and Freyd, 2014). Cox (2024) finds that a majority of Black Americans believe that societal institutions are designed to disadvantage them, a perception that can reduce reliance on government-backed health or climate messaging (Jaiswal et al., 2020). Such attitudes are grounded in documented histories of discrimination and institutional betrayal (particularly in medical and scientific domains, epitomized by the legacy of the Tuskegee Syphilis Study) which have long shaped patterns of epistemic mistrust (Bogart et al., 2021). From this perspective, distrust operates as an adaptive response to historically sedimented expectations shaped by repeated experiences of exclusion and harm.

Corruption and Institutional Credibility: institutional distrust is not only historically sedimented but is continuously reproduced by contemporary governance quality. In contexts where corruption is salient, conspiratorial explanations become structurally less implausible. Cordonier and Cafiero (2024) shows that higher levels of public sector corruption are associated with stronger endorsement of conspiracy beliefs, even when controlling for democracy, press freedom, inequality, and human development. Relatedly, populist attitudes - often rooted in anti-elite and anti-institutional rhetoric - predict reduced trust in political and scientific institutions and increased reliance on alternative media, thereby indirectly fostering lower compliance with expert-backed guidance (Ehrke et al., 2023).

Identity and Social Signaling: online platforms should not be conceptualized merely as information distribution systems, but as arenas for identity performance. Bail (2022) argues that social media environments encourage users to publicly signal group belonging, moral commitments, and political identities, rewarding expressions that are distinctive, emotionally charged, or normatively aligned with in-group expectations. In this context, sharing or affirming contested claims can function less as a statement of epistemic conviction than as an act of social positioning. As Mercier (2020) and Bassi et al. (2025) contend, individuals often endorse or publicly defend ideas - including implausible ones - not because they are epistemically persuaded, but because doing so reinforces alliances and distinguishes adversaries. At its extreme, this dynamic converges with the phenomenon of *conspirituality* (Ward and Voas, 2011), in which the search for alternative spiritual truths merges with the quest for hidden political ones, forming self-reinforcing epistemic communities organized around identity rather than evidence. Empirical data confirms the depth of this motivated selective exposure: individuals with highly unfavorable views on vaccines are nine times more likely to visit vaccine-skeptical websites than their favorable peers; during the 2016 US election, Trump supporters were twice as likely to visit unreliable news sites compared to Clinton supporters (Altay and Mercier, 2026). For these users, the problem is not a lack of data but a lack of institutional trust, and misinformation circulates in part because it serves relational and identity-building functions that accurate information does not.

5.1. Two Epistemic Positions: The Uninformed and the Disinformed

The structural conditions described above do not produce a single, homogeneous 'misinformed public'. Following a classical methodological distinction between ideal-typical categories (Weber, 1949), we argue that they instead generate two structurally distinct epistemic positions that are too often conflated in the literature.

The *uninformed* are those passively excluded from accurate information by structural barriers (e.g. access costs, digital skill deficits, cognitive bandwidth constraints) rather than by any motivated rejection of it. Their relationship to misinformation is one of exposure without resistance, driven by infrastructural asymmetry rather than epistemic choice.

The *disinformed*, by contrast, occupy an active position: they encounter accurate information but reject it, because acceptance would threaten group identity, undermine social alliances, or destabilize the compensatory narrative frameworks that give meaning to experiences of institutional betrayal and loss of agency.

This distinction carries direct implications for intervention design. Strategies premised on correcting an information deficit (e.g. fact-checking, content labeling, source literacy training) are well-suited to the uninformed, but structurally misaligned with the disinformed, for whom the problem is not epistemic but relational and political. Treating the disinformed as merely uninformed ignores the profound moral and social dimensions of resistance to institutional truth claims (Morozov, 2013). Effective intervention must instead engage with the structural conditions we outlined that make conspiratorial and misinformative frameworks not just available, but meaningful.

6. Reorienting NLP Research

If misinformation is not merely anomalous content to be removed, but a symptom of deeper structural transformations, research efforts must shift toward modeling the contextual and systemic conditions under which such narratives emerge and acquire meaning. While the macro-level determinants discussed in Section 5 have traditionally been investi-

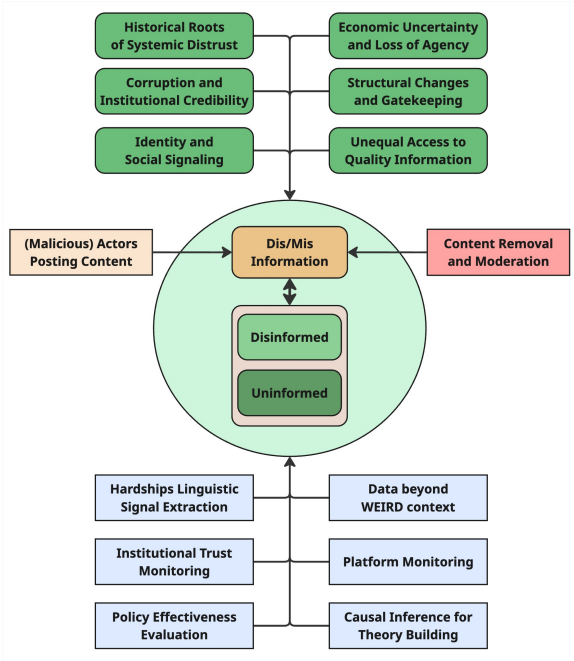


Figure 1: Structural model of mis- and disinformation. Dark green boxes represent structural causes shaping the socio-epistemic context (light green) in which disinformation (orange) is consumed by two publics: the *disinformed* and the *uninformed*. Blue boxes indicate proposed NLP contributions to tackle the issue.

gated within sociology, political science, and psychology, this raises a distinct question for computational research: what is the specific contribution that NLP can provide within a structurally informed framework? We contend that this contribution lies in moving beyond content-level classification toward the large-scale operationalization of structural variables. In the following, we outline concrete research directions through which the NLP community can support such a reorientation.

Data and Sampling: a central limitation of the current literature concerns its sampling bias. More than 80% of misinformation research is conducted in high-income Western democracies and typically relies on young, educated, urban, and digitally literate participants (Badrinathan et al., 2025; Blair et al., 2024). As emphasized by Ronzani (2025), this narrow empirical base risks generating globally generalized conclusions from demographically restricted populations, thereby obscuring the role of structural variance in misinformation dynamics. This issue is not merely geographical but socioeconomic. Schirmer et al. (2025) highlights how digital exclusion, economic precarity, language barriers, and historically rooted institutional mistrust shape differential exposure to and engagement with misinformation. Yet such dimensions remain underrepresented in large-scale computational studies. The main challenge in reorienting NLP research beyond Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations lies on the fact that in many contexts, misinformation circulates not primarily through public social media feeds but via encrypted messaging platforms (e.g., WhatsApp, Telegram), local radio broadcasts, SMS chains, community gatherings, and religious networks (Badrinathan et al., 2025; Blair et al., 2024). However, recent methodological innovations provide viable pathways for accessing these environments responsibly, including chatbot-based recruitment strategies (Bowles et al., 2020) and structured WhatsApp data-donation infrastructures (Garimella and Chauchard, 2025).

Sociological Signal Extraction from Discourse: NLP can contribute by operationalizing *structural variables as first-class prediction targets*, i.e., by devising models that detect when discourse expresses specific sociological dimensions. Rather than detecting misinformation for removal, models can be trained to identify explicit and implicit hardship narratives such as inability to afford basic goods, instability of work and housing, and perceived unfairness of the economic system. Operationally, this implies building datasets and taxonomies for categories like material deprivation, economic insecurity, downward mobility, and resource-strain complaints, enabling supervised or weakly supervised detection at scale. For instance, hardship detection targets discourse reporting scarcity and struggle; identity-threat detection can target rhetorical strategies used to signal belonging and mark disaffiliation (Bassi et al., 2025); and political alienation can be operationalized via stigmatizing and exclusionary language detection (Sap et al., 2020; Straton et al., 2020; Bassi et al., 2024). This approach aligns with (i) text-as-data methods that treat language as measurement of latent social conditions (Puklavec et al., 2023), and (ii) recent domain-specific efforts to formalize poverty-related social phenomena in NLP, including datasets and taxonomies targeting poverty-related bias and discourse about poverty (Curto et al., 2025). The resulting signals can be linked to regional socioeconomic indicators (e.g., unemployment, inequality), educational access, or infrastructure metrics, enabling multi-level modeling of how material constraints shape informational engagement and vulnerability pathways. Although modeling structural vulnerability in NLP entails non-trivial ethical risks related to surveillance and stigmatization (cf. Section 6.1), these must be weighed against the long-term benefits of making such constraints publicly legible. From a capabilities-based perspective (Nussbaum, 2011), rendering these constraints visible informs interventions, largely outside NLP itself, aimed at improving the struc-

tural environments in which epistemic vulnerability arises, thereby promoting full human functioning and meaningful participation in collective sense-making.

Institutional Trust Monitoring Systems: A complementary direction concerns longitudinal monitoring of institutional legitimacy via construct-level detection of trust-related social phenomena. Specifically, NLP systems can be designed to detect specific delegitimization and distrust frames (e.g., corruption, betrayal, capture, manipulation), as well as experiences of institutional abandonment and credibility rupture (Djouvas et al., 2023; Lattmann, 2025). In this sense, NLP tools become a diagnostic dashboard for epistemic infrastructure: not just tracking isolated claims, but monitoring structural erosion in legitimacy and belonging across time, communities, and platforms. NLP systems can be designed to track fluctuations in trust-related discourse across time, communities, and platforms.

Platform and Amplification Auditing NLP and computational methods can also contribute by modeling how platform architectures shape exposure and amplification. Rather than presenting the detection of misinformative content using its semantic properties as a solution to the epistemic crisis, research can investigate how engagement-based ranking systems, recommender algorithms, and network structures condition visibility, diffusion velocity, and cross-group exposure. Combining discourse analysis with network and ranking data would enable systematic auditing of amplification dynamics, identifying structural asymmetries in content propagation independent of truth value. Such approaches shift attention from "what is false" to "what becomes visible and why."

Policy and Intervention Evaluation: NLP-based measurement systems can support the evaluation of structural interventions (Nannini et al., 2025). Computational indicators of institutional trust, polarization dynamics, and exposure asymmetries can be tracked before and after regulatory changes (e.g., platform transparency mandates, media subsidy policies, or moderation reforms), enabling systematic assessment of their downstream epistemic effects. Rather than relying on platform self-reporting or short-term engagement metrics, such infrastructures would allow independent monitoring of how policy decisions reshape informational ecosystems over time; configuring NLP as a component of democratic accountability.

From Measurement to Causal Inference: finally, the value of these tools lies not in their descriptive power alone but in their integration within interdisciplinary causal frameworks. Computational outputs—structural discourse indicators, trust trajectories, network diffusion patterns—can be combined with socioeconomic data, policy interventions, and institutional metrics to support causal inference. Under this reorientation, detection would be subordinated to a broader epistemic objective: understanding the structural conditions under which misinformation emerges, circulates, and acquires meaning.

6.1. Methodological Considerations

To fully realize the clinical model proposed above, it is crucial to recognize the epistemological limits of text-based computational methods. If the root causes of information disorders are structural, then NLP cannot serve as the sole instrument of detection and intervention. Instead, NLP should be conceptualized as a complementary diagnostic tool. It provides a scalable layer for measuring discursive symptoms, but the actual interventions, aimed at restoring epistemic resilience and addressing socioeconomic marginalization, must inherently take place in other domains, such as public policy, institutional reform, and civic education.

Given the possible avenues for reorienting NLP outlined above, we now discuss some general methodological considerations to be kept in mind to correctly operationalize this shift:

Integrating Textual Signals with Socio-Institutional Data NLP-derived indicators acquire diagnostic value only when anchored to external data. Aforementioned discourse signals about institutional distrust, hardship narratives, identity-threat framing should be linked to regional socioeconomic indicators (e.g., unemployment, Gini coefficients, poverty indices) and survey-based trust measures (e.g., Eurobarometer, World Values Survey) to identify whether structural conditions predict discursive vulnerability patterns at scale (Khoun et al., 2025). Bor et al. (2026), for instance, combining cross-national survey data on online political hostility with macro-level democracy and inequality indicators across 30 countries, demonstrates that discourse patterns are structured by societal conditions rather than platform affordances alone.

Hybridizing Detection through Mixed-Methods First, the use of NLP for detection should be reoriented from a standalone classification task into a component of a hybrid, mixed-methods research architecture. Because the core determinants of misinformation lie largely outside the text, computational analysis is intended to be triangulated with external methodologies. This includes, but is not limited to: integrating NLP pipelines with digital ethnography to understand the localized cultural context of online communities (Kozinets and Gretzel, 2024); utilizing sociolinguistic models to distin-

guish expressions of material deprivation (Muneton-Santa et al., 2022); combining semantic classification with platform metadata to audit algorithmic amplification (Nannini et al., 2025); and leveraging LLM-driven swarm intelligence engines to simulate the sociological spread of these narratives *in silico* (i.e. *MiroFish*). In this hybrid model, automated detection is no longer the terminus of the research pipeline, but rather a sensor that informs broader qualitative, quantitative, and computational social investigations.

Reliability and Riskiness of Extracting Social Variables All measuring instruments have error margins, and NLP tools are no different. Modern LLMs are often overestimated in terms of how reliable their outputs are, because they produce very fluent and natural text. However, LLMs often produce factually incorrect responses (Ji et al., 2023) and poorly deal with social context (Hovy and Yang, 2021). This becomes especially salient in cross-cultural contexts (Moayeri et al., 2024), or when dealing with data from non-standard language varieties (often used by minoritized demographics) (Sap et al., 2019; Nguyen et al., 2025).

A large part of NLP research is currently focused on improving model performances on pre-established benchmarks (Schlangen, 2021). But when actually using NLP tools, it is at least equally critical to understand the error margins, both quantitatively and qualitatively. Error patterns are not random, they are the result of systemic biases in the training data and inherent ambiguities of natural language in use. One of the main tasks of NLP practitioners should be to help scholars and policy makers from other fields understand how to responsibly interpret results of NLP tools.

Extracting social variables from text also poses ethical risks in terms of profiling. We do not advocate for the development of systems to profile individuals for social class, wealth, or similar factors. However, we believe that traditional sociological research can benefit from NLP research by revealing aggregate patterns of, among others, social stratification, linguistic variation, and cultural capital.

7. Conclusions

This paper has argued that the prevailing *Detection Paradigm* configures misinformation as a pathological object to be excised from the informational body. Drawing on critical anthropology, we suggested that this mirrors earlier reductions of illness to isolated biological dysfunctions: symptoms are managed, while the social conditions that generate vulnerability remain intact. In clinical contexts, such reductionism has long been criticized for depoliticizing suffering, transforming historically produced inequalities into technical problems of correction. Analogously, when misinformation is framed primarily as a content-level anomaly, structural determinants such as economic precarity, institutional distrust and identity fragmentation are displaced from the analytic foreground. We do not deny that misinformation can exert causal effects (Tay et al., 2024). Yet recognizing multicausality does not imply symmetry of intervention. Even if misinformation contributes to downstream harms, treating detection as the primary endpoint risks confusing proximal triggers with generative conditions. In an 'ideal clinical model,' symptoms are diagnostically meaningful signals of deeper configurations of power, exclusion, and legitimacy. Hence, detection should function as an instrument of structural diagnosis: an '*epidemiology of discourse*', rather than as a self-sufficient cure. This reorientation carries institutional implications. Detection-centered governance presupposes sustained alignment between public epistemic goods and platform incentives, a fragile and politically contingent condition. The proposed approach, instead, seeks to equip public institutions with computational tools capable of identifying patterns of distrust, marginalization, and socio-economic stress that shape the resonance of misinformative narratives. Under this view, NLP shifts from boundary enforcement to infrastructural analysis, contributing to a diagnostic framework that reconnects epistemic disorder to its political–economic substrates.

Concluding, we emphasize how our contribution is intentionally programmatic. The determinants and research directions outlined are illustrative rather than comprehensive, and detection can remain operationally valuable in circumscribed contexts. However, if misinformation is understood not only as distortion but as expression, its analysis must be repoliticized: embedding it within the structural transformations of the public sphere that condition both belief formation and institutional legitimacy. Only within such a framework can mitigation move beyond symptom management toward the cultivation of long-term epistemic resilience.

8. Ethical Considerations

Current NLP Ethics is largely informed by principalist and applied-ethics frameworks, most notably, Value Sensitive Design (Friedman and Kahn, 2002; Friedman et al., 2013). Attention has also been paid to applying explicitly deontological (Prabhumoye et al., 2021) or consequentialist (Card and Smith, 2020) analyses to NLP practices, but even so, ethical considerations in the NLP community have primarily been oriented towards design and deployment practices of researchers and developers (cf. the ACM Code of Ethics and IEEE Global Initiative

on Ethics of Autonomous and Intelligent Systems (Chatila and Havens, 2019)). Assessment of social impact (Hovy and Spruit, 2016), such as harm, risk, and algorithmic fairness, is crucial, but if we are concerned with understanding the mechanisms that drive the expressions of information disorders, we need to re-conceptualize the role of NLP researchers not just as builders of potentially dangerous tools, or as moderators of undesirable content, but as curious caretakers who play a role in improving the structural conditions that shape our information environments for the better. If we instead think of the role of NLP as facilitating the diagnoses of diseases and disorders, it would be appropriate to move away from the beneficent paternalism implied by the Detection Paradigm, towards an attitude that favors autonomy of and care for the populations that we model. Biomedical ethics changed from framing the doctor-patient relationship as one marked by the well-meaning authority of the doctor over their patient to one increasingly characterized by the respect for the autonomy of and interpersonal care for the patient (Will, 2011), largely informed by the Ethics of Care (Gilligan, 1984; Noddings, 1984). NLP research should adopt a comparable attitude towards the individuals and populations we model with our data and tools. Discussions of social harms (Hovy and Spruit, 2016) and ethics-by-design (Leidner and Plachouras, 2017) pushed the ethics of NLP in necessary directions. We believe the Ethics of Care could offer a similar path forward for improving our ethical framework of NLP, just as it has been suggested for AI Ethics for decision making (Villegas-Galaviz and Martin, 2023). Precisely how to faithfully support the autonomy and flourishing of the participants and populations modeled in NLP research applied to phenomena like misinformation remains for now an open question for future research.

Limitations

This paper is intentionally programmatic and therefore subject to several limitations. The structural conditions outlined in Section 5 and the NLP research directions proposed in Section 6 are illustrative rather than exhaustive: both represent theoretically grounded selections, not comprehensive inventories.

Additionally, although the paper critiques the WEIRD bias of existing misinformation research, its own argumentation draws predominantly on evidence and policy examples from European and North American contexts; the structural factors identified may differ substantially in salience and configuration elsewhere.

At the methodological level, the considerations raised in Section 6.1 apply with particular force to the sociological signal extraction direction. NLP tools operate with non-trivial error margins and underperform systematically on non-standard language varieties and cross-cultural contexts. Indicators of institutional distrust, hardship, or identity threat should therefore be treated as noisy proxies requiring external validation before informing policy.

Finally, extracting socioeconomic signals from discourse at scale carries substantive ethical risks. Even when intended for aggregate analysis, such systems can be repurposed for individual profiling or surveillance of vulnerable populations. The boundary between aggregate monitoring and individual-level stigmatization is technically fragile and institutionally contingent — a risk that any operationalization of the proposed framework must explicitly address.

Acknowledgments

This work was supported by the HYBRIDS project, which has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee (Grant Number: EP/X036758/1).

References

Alberto Acerbi, Sacha Altay, and Hugo Mercier. 2022. Research note: Fighting misinformation or fighting for information? *Harvard Misinformation Review*.

Magdalena Adamus, Eva Ballová Mikušková, Pavol Kačmár, Martin Guzi, Matuš Adamkovič, Maria Chayinska, and Jais Adam-Troian. 2024. The mediating effect of institutional trust in the relationship between precarity and conspiracy beliefs: A conceptual replication of adam-troian et al.(2023). *British Journal of Social Psychology*, 63(3):1207–1225.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138, Cham. Springer International Publishing. TLDR: A fake news detection model that use n-gram analysis and machine learning techniques is proposed, which investigates and compares two different features extraction techniques and six different machine classification techniques.

Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. 2023. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, 4(4):1–34.

Sacha Altay and Hugo Mercier. 2026. Misinformation is a symptom: Commentary on Ecker et al. (2025). *American Psychologist*, 81(2):287–289.

Michelle A Amazeen, Rosalynn A Vasquez, Arunima Krishna, Yi Grace Ji, Chao Chris Su, and James J Cummings. 2024. Missing voices: examining how misinformation-susceptible individuals from underrepresented communities engage, perceive, and combat science misinformation. *Science communication*, 46(1):3–35.

Amazon Alexa AI, Seunghak Yu, Giovanni Da San Martino, Department of Mathematics, University of Padova, Italy, Mitra Mohtarami, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA, James Glass, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA, Preslav Nakov, and Qatar Computing Research Institute, HBKU, Qatar. 2021. Interpretable Propaganda Detection in News Articles. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*, pages 1597–1605. INCOMA Ltd. Shoumen, BULGARIA.

Sumitra Badrinathan, Simon Chauchard, and Niloufer Siddiqui. 2025. Misinformation and support for vigilantism: An experiment in india and pakistan. *American Political Science Review*, 119(2):947–965.

Chris Bail. 2022. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Davide Bassi, Giovanni Da San Martino, Renata Vieira, and Martín Pereira-Fariña. 2025. Drawing digital lines: pattern analysis of divisive rhetoric in social network discussions. *Humanities and Social Sciences Communications*, 12(1):2009.

Davide Bassi, Luisa Orrù, Christian Moro, Davide Salvarani, and Gian Piero Turchi. 2024. Investigating avhs narratives through text analysis: the proposal of dialogic science for tackling stigmatization. *BMC psychology*, 12(1):434.

Robert D Benford and David A Snow. 2000. Framing processes and social movements: An overview and assessment. *Annual review of sociology*, 26(2000):611–639.

Robert A Blair, Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J Stainfield. 2024. Interventions to counter misinformation: Lessons from the global north and applications to the global south. *Current Opinion in Psychology*, 55:101732.

Laura M Bogart, Bisola O Ojikutu, Keshav Tyagi, David J Klein, Matt G Mutchler, Lu Dong, Sean J Lawrence, Damone R Thomas, and Sarah Kellman. 2021. Covid-19 related medical mistrust, health impacts, and potential vaccine hesitancy among black americans living with hiv. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 86(2):200–207.

Alexander Bor, Antoine Marie, Lea Pradella, and Michael Bang Petersen. 2026. Social media users experience more political hostility in less economically equal and less democratic societies. *Nature Human Behaviour*.

Jeremy Bowles, Horacio Larreguy, and Shelley Liu. 2020. Countering misinformation via whatsapp: Preliminary evidence from the covid-19 pandemic in zimbabwe. *PloS one*, 15(10):e0240005.

Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. 2014. Detecting Online Harassment in Social Networks. *ICIS 2014 Proceedings*.

T Brewster. 2023. Google cuts company protecting people from surveillance to a 'skeleton crew,' say laid off workers. *Forbes*.

Sergiu Burlacu, Austėja Kažemekaitytė, Piero Ronzani, and Lucia Savadori. 2022. Blinded by worries: sin taxes and demand for temptation under financial worries. *Theory and Decision*, 92(1):141–187.

Dallas Card and Noah A. Smith. 2020. On Consequentialism and Fairness. *Frontiers in Artificial Intelligence*, 3.

Bruno Gabriel Salvador Casara, Caterina Suitner, and Jolanda Jetten. 2022. The impact of economic inequality on conspiracy beliefs. *Journal of Experimental Social Psychology*, 98:104245.

Raja Chatila and John C. Havens. 2019. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Osman Tokhi, and Endre E. Kadar, editors, *Robotics and Well-Being*, volume 95, pages 11–16. Springer International Publishing,

Cham. Series Title: Intelligent Systems, Control and Automation: Science and Engineering.

Laurent Cordonier and Florian Cafiero. 2024. Public sector corruption is fertile ground for conspiracy beliefs: A comparison between 26 western and non-western countries. *Social Science Quarterly*, 105(3):843–861.

Kiana Cox. 2024. *Most Black Americans believe US institutions were designed to hold Black people back*. Pew Research Center.

Georgina Curto, Svetlana Kiritchenko, Muhammad Hammad Fahim Siddiqui, Isar Nejadgholi, and Kathleen C. Fraser. 2025. Tackling social bias against the poor: a dataset and a taxonomy on aporophobia. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7010–7031, Albuquerque, New Mexico. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

Ernst-Jan De Bruijn and Gerrit Antonides. 2022. Poverty and economic decision making: a review of scarcity theory. *Theory and Decision*, 92(1):5–37.

Leon Derczynski and A. Zubiaga. 2020. Detection and Resolution of Rumors and Misinformation with NLP. In *COLING*. ECC: No Data (logprob: -123.494) tex.ids: derczynskiDetectionResolutionRumors2020a, derczynskiDetectionResolutionRumors2020b.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the Detection of Textual Cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(3):11–17.

Constantinos Djouvas, Christos Christodoulou, Antonis Charalampous, and Nikandros Ioannidis. 2023. Identifying euroscepticism using a text-as-data approach: An experimental study employing parliamentary speeches. In *18th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP 2023)*, pages 1–6.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 29–30, New York, NY, USA. Association for Computing Machinery.

Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political psychology*, 40:3–35.

A Down. 2026. Us builds website that will allow europeans to view blocked content. *The Guardian*.

Franziska Ehrke, Gloria Grommisch, Emma Penelope Busch, and Magdalena C. Kaczmarek. 2023. Populist attitudes predict compliance-related attitudes and behaviors during the covid-19 pandemic via trust in institutions. *Social Psychology*.

European Commission. 2025. The code of conduct on disinformation.

European Commission. Directorate General for Communications Networks, Content and Technology. 2018. *A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation.* Publications Office, LU.

European Parliament. Directorate General for Parliamentary Research Services. 2019. *Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism.* Publications Office, LU.

Cataldo F, Redecker C, Montanari E, Galariotis I, Greidanus H, and Quetel C. 2025. Rethinking societal resilience in a time of polycrisis. Technical report, Joint Research Center, Ispra (Italy).

Paul Farmer. 2004. An anthropology of structural violence. *Current anthropology*, 45(3):305–325.

H Field and J Vanian. 2023. Tech layoffs ravage the teams that fight online misinformation and hate speech. *CNBC*.

Batya Friedman and Peter H. Kahn. 2002. Human Values, Ethics, and Design. In *The Human-Computer Interaction Handbook*, pages 1177–1201. CRC Press.

Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman, editors, *Early Engagement and New Technologies: Opening up the Laboratory*, Philosophy of Engineering and Technology, pages 55–95. Springer Netherlands, Dordrecht.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting Political Bias in News Articles Using Headline Attention. In *Proceedings of the 2019 ACL Workshop*

BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 77–84, Florence, Italy. Association for Computational Linguistics.

Kiran Garimella and Simon Chauchard. 2025. Whatsapp explorer: A data donation tool to facilitate research on whatsapp. *Mobile Media & Communication*, 13(3):481–503.

Souvick Ghosh and Chirag Shah. 2019. Toward Automatic Fake News Classification. In *HICSS*.

Carol Gilligan. 1984. *In a different voice*. Harvard University, [Place of publication not identified]. Section: 174 pages.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. Ambifc: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Antonio Gramsci. 1929/2020. Selections from the prison notebooks. In *The applied theatre reader*, pages 141–142. Routledge.

Leon Graumas, R Latulippe David, and Tommaso Caselli. 2019. Twitter-based Polarised Embeddings for Abusive Language Detection. *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–7.

K. Gupta. 2025. Comparative Analysis of Encoder-Based and Decoder-Based Architectures for Automatic Conspiracy Theory Identification. In *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*, pages 394–398.

Matthew J Hornsey and Samuel Pearson. 2022. Cross-national differences in willingness to believe conspiracy theories. *Current Opinion in Psychology*, 47:101391.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602.

Jessica Jaiswal, Caleb LoSchiavo, and David C Perlman. 2020. Disinformation, misinformation and inequality-driven mistrust in the time of covid-19: lessons unlearned from aids denialism. *AIDS and Behavior*, 24(10):2776–2780.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Andreas Jungherr and Ralph Schroeder. 2021. Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media+ Society*, 7(1):2056305121988928.

Theara Khoun, Ate Poortinga, Nyein Soe Thwal, Iván González de Alba, Andrea McMahon, and Carlos Mendez. 2025. Mapping the dimensions of poverty through big data, socioeconomic surveys and machine learning in cambodia. *Social Indicators Research*, 180(3):1593–1618.

Robert V Kozinets and Ulrike Gretzel. 2024. Netnography evolved: New contexts, scope, procedures and sensibilities. *Annals of Tourism Research*, 104:103693.

Kimberley Kruijver, Neill Bo Finlayson, Beatrice Cadet, and Sico van der Meer. 2025. The disinformation lifecycle: An integrated understanding of its creation, spread and effects. *Discover Global Society*, 3(1):58.

David La Barbera, G Milanese, Georgios Peikos, Gabriella Pasi, Marco Viviani, et al. 2025. Beyond binary classification: ranking for information access in misinformation contexts. In *CEUR WORKSHOP PROCEEDINGS*, volume 4121, pages 1–7. CEUR-WS.

Bruno Latour. 2012. *We have never been modern*. Harvard university press.

Johannes Lattmann. 2025. Detecting eu sentiment in texts: A llm machine learning application for euroscepticism research. *OSF*.

Jochen L Leidner and Vassilis Plachouras. 2017. Ethical by Design: Ethics Best Practices for Natural Language Processing. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. ACM.

Robert Litschko, Max Müller-Eberstein, Rob Van Der Goot, Leon Weber-Genzel, and Barbara

Plank. 2023. Establishing trustworthiness: Rethinking tasks and model evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Erik Bran Marino, Jesus M Benitez-Baleato, and Ana Sofia Ribeiro. 2024. The polarization loop: How emotions drive propagation of disinformation in online media—the case of conspiracy theories and extreme right movements in southern europe. *Social Sciences*, 13(11):603.

Abraham Harold Maslow. 1966. *The psychology of science*. Harper and Row.

Hugo Mercier. 2020. *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Meta Transparency Center. 2024. Update on Meta's policy on the term 'Zionist'.

Todor Mihaylov and Preslav Nakov. 2016. Hunting for Troll Comments in News Community Forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Berlin, Germany. Association for Computational Linguistics.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1211–1228.

Evgeny Morozov. 2013. *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

Guberney Muneton-Santa, Daniel Escobar-Grisales, Felipe Orlando López-Pabón, Paula Andrea Pérez-Toro, and Juan Rafael Orozco-Arroyave. 2022. Classification of poverty condition using natural language processing. *Social Indicators Research*, 162(3):1413–1435.

Luca Nannini, Eleonora Bonel, Davide Bassi, and Michele Joshua Maggini. 2025. Beyond phase-in: assessing impacts on disinformation of the eu digital services act. *AI and Ethics*, 5(2):1241–1269.

Philip M Napoli. 2019. *Social media and the public interest: Media regulation in the disinformation age*. Columbia university press.

Nic Newman, Richard Fletcher, Craig T. Robertson, Amy Ross Arguedas, and Rasmus Kleis Nielsen. 2024. Reuters institute digital news report 2024. Report, Reuters Institute for the Study of Journalism, Oxford, UK.

Ilana Nguyen, Harini Suresh, and Evan Shieh. 2025. Representational harms in llm-generated narratives against nationalities located in the global south. In *HEAL Workshop, CHI*, volume 2025.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Nel Noddings. 1984. *Caring: a feminine approach to ethics and moral education*. University of California Press.

Martha C Nussbaum. 2011. Creating capabilities: The human development approach and its implementation. *Hypatia*, 24(3):211–215.

Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2021. Case Study: Deontological Ethics in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3784–3798, Online. Association for Computational Linguistics.

Žiga Puklavec, Christoph Kogler, Olga Stavrova, and Marcel Zeelenberg. 2023. What we tweet about when we tweet about taxes: A topic modelling approach. *Journal of Economic Behavior & Organization*, 212:1242–1254.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive Language Detection Using Multi-level Classification. In *Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer.

Michael J Reddy. 1979. *The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language*. Cambridge: Cambridge UP, Cambridge.

Timothy S Rich, Ian Milden, and Mallory Treece Wagner. 2020. Research note: Does the public support fact-checking social media? it depends who and how you ask. *The Harvard Kennedy School Misinformation Review*, 1(8).

Ronald E Robertson, Jon Green, Damian J Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. 2023. Users choose to engage with more partisan news than they are exposed to on google search. *Nature*, 618(7964):342–348.

Piero Ronzani. 2025. Towards the study of world misinformation. *Harvard Kennedy School Misinformation Review*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Miriam Schirmer, Nathan Walter, and Emőke-Ágnes Horvát. 2025. Disparities by design: Toward a research agenda that links science misinformation and socioeconomic marginalization in the age of ai. *Harvard Kennedy School Misinformation Review*.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674.

Merrill Singer. 2009. *Introduction to syndemics: A critical systems approach to public and community health*. John Wiley & Sons.

Carly Parnitzke Smith and Jennifer J Freyd. 2014. Institutional betrayal. *American psychologist*, 69(6):575.

Nadiya Straton, Hyeju Jang, and Raymond Ng. 2020. Stigma annotation scheme and stigmatized language detection in health-care discussions on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1178–1190, Marseille, France. European Language Resources Association.

Li Qian Tay, Stephan Lewandowsky, Mark J Hurlstone, Tim Kurz, and Ullrich KH Ecker. 2024. Thinking clearly about misinformation. *Communications Psychology*, 2(1):4.

U.S. House of Representatives. 2023. H.res.894 - Strongly condemning and denouncing the drastic rise of antisemitism in the United States and around the world. Technical report, 118th Congress. Accessed: 2026-02-10.

Carolina Villegas-Galaviz and Kirsten Martin. 2023. Moral distance, AI, and the ethics of care. *AI & society*, pages 1–12.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Silvio Waisbord. 2018. Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism studies*, 19(13):1866–1878.

Charlotte Ward and David Voas. 2011. The emergence of conspirituality. *Journal of Contemporary Religion*, 26(1):103–121.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 88–93.

Max Weber. 1949. *Max Weber on the methodology of the social sciences*. Free Press.

Jonathan F. Will. 2011. A Brief Historical and Theoretical Perspective on Patient Autonomy and Medical Decision Making. *Chest*, 139(3):669–673.

Shehel Yoosuf and Yin Yang. 2019. Fine-Grained Propaganda Detection with Fine-Tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China. Association for Computational Linguistics.

Unraveling Deceptive Narratives: A Study of Conceptual Frameworks

Elie Alhajjar
RAND
Arlington, VA, USA
ealhajjar@rand.org

Abstract

In the era of digital communication, the rapid spread of information presents significant challenges to society. This paper provides an in-depth examination of the existing frameworks developed to understand and address these phenomena. More precisely, this paper categorizes and compares various frameworks, including typology-based, process-oriented, impact-oriented, and actor-centric approaches. It highlights the strengths and limitations of each framework type, with a particular focus on their applicability to combat false information in diverse contexts. The paper underscores the importance of adopting a holistic and flexible approach that integrates multiple frameworks and adapts to the evolving nature of technology, particularly AI-driven false and misleading content.

Keywords: Information spread, frameworks, policy.

1. Introduction

As false information spreads more rapidly and widely than ever before, understanding the frameworks that scholars and practitioners use to analyze and address these phenomena is essential for developing effective strategies to combat its spread, protect public trust, and mitigate its societal impacts. The spread of falsehoods, both unintentional and strategically engineered, has become increasingly significant. Both aspects refer to the spread of false or misleading information, but they differ fundamentally in purpose and impact. Understanding these differences is crucial for comprehending the broader challenges they pose to society.

Some forms of false content are shared without malicious intent (Baines and Elliott, 2020). Often, individuals or organizations believe what they are sharing to be true, despite it being incorrect. This can stem from misunderstandings, insufficient verification, or the use of unreliable sources. For example, during fast-moving news events, early reports that lack confirmation are frequently circulated by well-meaning individuals trying to make sense of unfolding developments. Even when shared in good faith, such content can perpetuate errors and mislead others.

In contrast, other forms are deliberately constructed and disseminated to deceive or manipulate (Baines and Elliott, 2020). These efforts are typically intentional and often orchestrated by actors seeking specific outcomes such as shaping public opinion, influencing elections, harming reputations, or sowing discord. These campaigns are frequently sophisticated, exploiting digital infrastructure and cognitive biases to extend their reach. Their strategic and deliberate nature makes them especially dangerous, as they erode institutional trust, deepen social divisions, and threaten the stability of democratic systems.

The complexity of detecting and responding to false information further underscores the importance of studying these phenomena. The methods used to spread false information are becoming more sophisticated, employing technologies such as bots, deepfakes, and algorithmic amplification to reach wider audiences and evade detection (Moral, 2022). This technological evolution makes it more challenging to identify and counteract misleading content effectively. In addition, the proliferation of such content raises important ethical and legal questions (Dalessandro et al., 2019). Society must navigate the delicate balance between protecting free speech and preventing harm, determining the responsibilities of platforms and governments in regulating content, and upholding the public's right to accurate information.

This paper aims to provide a comprehensive overview of the key frameworks developed to understand and address inaccurate or deceptive narratives. It starts with an introduction that sets the context, followed by an overview of historical context. The core sections present and categorize various theoretical and methodological frameworks, offering a comparative analysis to highlight their strengths, weaknesses, and applicability. The discussion synthesizes key insights, addressing the broader implications and suggesting future research directions.

2. Historical Context

The historical context of false and misleading content highlights the persistent nature of these

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 16–24

challenges and underscores the importance of understanding their evolution. The phenomena of circulating deceptive or inaccurate narratives are not new; they have existed in various forms throughout history. However, the methods and speed at which false information can spread have evolved dramatically with the advent of new technologies.

Historically, such content was often spread through word of mouth, pamphlets, or print media. In ancient civilizations, this often took the form of rumors or oral traditions that were passed down through generations. These could include exaggerated tales of heroism, false reports of enemy movements during wars, or distorted accounts of natural disasters. One of the earliest documented cases of strategic falsehoods dates back to the Roman Empire (Frasier, 2020), where political rivals would spread false rumors to undermine each other's credibility.[1]

In the medieval period, the Catholic Church played a significant role in controlling the flow of information. The selective use or distortion of truth became a tool used to enforce religious orthodoxy and suppress dissent. For instance, during the Inquisition, accusations of heresy were often based on exaggerated information, leading to the persecution of individuals or groups who were seen as threats to the church's authority. The church also controlled the dissemination of knowledge, often suppressing scientific discoveries that contradicted religious doctrine, such as the heliocentric theory proposed by Copernicus (Hess and Allen, 2008).

The invention of the printing press in the 15th century marked a significant turning point in this context. While the printing press revolutionized the spread of information, it also made it easier to disseminate false information on a larger scale. Pamphlets, leaflets, and books could now be mass-produced and distributed widely, often without any means of verifying the accuracy of the content. This period saw the rise of propaganda as a powerful tool in religious and political conflicts, such as during the Protestant Reformation, where both sides used printed materials to spread their messages and discredit their opponents (Edwards, 1994). Martin Luther's 95 Theses, for example, were widely circulated, but so were counter-pamphlets filled with distorted portrayals of his arguments and intentions.

The 19th and early 20th centuries witnessed the growth of mass media, particularly newspapers, which became the primary means of information dissemination. This era also saw the emergence of yellow journalism, a term used to describe sensationalist and often misleading news designed to attract readers and increase sales. Publishers like William Randolph Hearst and Joseph Pulitzer engaged in fierce competition, leading to the publication of fabricated stories (Spencer, 2007). The Spanish-American War is a prime example of how yellow journalism contributed to the spread of misleading information, as reports in American newspapers fueled public support for the war, often based on false accounts of events.

During World Wars I and II, propaganda was widely used by various governments to influence public perception and morale, often blurring the lines between true and falsified information. The Cold War era saw the rise of sophisticated information campaigns, particularly by state actors. The Soviet Union, for instance, engaged in active measures, a term used to describe a range of activities aimed at influencing global opinion and destabilizing Western governments. An infamous example was Operation INFEKTION,[2] which falsely claimed that the U.S. government had created the HIV/AIDS virus as a biological weapon (Selvage, 2021). These activities were meticulously planned and executed, often using forgeries, false media reports, and manipulated documents to create and spread false narratives.

In the modern era, the internet and social media have fundamentally reshaped how inaccurate and manipulative content is produced and circulated. The internet has made it easier than ever to create and spread false information on a global scale with minimal cost and effort (Vosoughi et al., 2018). Social media platforms, in particular, have become fertile ground for the rapid dissemination of all kinds of information, as content can go viral within minutes, reaching millions of people around the world. The decentralized nature of the internet also means that it is more challenging to control or regulate the spread of false information, leading to a proliferation of conspiracy theories, fake news, and other forms of misleading content.

3. Existing frameworks

A variety of frameworks have been developed to understand, categorize, and address how false or misleading content is created, spread, and mitigated. In this section, we explore the major existing frameworks, categorizing them based on their focus and approach.

[1] Octavian (later Augustus) waged a propaganda campaign against his rival Mark Antony by spreading rumors portraying Antony as a traitor and a puppet of Cleopatra.

[2] Also known as Operation Denver.

3.1 Typology-Based Frameworks

They focus on categorizing falsehoods into distinct types based on various factors such as intent, content, and dissemination method.

1) Intent-Based Typologies: One of the most common typological approaches is to classify false information based on the intent behind its creation and dissemination. For example, some frameworks distinguish between unintentional and intentional spread as primary categories. Further subcategories might include the deliberate spread of truthful information with the intent to cause harm, such as doxing or releasing private information (Carmi et al., 2020).

2) Content-Based Typologies: Another approach focuses on the nature of the content itself. These frameworks classify information spread based on the type of falsehood or distortion present in the content (First Draft, 2017). For example, Wardle introduced a typology that categorizes false information into seven types: satire or parody, false connection, misleading content, false context, impostor content, manipulated content, and fabricated content. Each type represents a different way in which the truth is distorted, providing a detailed map of the information landscape.

3) Dissemination Method-Based Typologies: Some frameworks classify falsehoods based on the methods and channels used to spread them. These might include distinctions between organic spread (e.g., via social media sharing) and coordinated campaigns (e.g., through bot networks or paid advertisements). Understanding the dissemination methods helps in identifying the mechanisms by which false information reaches and influences audiences (Tsfati et al., 2020).

3.2 Process-Oriented Frameworks

They focus on the lifecycle of false and misleading narratives, examining how these phenomena are created, disseminated, consumed, and ultimately affect audiences. These frameworks often draw from communication and media studies to map out the stages through which false information travels, and the factors that influence each stage.

1) The Information Disorder Framework: This type of frameworks identifies three key stages in the lifecycle of false information: creation, production, and distribution (Wardle and Derakhshan, 2017). It also distinguishes between three elements involved: agents (creators, producers, and distributers), messages (the content itself), and interpreters (audiences). This framework is useful for understanding how deceptive narratives are constructed and spread across different platforms and contexts.

2) The Information Lifecycle Model: Another process-oriented approach is the information lifecycle model, which outlines the stages through which false or misleading content moves from its initial creation to its eventual impact on public perception. These stages typically include creation, amplification, dissemination, and correction (Puska and Pereira, 2023). This model emphasizes the role of social media algorithms, news cycles, and audience engagement in the amplification and spread of such content.

3) The Knowledge-Based Approach: This type of framework focuses on how individuals process and interpret information (Bode and Vraga, 2017). It examines the cognitive processes that occur when people encounter false information, including how they decide whether to believe it or share it. The model suggests interventions at different stages of information processing, such as providing corrective information or promoting critical thinking skills, to reduce the spread and impact of misleading information.

3.3 Impact-Oriented Frameworks

They are concerned with the consequences of falsehoods on individuals, communities, and societies. They assess the effects of false information and help identify the broader implications on public opinion. Each model below represents a cluster of frameworks rather than a single framework, considered as a many-to-one mapping.

1) The Trust Erosion Model: This family of frameworks explores how false information campaigns erode public trust in institutions, media, and democracy (Atele-Williams and Marsh, 2023). It posits that repeated exposure to false information, especially when it aligns with existing biases or distrust, leads to a gradual decline in trust. The model is particularly relevant for understanding the long-term societal impacts of falsehoods and the challenges in restoring trust once it has been damaged.

2) The Public Health Impact Model: This family of models examines the spread of health-related narratives (e.g., about vaccines or treatments) and its impact on public health, such as vaccine hesitancy or non-compliance with health guidelines (Pulido et al., 2020). The framework also considers the role of public health communication in countering such narratives and promoting accurate information.

3) Behavioral Impact Model: This cluster of frameworks looks at how information flows

influence individual and collective behavior (Alhajjar, 2022). It considers factors such as cognitive biases, social influence, and emotional responses that lead individuals to accept or act on false information. The framework is useful for designing interventions that address the behavioral drivers of false information spread, such as social norms campaigns or behavioral nudges.

3.4 Actor-Centric Frameworks

They focus on the roles and motivations of different actors involved in the creation, dissemination, and consumption of distorted narratives. These frameworks analyze the behaviors, strategies, and networks of various stakeholders, including individuals, organizations, governments, and platforms.

1) The Actor-Network Theory (ANT) : This sociological framework explores the complex relationships between different actors (both human and non-human, such as algorithms) involved in the spread of falsehoods (Latour, 2005). ANT examines how these actors form networks that facilitate the dissemination of false information and how power dynamics within these networks influence the spread and impact of information. The framework is useful for understanding the interconnectedness of different actors and the systemic nature of information ecosystems.

2) The Political Economy Framework: This approach focuses on the economic and political motivations behind false information campaigns (Dobson and Hunsinger, 2016). It examines how state and non-state actors use information as a tool for political gain, financial profit, or social influence. The framework also considers the role of media ownership, advertising revenue models, and regulatory environments in shaping the spread of misleading information. Understanding these motivations is crucial for designing policies and interventions that address the root causes of falsified information.

3) The Platform Responsibility Framework: With the rise of social media and digital platforms, this framework addresses the responsibilities of these platforms in managing falsehoods (Ramdas, 2022). It examines the role of algorithms, content moderation policies, and platform governance in either exacerbating or mitigating the spread of false information. The framework also explores the ethical and legal implications of platform actions, such as content removal or algorithmic transparency.

4. Categorization of frameworks

The diverse array of frameworks for understanding false and misleading content can be overwhelming due to the various perspectives and methodologies they encompass. In this section, we divide the existing frameworks into thematic, methodological, and geographical or cultural considerations. Practically, when frameworks are categorized by themes, methods, or regional contexts, the focus is on tailoring research approaches and interventions to specific settings or problems. In contrast, when frameworks are grouped into actor, process, impact, and typology categories, they are examined based on what aspect of the problem they target.

4.1 Thematic Categorization

It involves grouping frameworks based on the primary themes or issues they address. This approach helps illuminate the specific domains each model concentrates on, whether it be political, social, health-related, or technological (Tsfati et al., 2020; Pulido et al., 2020).

1) Political Frameworks: Frameworks in this category focus on the role of false narratives in political contexts. They examine how false information is used to influence elections, shape public opinion, and destabilize political systems (Howard et al., 2018; Wardle & Derakhshan, 2017). For example, frameworks that analyze information campaigns during elections or state-sponsored propaganda efforts fall into this category. These frameworks often emphasize the strategic use of deceptive arguments by political actors to achieve specific goals, such as voter manipulation or undermining opponents.

2) Social Frameworks: Socially oriented frameworks explore how information can affect social dynamics and relationships. They may focus on how false information spreads within communities, influences social norms, or exacerbates societal divisions. Frameworks in this category often address issues like the role of social media in amplifying falsehoods, the formation of echo chambers, and the impact of misleading information on social cohesion (Sunstein, 2017; Vosoughi et al., 2018). These frameworks are particularly relevant for understanding how information contributes to polarization and the fragmentation of public discourse.

3) Health-Related Frameworks: Given the significant impact of false information on public health, several frameworks specifically address the spread and effects of health-related falsehoods (Pulido et al., 2020; Bode & Vraga,

2017). Health-related frameworks often emphasize the need for accurate communication, the dangers in undermining public health efforts, and strategies for combating inaccurate health narratives through education and public awareness campaigns.
4) Technological Frameworks: Technological frameworks focus on the role of digital platforms, algorithms, and artificial intelligence (AI) in the spread of information. They explore how technology facilitates the rapid dissemination of false information, the role of social media algorithms in promoting sensationalist content, and the potential for automated tools like bots and deepfakes to spread information (Gradoń et al., 2021; Gillespie, 2018). These frameworks often address the challenges of regulating digital platforms and the ethical implications of technological interventions designed to counter misleading information.

4.2 Methodological Approaches

They group frameworks based on the research methods they employ. This categorization highlights the diversity of techniques used to study misleading content, ranging from qualitative analyses to quantitative data-driven models (Lazer et al., 2018).
1) Qualitative Frameworks: Qualitative frameworks often involve case studies, interviews, content analysis, and other non-numerical methods to explore false and misleading content (Wardle & Derakhshan, 2017). These frameworks are valuable for understanding the nuanced and contextual factors that influence how false information is created, spread, and received. For example, qualitative studies may examine the narratives used in information campaigns, the motivations of actors involved in spreading false information, or the experiences of individuals who encounter false and misleading content.
2) Quantitative Frameworks: Quantitative frameworks rely on numerical data and statistical analysis to study false and misleading content. These frameworks often involve large-scale data collection, such as social media analytics, survey data, or experiments designed to measure the effects of false and misleading content (Vosoughi et al., 2018). Quantitative approaches are useful for identifying patterns in the spread of false and misleading content, assessing the prevalence of false information, and evaluating the effectiveness of interventions.
3) Mixed-Methods Frameworks: Some frameworks combine qualitative and quantitative approaches to offer a more comprehensive understanding of false and misleading content. Mixed-methods frameworks might use qualitative research to explore the context and motivations behind false and misleading content, followed by quantitative analysis to measure the scale and impact of these phenomena (Tsfati et al., 2020). This approach allows for a more holistic view, capturing both the detailed, context-specific elements and the broader trends in false and misleading content spread and impact.

4.3 Computational Frameworks

With the rise of big data and machine learning, computational frameworks have become increasingly important in the study of false and misleading content. These frameworks use algorithms, network analysis, and other computational tools to model the spread of false and misleading content, detect false information, and simulate the effects of different interventions (Gradoń et al., 2021; Lazer et al., 2018).
1) Geographical and Cultural Considerations involve categorizing frameworks based on the regions or cultural contexts in which they are applied. False and misleading content does not operate in a vacuum; it is deeply influenced by the social, cultural, and political environments in which it spreads (Tsfati et al., 2020).
2) Regional Frameworks: Some frameworks are designed to address false and misleading content in specific geographical regions, such as North America, Europe, Asia, or Africa. These frameworks consider the unique political, social, and media landscapes of each region, which influences how false and misleading content spreads and is perceived (Howard et al., 2018). For example, frameworks developed for Western democracies might focus on the role of free speech and the media, while those for authoritarian regimes might emphasize state control and censorship.
3) Cultural Frameworks: Cultural frameworks examine how cultural factors, such as language, values, and traditions, shape the creation and spread of false and misleading content. These frameworks recognize that false and misleading content is often tailored to resonate with specific cultural beliefs or biases, making it more effective in certain communities. For instance, information campaigns may exploit cultural tensions or stereotypes to create division or mistrust.
4) Cross-Cultural Frameworks: Cross-cultural frameworks compare the spread and impact of false and misleading content across different cultural contexts. These frameworks are useful for identifying universal patterns in false and misleading content spread, as well as context-specific factors that influence how false and

misleading content is received and acted upon (Pennycook and Rand, 2021). Cross-cultural studies can reveal how different societies respond to false and misleading content and what lessons can be learned from various approaches to combating false information.

5. Discussion

In this section, we undertake a comparative analysis of the existing frameworks for understanding false and misleading content. By systematically comparing these frameworks, we aim to identify their strengths, weaknesses, and areas of overlap or divergence. This analysis will help clarify which frameworks are most effective in addressing specific aspects of false and misleading content based on scope and effectiveness and summarize key insights from the study and implications for future research in this field.

5.1 Comparative Analysis

Typology-based frameworks are designed to offer a comprehensive categorization of false and misleading content, attempting to classify all forms of false information into a structured taxonomy. By creating categories based on criteria such as intent, content, and medium, typology frameworks allow for a systematic analysis of the different types of false and misleading content that exist. This broad classification system is advantageous because it provides a high-level overview that can help in identifying patterns and trends in the spread of false and misleading content. However, the very breadth of typology frameworks can also be a limitation as they may not delve deeply into the specific processes that lead to the creation and dissemination of false and misleading content.

Process-oriented frameworks, in contrast, focus on the lifecycle of false and misleading content. This narrower focus allows for detailed insights into the stages of false and misleading content spread, identifying critical points where interventions could be most effective. By understanding these processes, stakeholders can develop targeted strategies to disrupt the spread of false and misleading content at key stages. However, the focus on processes can limit the ability of these frameworks to account for the broader social, political, or cultural contexts that influence the spread of false and misleading content. While they provide valuable insights into the mechanics of false and misleading content dissemination, process frameworks may not fully capture the external factors that shape the environment in which false and misleading content thrives.

Impact-oriented frameworks take a different approach by concentrating on the consequences of false and misleading content, rather than its classification or lifecycle. These frameworks are particularly effective in highlighting the tangible effects of false and misleading content by linking false information to specific outcomes. By focusing on the measurable consequences of false and misleading content, impact frameworks provide critical insights into the harm caused by false information and the importance of addressing it. However, the reliance on measurable outcomes can be both a strength and a limitation. While impact frameworks excel in demonstrating the immediate and direct effects of false and misleading content, they may struggle to capture the full range of impacts, particularly those that are long-term, indirect, or difficult to quantify.

Finally, actor-centric frameworks offer a broad scope by considering the wide range of players involved in the creation, dissemination, and consumption of false and misleading content, as well as the complex relationships between them. By focusing on the motivations and behaviors of key actors, actor-centric frameworks can reveal the underlying drivers of false and misleading content. However, the inherent complexity of actor-centric frameworks can make them challenging to apply. The interactions between various actors are often intricate and not easily discernible, especially when motivations are hidden or intentionally obscured. This complexity requires significant resources and expertise to untangle, making actor-centric frameworks more difficult to implement effectively compared to other frameworks that focus on more straightforward aspects of false and misleading content.

5.2 Key Insights

Each framework brings unique strengths to the table, contributing valuable perspectives on how false information is generated, disseminated, and impacts society. However, the analysis also highlights the limitations of each approach, suggesting that a multifaceted strategy combining elements from multiple frameworks may be the most effective way to combat false and misleading content. One of the most significant insights is that no single framework can fully address the complexity of false and misleading content. This suggests that relying on one framework alone may lead to an incomplete understanding of the problem and potentially ineffective interventions.

Another important insight is the critical role that context plays in the effectiveness of different frameworks. Effectiveness in this context is

measured by how well a framework achieves its intended purpose, which can vary depending on the framework's focus. False and misleading narratives are deeply influenced by social, political, and cultural factors in different environments. In typology-based frameworks, effectiveness is measured by how well the framework categorizes different types of false and misleading content based on key factors like intent, content, or dissemination method. A typology framework is considered effective if it provides a clear, comprehensive, and useful classification system that helps researchers and practitioners distinguish between various forms of false information. For process-oriented frameworks, effectiveness is determined by their ability to map the lifecycle of false and misleading content, identify critical intervention points, and develop strategies to disrupt its dissemination. Impact-oriented frameworks are judged by how accurately they assess the consequences of false and misleading content, such as changes in public opinion or behavior, while actor-centric frameworks are evaluated based on their capacity to reveal the motivations and behaviors of those involved in spreading false and misleading content.

False and misleading content that resonates in one cultural setting may not have the same impact in another, and the strategies used to combat it must be tailored accordingly. The analysis also highlights that false and misleading content is not solely a communication issue but also intersects with other disciplines, bringing diverse methodologies and insights to the discussion. By combining these perspectives, a more robust and comprehensive understanding of false and misleading content can be developed. Moreover, the analysis reveals that the rapid evolution of digital technologies necessitates continuous adaptation of existing frameworks. False and misleading narratives are increasingly spread through new and evolving platforms, such as social media, where traditional approaches may no longer be sufficient. This dynamic environment requires frameworks that are not only comprehensive but also flexible and adaptable to change, and that can keep pace with technological advancements as well as the changing nature of information dissemination.

5.3 Implications for Future Research

One of the primary implications for future research is the need for more integrative approaches that combine the strengths of multiple frameworks. For instance, combining typology frameworks with process frameworks could provide a more comprehensive understanding of both the classification of false and misleading content and the mechanisms by which it spreads. Similarly, integrating actor frameworks with impact frameworks could help elucidate how the motivations of key players influence the tangible outcomes of false and misleading content. Future research should prioritize developing hybrid frameworks that draw on the strengths of existing models while addressing their respective shortcomings.

On the level of contextualization of false and misleading content, future research should focus on comparative studies that examine how false and misleading content operates across diverse contexts, including non-Western societies that are often underrepresented in the literature. This would not only broaden the understanding of false and misleading content globally but also inform the development of context-specific interventions that are more likely to be effective in diverse environments.

In addition to the plethora of false and misleading content instances, a new phenomenon has emerged in the last couple of years: AI-enabled false and misleading content (Newsguard, 2025). AI-driven technologies, which include everything from automated news outlets that produce content with minimal or no human intervention to sophisticated AI image generators that create convincing but entirely fabricated visuals, have opened new avenues for the production and dissemination of misleading information (Alhajjar and Lee, 2022). AI's capabilities to generate large volumes of content quickly and convincingly, false and misleading content purveyors now have powerful tools at their disposal to create and spread false narratives on an unprecedented scale. This development poses serious challenges to the integrity of information ecosystems. The line between genuine and fabricated content increasingly blurs, making it harder for the public to distinguish truth from falsehood. The ease with which these tools can be used to produce deceptive content underscores the urgent need for robust strategies to detect and counteract AI-generated false and misleading content.

Finally, there is a pressing need for interdisciplinary research that brings together scholars from various fields to tackle the complex problem of false and misleading content. Future research should encourage collaboration across these fields to develop more comprehensive and multidimensional frameworks. This interdisciplinary approach would facilitate a deeper understanding of the psychological, social, and technological factors that drive false and misleading content, leading to more effective strategies for prevention and intervention.

6. Conclusion

The challenges posed by false and misleading content are among the most pressing issues facing societies today. The frameworks discussed in this paper provide valuable tools for analyzing the spread of false and misleading content and pinpoint the limitations of any single approach.

The selection and improvement of frameworks for combating false and misleading content requires a nuanced understanding of the problem's complexity. Each framework, whether typology-based, process-oriented, impact-focused, or actor-centric, offers distinct advantages and limitations. By choosing the right framework for the specific context and continuously improving upon existing models, researchers and practitioners can develop more effective strategies to counter false and misleading content. The key to addressing new challenges, including AI-enabled false and misleading content, lies in embracing a holistic and flexible approach. By integrating multiple frameworks, adapting strategies to specific contexts, and fostering interdisciplinary collaboration, more effective methods can be developed for combating false and misleading content.

As technology continues to evolve, it is essential to remain vigilant and proactive, continuously updating and refining our frameworks to keep pace with new developments. This ongoing effort requires not only innovation but also a commitment to transparency, accountability, and public trust. By staying ahead of emerging threats and fostering a culture of critical thinking, we can build a more resilient information ecosystem for the future.

7. Acknowledgments

The author would like to thank Bridget Chan at New America, DC for her support during the project. Her comments made major improvements to the clarity of the document.

8. Bibliographical References

Alhajjar, E. (2022). Alternate reality: The use of disinformation to normalize extremism. In The Great Power Competition Volume 3: Cyberspace: The Fifth Domain (pp. 157-165). Springer International Publishing.

Alhajjar, E., & Lee, K. (2022). The U.S. cyber threat landscape. In European Conference on Cyber Warfare and Security (Vol. 21, No. 1, pp. 18-24).

Atele-Williams, T., & Marsh, S. (2023). Information trust model. Cognitive Systems Research, 80, 50-70.

Baines, D., & Elliott, R. J. R. (2020). Defining misinformation, disinformation and malinformation: An urgent need for clarity during the COVID-19 infodemic. Discussion Papers, 20(06), 20-06.

Bode, L., & Vraga, E. K. (2017). See something, say something: Correction of global health misinformation on social media. Health Communication, 33(9), 1131-1140. https://doi.org/10.1080/10410236.2017.1331312

Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. Internet Policy Review, 9(2), 1-22.

Dalessandro, R. C., Guimarães, J. A. C., & Campbell, D. G. (2019). Fake news as an emergent subject domain: Conceptual and ethical perspectives for the development of a critical knowledge organization. In The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization (pp. 208-217). Ergon-Verlag.

Dobson, K., & Hunsinger, J. (2016). The political economy of WikiLeaks: Transparency and accountability through digital and alternative media. Interactions: Studies in Communication & Culture, 7(2), 217-233.

Edwards Jr, M. U. (1994). Printing, Propaganda, and Martin Luther. Fortress Press.

First Draft (2017). Fake news. It's complicated. https://www.firstdraftnews.org/articles/fake-news-complicated

Fraser, M. (2020). In truth: a history of lies from ancient Rome to modern America. Rowman & Littlefield.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven: Yale University Press. https://doi.org/10.12987/9780300235029

Gradoń, K. T., Hołyst, J. A., Moy, W. R., Sienkiewicz, J., & Suchecki, K. (2021). Countering misinformation: A multidisciplinary approach. Big Data & Society, 8(1).

Hess, P. M., & Allen, P. L. (2008). Catholicism and science.

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). The IRA, social media and political polarization in the United States, 2012-2018. Computational Propaganda Research Project, University of Oxford.

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford University Press.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096.

Moral, P. (2022). The challenge of disinformation for national security. In Security and Defence: Ethical and Legal Challenges in the Face of Current Conflicts (pp. 103-119). Springer International Publishing.

Newsguard (2025). AI Tracking Center. https://www.newsguardtech.com/special-reports/ai-tracking-center/

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. Trends in cognitive sciences, 25(5), 388-402.

Pidot, J. R. (2019). Environmental nihilism. Arizona Journal of Environmental Law & Policy, 10(1).

Pulido, C. M., Ruiz-Eugenio, L., Redondo-Sama, G., & Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. International Journal of Environmental Research and Public Health, 17(7), 2430.

Puska, A. A., & Pereira, R. (2023). Exploring digital misinformation as a sociotechnical phenomenon: Insights from a small-scale study. In Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems (pp. 1-12).

Ramdas, V. (2022). Identifying an actionable algorithmic transparency framework: A comparative analysis of initiatives to enhance accountability of social media platforms. National Law University Delhi Student Law Journal, 4, 74.

Selvage, D. (2021). Operation “Denver” The East German Ministry for State Security and the KGB’s AIDS Disinformation Campaign, 1986–1989 (Part 2). Journal of Cold War Studies, 23(3), 4-80.

Spencer, D. R. (2007). The yellow journalism: The press and America’s emergence as a world power. Northwestern University Press.

Sunstein, C. R. (2018). #Republic: Divided Democracy in the Age of Social Media (New edition). Princeton University Press. https://doi.org/10.2307/j.ctv8xnhtd

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.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27). Council of Europe.

High Accuracy, Low Generalization: Structural Homogeneity and Cross-Dataset Evaluation in Fake-News Benchmarks

Hiram Calvo, Mayte H. Laureano
Center for Computing Research
Instituto Politécnico Nacional
hcalvo@cic.ipn.mx, mhernandezl2021@cic.ipn.mx

Abstract

State-of-the-art fake-news classifiers frequently report near-ceiling accuracy on widely used benchmarks such as ISOT, Misinfo, and WELFake. We argue that such results often reflect structural homogeneity and provenance-based separability rather than robust claim-level veracity inference. Anchored in the Information Disorder framework, we analyze how dataset construction operationalizes the notion of "fake" and how this shapes model behavior. We conduct systematic bidirectional cross-dataset experiments across six transfer directions and evaluate performance not only by mean accuracy, but also by variance and directional asymmetry. Results reveal substantial degradation under distribution shift and pronounced transfer asymmetries between dataset pairs. Although not always achieving the highest mean accuracy, affective augmentation combining dimensional (VAD) and categorical (Ekman) representations yields the lowest variance and smallest directional gap, indicating superior cross-domain stability. Our findings expose the disconnect between accuracy-driven benchmarking and construct-valid evaluation. We argue that progress in fake-news detection requires shifting from isolated in-domain optimization toward robustness-oriented, bidirectional, and distribution-aware assessment practices.

Keywords: fake news detection, information disorder, dataset bias, cross-dataset evaluation, distribution shift

1. Introduction

Fake-news classification has become a canonical supervised NLP task, frequently reporting accuracy values exceeding 98%—and in some cases approaching or surpassing 99%—on benchmark datasets. Systematic reviews confirm the prevalence of such near-ceiling results across a wide range of machine-learning architectures (Villela et al., 2023). At face value, these figures may suggest that automated veracity detection is largely solved.

At the same time, the notion of "fake news" is not a well-defined natural category. Prior work has emphasized its overlap with broader forms of information disorder (Lazer et al., 2018), the role of deliberate deception and imitation of journalistic style (Gelfert, 2018), and the diversity of phenomena grouped under the term, including satire, propaganda, and fabrication (Tandoc et al., 2018). These perspectives highlight that the target label in fake-news detection is conceptually heterogeneous and socially constructed.

In practice, however, machine-learning research depends heavily on publicly available supervised datasets. Among these, the ISOT dataset (Ahmed et al., 2017) has become particularly influential. The original ISOT study reported approximately 92% accuracy using n-gram features and classical classifiers, while more recent work using deep learning and transformer-based architectures reports near-ceiling performance, often exceeding 98% (Lawal and Abdulrauf, 2025; Villela et al., 2023). Such results raise a methodological question: what exactly is being learned? In ISOT, true articles are largely crawled from Reuters, whereas fake articles are collected from outlets flagged as unreliable. This provenance-based construction introduces systematic lexical, stylistic, and formatting differences between classes, allowing models to exploit ecosystem-level regularities rather than perform claim-level epistemic inference.

Three corpora—ISOT, Misinfo, and WELFake (Ahmed et al., 2017; Peutz, 2023; Verma et al., 2021)—are widely used due to their size and binary labeling schemes. However, each encodes different assumptions about the negative class and is constructed through distinct pipelines, potentially introducing spurious correlations between labels, sources, and stylistic patterns.

When class labels correlate strongly with source identity or stylistic conventions, classification may approximate corpus discrimination rather than epistemic reasoning. This raises a central validity question: what do near-ceiling benchmark results actually measure?

We argue that high intra-corpus accuracy often reflects structural separability and dataset homogeneity rather than transferable veracity reasoning. To examine this, we benchmark representative modeling setups on ISOT, Misinfo, and WELFake and complement them with controlled cross-dataset transfer experiments. The results reveal a substantial generalization gap: models that appear nearly perfect under random within-dataset splits degrade markedly when evaluated

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 25–33
12 May 2026.

across corpora. By situating our findings within the Information Disorder framework (Wardle and Derakhshan, 2017), we shift the focus from raw accuracy to construct validity and robustness under distributional shift.

2. Related Work

2.1. Fake-News Detection Models

Early work on fake-news detection relied on lexical features such as n-grams, TF–IDF representations, and classical classifiers including SVM, logistic regression, and decision trees (Ahmed et al., 2017). On datasets such as ISOT, these approaches already achieved accuracy above 90%, suggesting strong lexical separability between classes.

With the rise of deep learning, convolutional and recurrent architectures (CNN, LSTM, BiLSTM) became common (Kaliyar et al., 2020). Subsequent transformer-based approaches, particularly BERT and its variants, further increased reported performance (Kaliyar et al., 2021). Systematic reviews confirm that transformer-based architectures now dominate the field and frequently report accuracy exceeding 95%, with some studies approaching or surpassing 99% on benchmark corpora (Villela et al., 2023).

Hybrid models combining contextual embeddings with gradient boosting or ensemble methods also report near-ceiling performance. For example, BERT embeddings combined with LightGBM achieve 99% accuracy on ISOT under random splits (Essa et al., 2023). Similarly, recent BiLSTM-based approaches on ISOT report 98.98% accuracy and recall above 99% (Lawal and Abdulrauf, 2025). Such figures suggest that, at least under standard evaluation protocols, separability between classes is extremely high.

Beyond English-language corpora, multilingual and cross-lingual fake-news detection has also expanded. Large language models (LLMs) and instruction-tuned transformers have recently been explored for misinformation detection, often achieving strong zero-shot or few-shot results (Li et al., 2024; Koka et al., 2024). However, these studies typically evaluate within-dataset performance and rarely stress-test cross-domain transfer.

Our work differs from prior studies by explicitly quantifying directional asymmetry and demonstrating that near-ceiling performance can be reproduced by minimal lexical rules.

2.2. Dataset Bias, Distribution Shift, and Shortcut Learning

Parallel to advances in model architecture, research in NLP and machine learning has highlighted the risk of shortcut learning, whereby models exploit highly predictive but semantically shallow correlations present in the training data (Geirhos et al., 2020). In text classification, such shortcuts may include stylistic markers, formatting artifacts, or source-specific lexical cues.

In the context of fake-news detection, provenance-based labeling is particularly susceptible to shortcut exploitation. If true and fake classes are drawn from distinct outlet ecosystems, models may learn outlet identification rather than veracity inference. Surveys of misinformation detection acknowledge the presence of dataset bias and limited generalization, though systematic stress-testing remains rare (Islam et al., 2020; Villela et al., 2023).

Recent work has explicitly investigated cross-dataset generalization. Pszona et al. (2023) show that models trained on one fake-news corpus often degrade significantly when evaluated on another, suggesting that dataset-specific artifacts drive performance. Similarly, studies on domain adaptation for misinformation detection report substantial drops under distribution shift (Shu et al., 2022; Huang et al., 2021). These findings align with broader observations in NLP that benchmark performance does not necessarily imply robustness under temporal, topical, or source shift.

Despite these concerns, many papers reporting near-ceiling accuracy rely on random within-dataset splits without source-held-out or cross-dataset evaluation. As a result, systematic evidence about transferability remains limited relative to the volume of high-accuracy claims. Our work aims to address this gap by directly quantifying cross-dataset degradation across ISOT, Misinfo, and WELFake.

3. Dataset Semantics and Structural Homogeneity

3.1. ISOT

The ISOT Fake News Dataset (Ahmed et al., 2017) consists of two files: *True.csv* (21,417 articles) and *Fake.csv* (23,481 articles), primarily covering 2016–2017 political and world news. True articles are largely sourced from Reuters, while fake articles are collected from outlets flagged as unreliable.

This provenance-based construction introduces strong structural regularities. The true class reflects wire-service style and formatting, whereas the fake class aggregates politically charged content from heterogeneous sources. As a result, classification can rely on stylistic and source cues rather than claim-level reasoning.

3.2. Misinfo (79k)

The Misinfo dataset contains 78,617 articles (34,975 true; 43,642 fake/misinformation/propaganda). True articles originate from mainstream outlets such as Reuters, The New York Times, and The Washington Post, while the negative class aggregates diverse sources, including extremist websites and curated disinformation cases.

Although this mixture increases semantic heterogeneity, structural clustering remains: different source ecosystems exhibit consistent stylistic and rhetorical patterns. Models may therefore still exploit these cues instead of factual inconsistencies.

3.3. WELFake

WELFake contains 72,134 articles (35,028 real; 37,106 fake), constructed by merging multiple datasets (Kaggle, McIntire, Reuters, BuzzFeed) to increase diversity (Verma et al., 2021).

While this reduces reliance on a single source, binary provenance signals persist across merged corpora. Moreover, substantial textual overlap with ISOT is observed due to shared sources. This challenges dataset independence and motivates explicit deduplication in cross-dataset evaluation.

3.4. Structural Implications

Across ISOT, Misinfo, and WELFake, class labels correlate strongly with source provenance and topic distributions. Such correlations create conditions conducive to shortcut learning (Geirhos et al., 2020), where models exploit stable stylistic or ecosystem-level cues. High within-dataset accuracy under random splits may therefore reflect structural homogeneity rather than transferable epistemic reasoning.

4. Intra-corpus classification and shortcut analysis

We first evaluated intra-corpus classification on the ISOT dataset under a standard train/validation split (70/15/15, stratified by label). Models were trained and validated exclusively on ISOT, without cross-domain mixing. Performance was measured using accuracy, macro F1, and weighted F1.

We focus on ISOT as a maximal case of structural separability, where class labels are strongly aligned with source-specific lexical and stylistic cues. This makes it particularly suitable for analyzing the extent to which near-ceiling performance can be explained by shallow signals rather than semantic understanding.

4.1. Transformer-based models

Across configurations, transformer models achieved near-perfect validation performance:

- Accuracy: 0.99+
- Macro F1: 0.99+
- Weighted F1: 0.99+

These results are consistent with prior literature reporting $> 98\%$ accuracy on ISOT. However, such performance alone does not establish robust facticity discrimination, as intra-corpus evaluation may permit learning of dataset-specific regularities.

4.2. Shallow stylistic modeling

To assess whether intra-corpus performance on ISOT requires deep contextual representations, we trained a logistic regression classifier using only low-dimensional stylistic features extracted from raw text. No lexical n-grams or embeddings were used.

Feature extraction. For each document, we computed a set of normalized surface features designed to capture stylistic and provenance-related regularities. These included binary indicators such as `has_reuters` (presence of the token "(Reuters)") and `has_video` (presence of markers such as "[Video]" or "video"), as well as continuous features such as `upper_ratio` (proportion of uppercase letters over alphabetic characters), `qm_per_100` and `excl_per_100` (question and exclamation marks per 100 characters), `log_chars`, and `log_words`. Additional cues included punctuation density, quotation density, source-name matches, and counts of emphatic uppercase tokens. All continuous features were standardized before training. The classifier used L_2-regularized logistic regression with default scikit-learn settings.

Model	Accuracy	Macro F1
Logistic (features)	0.9960	0.9960
Logistic (w/o Reuters)	0.8935	0.8935

Table 1: Intra-corpus validation performance on ISOT using shallow stylistic features.

Using the full feature set, performance matches transformer-based models within rounding error, indicating that ISOT is nearly linearly separable in a low-dimensional stylistic space. To test whether this result is dominated by a single provenance marker, we performed an ablation in which the token "(Reuters)" was removed from the text before feature extraction, and the derived features `has_reuters` and `has_dateline` were excluded from the model. Even after this removal, the

classifier still achieved 0.8935 accuracy and 0.8935 Macro F1 on validation, showing that Reuters is the strongest cue but not the only source of separability.

Feature importance. Table 2 shows the largest-magnitude coefficients from the full logistic regression model (positive values predict *true*, negative values predict *fake*). As expected, `has_reuters` is by far the dominant predictor.

Feature	Coefficient
`has_reuters`	+6.36
`log_words`	+1.38
`upper_ratio`	-0.38
`has_video`	-1.49
`qm_per_100`	-1.37
`log_chars`	-1.59

Table 2: Selected coefficients from the full logistic regression model on ISOT. Positive values favor the *true* class; negative values favor the *fake* class.

Table 3 summarizes the most influential coefficients after removing Reuters-related cues. The remaining decision boundary is still driven by simple stylistic properties, especially document length, uppercase usage, question-mark density, exclamation frequency, and video markers.

Feature	Coefficient
Most predictive of fake	
`log_words`	-5.90
`upper_ratio`	-4.47
`qm_per_100`	-2.66
`excl_per_100`	-1.32
`has_video`	-1.17
Most predictive of true	
`log_chars`	+4.00
`quote_ratio`	+1.84
`n_chars`	+0.74
`punct_per_100`	+0.28

Table 3: Selected coefficients from the logistic regression model after removing Reuters-related cues. Positive values favor the *true* class; negative values favor the *fake* class.

These results show that the near-ceiling separability of ISOT is heavily driven by provenance markers, but also supported by additional shallow stylistic differences between classes. In other words, removing Reuters substantially weakens the shortcut, yet intra-corpus classification remains far above chance without requiring semantic representations.

4.3. Ablation and minimal baselines

To further analyze the source of near-ceiling performance in ISOT, we evaluated minimal classifiers based on shallow stylistic cues.

Building on the shallow stylistic model above, we further evaluated a one-rule classifier based on a single binary feature:

```
if has_reuters == 1 → TRUE
        else → FAKE
```

This classifier achieved 0.9963 accuracy on ISOT validation. Table 4 summarizes the intra-corpus performance of transformer-based models and minimal baselines. A single binary lexical cue is therefore sufficient to reproduce near-ceiling performance on ISOT.

Model	Acc.	M. F1	W. F1
BERT (baseline)	0.999+	0.999+	0.999+
Logistic (style features)	0.9960	0.9960	0.9960
1-rule (has_reuters)	0.9963	0.9963	0.9963

Table 4: Intra-corpus results on ISOT.

4.4. Temporal considerations.

The datasets used in this study span a relatively narrow and overlapping time period, primarily covering news events from 2016–2017 in the case of ISOT, with partially overlapping periods in Misinfo and WELFake due to shared or reused sources. Following common practice in the literature, we employed stratified random splits rather than temporal splits.

We acknowledge that the absence of temporal partitioning may allow models to be exposed to similar or related events across training and validation sets. Prior work has highlighted the importance of temporal separation in fact-checking and misinformation detection, arguing that models can otherwise exploit event-level overlap and lack of counter-evidence (Glockner et al., 2022).

However, the ablation results presented in this section indicate that near-ceiling performance can be achieved using shallow stylistic cues alone, including a single provenance marker. This suggests that high intra-corpus performance is not primarily driven by temporal memorization of events, but by structural regularities and source-specific artifacts embedded in the datasets.

Therefore, while temporal splits are important for evaluating robustness in realistic fact-checking scenarios, the evidence presented here indicates that the dominant factor behind near-perfect intra-corpus results in ISOT is dataset-level separability rather than event-level generalization.

5. Cross-Dataset Experiments

We evaluate several BERT-based configurations under cross-dataset transfer conditions. Models are trained on one dataset and evaluated on another without target fine-tuning.

To evaluate robustness under dataset shift, we conducted systematic cross-dataset transfer experiments using three widely adopted fake-news corpora: WELFake, Misinfo, and ISOT.

To assess whether affective enrichment improves cross-domain generalization, we additionally evaluate several BERT-based configurations, including variants augmented with VAD (Mohammad, 2018), Ekman emotion categories (Ekman, 1992), and SenticNet features (Cambria et al., 2018), which provide complementary affective signals at lexical and conceptual levels.

Affective feature integration. VAD features encode continuous affective dimensions (valence, arousal, dominance) at the lexical level. Ekman categories provide a discrete representation of basic emotions, capturing coarse affective states such as anger, fear, or joy. SenticNet supplies concept-level affective and semantic features derived from commonsense knowledge bases.

In all cases, these signals are computed from the input text and concatenated with the BERT representation prior to classification.

For each pair of datasets (A, B), we trained on A and evaluated on B, yielding six transfer directions:

1. WELFake $\rightarrow$ Misinfo
2. Misinfo $\rightarrow$ WELFake
3. ISOT $\rightarrow$ WELFake
4. WELFake $\rightarrow$ ISOT
5. ISOT $\rightarrow$ Misinfo
6. Misinfo $\rightarrow$ ISOT

All models were fine-tuned under identical conditions (BERT-base backbone, max length 128, batch size 128, three epochs, class weighting, mixed precision training). We report Accuracy and Macro F1 on the external test set in each direction. Macro F1 is considered as it captures class balance under distributional shift.

Full per-direction classification reports (including class-wise precision, recall, and F1 scores) are provided in the Appendix (Tables 8–13).

Table 5 summarizes Accuracy and Macro F1 for compact comparison across all six transfer directions. Across domains, performance drops relative to within-dataset evaluation, indicating limited transferability.

5.1. Duplicate Analysis and Data Leakage Control

Before conducting cross-dataset evaluation, we performed exact-match duplicate detection between corpora. Between WELFake (train) and ISOT (test), we identified 39,687 exact duplicates out of 44,898 ISOT instances (approximately 88.4%). To prevent leakage, all overlapping instances were removed prior to training.

This finding indicates that WELFake partially subsumes ISOT content, rendering naïve cross-dataset evaluation misleading unless strict deduplication is applied. The persistence of near-ceiling WELFake→ISOT performance after deduplication further suggests that ISOT remains structurally easy as a target domain.

6. Discussion

High intra-dataset accuracy does not imply generalization under distribution shift. When datasets are structurally homogeneous, classification approximates ecosystem discrimination.

Binary provenance-based labels risk conflating source identification with epistemic inference.

6.1. Cross-Domain Stability Analysis

To evaluate robustness under dataset shift, we computed the Macro F1 score across all six cross-dataset transfer directions.

For each model, we report the mean Macro F1 across directions and its standard deviation. Lower variance indicates higher cross-domain stability.

To quantify directional asymmetry, we compute the absolute Macro F1 difference between both transfer directions of each dataset pair (e.g., W→M vs M→W). Lower values indicate greater bidirectional consistency and thus stronger cross-domain robustness. Table 7 reports the directional gaps for the three dataset pairs, as well as the mean gap per model.

Overall, BERT+VAD+Ekman exhibits the lowest average directional gap, indicating the most symmetric transfer behaviour. Ekman alone ranks second. In contrast, SenticNet and VAD show the largest asymmetries, particularly in the ISOT–Misinfo pair.

6.2. Ranking of Cross-Domain Stability

Ranking models by increasing standard deviation (see Table 6) yields:

1. BERT + VAD + Ekman (most stable)
2. BERT + Ekman
3. BERT simple
4. BERT + SenticNet
5. BERT + VAD (least stable)

Model	W→M		M→W		I→W		W→I		I→M		M→I	
	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1
BERT	0.851	0.846	0.914	0.913	0.794	0.789	0.996	0.996	0.638	0.548	0.965	0.964
VAD	0.854	0.846	0.695	0.659	**0.812**	**0.810**	0.997	0.997	0.576	0.424	0.879	0.875
Ekman	0.846	0.842	0.952	0.952	0.745	0.733	**0.999**	**0.998**	**0.719**	**0.685**	0.669	0.614
VAD+Ekman	**0.863**	**0.858**	0.849	0.845	0.803	0.800	0.998	0.997	0.686	0.641	0.828	0.819
SenticNet	0.858	0.854	**0.958**	**0.958**	0.696	0.674	0.997	0.997	0.636	0.543	**0.987**	**0.987**

Table 5: Accuracy (Acc) and Macro F1 (MF1) across all six cross-dataset transfer directions. Best value per column is shown in bold.

Model	Std. Dev. (Macro F1)
BERT simple	0.1630
BERT + VAD	0.2012
BERT + Ekman	0.1526
BERT + VAD + Ekman	**0.1147**
BERT + SenticNet	0.1875

Table 6: Sample standard deviation of Macro F1 across six transfer directions. Lower values indicate greater cross-domain stability.

Model BERT+	W↔M	W↔I	I↔M	M.Gap
Base	0.0671	0.2073	0.4159	0.2301
VAD	0.1871	**0.1869**	0.4516	0.2752
Ekman	0.1104	0.2655	**0.0702**	0.1487
VAD+Ekman	**0.0138**	0.1973	0.1787	**0.1299**
SenticNet	0.1036	0.3231	0.4434	0.2900

Table 7: Directional Macro F1 gap (absolute difference between bidirectional transfers). Lower values indicate stronger symmetry and cross-domain robustness.

Several structural patterns emerge from the analysis.

First, raw in-dataset performance does not predict cross-domain robustness. Models achieving near-ceiling accuracy on certain datasets exhibit substantial degradation when evaluated under domain shift.

Second, affective enrichment behaves differently depending on the representation used. Ekman-based features consistently reduce performance variance across datasets. When combined with VAD, the resulting model achieves the lowest cross-domain fluctuation (Std. Dev. = 0.113), suggesting a regularization effect that stabilizes decision boundaries across stylistic and source-based shifts.

In contrast, SenticNet produces the highest peaks in some transfer directions (e.g., Misinfo → ISOT and WELFake → ISOT) but also sharp collapses in others (e.g., ISOT → Misinfo). This indicates high sensitivity to training-domain lexical-affective alignment.

VAD alone exhibits the highest instability (Std. Dev. = 0.207), suggesting that coarse affective dimensions without discrete emotional categories may insufficiently constrain cross-domain generalization.

Finally, the consistent asymmetry between directions (e.g., ISOT → Misinfo vs. Misinfo → ISOT) confirms that dataset provenance and stylistic priors strongly influence transferability. Training on more heterogeneous corpora appears to improve outward generalization, whereas training on more homogeneous datasets leads to brittle representations.

These observations align with the hypothesis that dataset structure plays a dominant role in model behavior. In particular, ISOT appears to be largely separable using source-specific lexical cues, such as the presence of the token "(Reuters)". This explains why both shallow and deep models achieve near-perfect intra-corpus performance, while cross-dataset generalization collapses. The phenomenon is consistent with shortcut learning and dataset bias, but here it emerges under minimal feature assumptions.

Taken together, these results indicate that cross-domain evaluation and variance analysis should complement standard accuracy reporting in fake-news classification. Stability under dataset inversion provides a stronger robustness criterion than isolated peak performance.

7. Conclusions

Fake-news benchmarks frequently report near-ceiling performance, yet cross-dataset evaluation reveals substantial and asymmetric generalization gaps. These results show that accuracy on isolated benchmarks is not a reliable indicator of robustness, but largely reflects dataset separability rather than genuine veracity understanding.

Across six cross-dataset transfer directions involving WELFake, Misinfo, and ISOT, we observe three systematic phenomena. First, average performance and stability are not equivalent objectives: in some cases, BERT+SenticNet achieves the highest mean Macro F1, yet exhibits considerable directional asymmetry and variance. Second, BERT+SenticNet attains near-ceiling results in cer-

tain directions but collapses under domain reversal, resulting in the highest mean directional gap. This pattern suggests strong sensitivity to corpus-specific stylistic and lexical distributions. Third, the combined VAD+Ekman representation yields the lowest standard deviation and the smallest bidirectional gap, indicating the most stable cross-domain behavior.

The superior robustness of VAD+Ekman may stem not only from feature concatenation, but from the complementarity of affective representations. VAD encodes continuous affective gradients (valence, arousal, dominance), providing smooth global signals, whereas Ekman categories introduce discrete emotional anchors. Their joint integration may provide an implicit regularizing effect: when distribution shift destabilizes one affective dimension, the other may preserve transferable structure. This hybrid affective embedding reduces directional asymmetry and mitigates collapse under domain shift.

These results suggest that robustness emerges not from maximizing in-domain accuracy, but from constraining models with semantically meaningful inductive biases. Emotional structure, when represented through complementary continuous and categorical spaces, appears to provide such a constraint. Importantly, robustness must be evaluated not only via mean performance, but through variance and directional gap metrics that explicitly capture asymmetry under transfer.

Progress in fake-news detection therefore requires a methodological shift: from accuracy maximization on isolated benchmarks toward cross-domain robustness, bidirectional evaluation, and construct-valid modeling. Without such a shift, near-ceiling scores risk reflecting dataset artifacts rather than genuine generalizable understanding.

Acknowledgements

The authors wish to thank the anonymous reviewers for their useful discussion, and Instituto Politécnico Nacional (COFAA, SIP-IPN) and the Mexican Government (SECIHTI, SNII) for their financial support for this work.

A. Full Cross-Domain Classification Results

This appendix reports the complete cross-domain evaluation results for all six transfer directions. For each experiment, we provide Accuracy (Acc), Macro F1 (MF1), Weighted F1 (WF1), per-class F1 scores, and per-class Recall. The best value in each column is highlighted in bold.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138. Springer.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *Proceedings of AAAI Conference on Artificial Intelligence*, 32(1).

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Ehab Essa, Karima Omar, and Ali Alqahtani. 2023. Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems*, 9(6):6581–6592.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673.

Axel Gelfert. 2018. Fake news: A definition. *Informal Logic*, 38(1):84–117. Accessed 2026-02-25.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinqiu Huang, Min Gao, Jia Wang, and Kai Shu. 2021. Dafd: Domain adaptation framework for fake news detection. In *International conference on neural information processing*, pages 305–316. Springer.

Md. Rafiqul Islam, M. M. Islam, Md. Shafiqul Azad, M. S. Uddin, Kamal Daud, and M. A. Hossain. 2020. A survey on fake news detection using machine learning and deep learning techniques. *IEEE Access*, 8:178007–178025.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80:11765–11788.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.8509	0.8460	0.8489	0.8733	0.8188	0.9303	0.7528
VAD	0.8536	0.8462	0.8498	0.8799	0.8126	**0.9702**	0.7095
Ekman	0.8463	0.8419	0.8447	0.8683	0.8156	0.9167	0.7594
VAD+Ekman	**0.8625**	**0.8583**	**0.8608**	**0.8827**	**0.8338**	0.9363	0.7713
SenticNet	0.8578	0.8539	0.8564	0.8777	0.8301	0.9237	**0.7765**

Table 8: Full classification results for the transfer experiment WELFake → Misinfo. Best values per column are shown in bold.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.9140	0.9131	0.9133	0.9221	0.9041	0.9951	0.8293
VAD	0.6949	0.6591	0.6616	0.7695	0.5487	0.9967	0.3794
Ekman	0.9524	0.9523	0.9524	0.9547	0.9500	0.9800	0.9236
VAD+Ekman	0.8489	0.8445	0.8451	0.8708	0.8182	0.9959	0.6953
SenticNet	**0.9577**	**0.9575**	**0.9576**	**0.9602**	**0.9549**	**0.9981**	**0.9154**

Table 9: Full classification results for the transfer experiment Misinfo → WELFake. Best values per column are shown in bold.

Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumya Sinha. 2020. Fndnet – a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44.

Sahas Koka, Anthony Vuong, and Anish Kataria. 2024. Evaluating the efficacy of large language models in detecting fake news: a comparative analysis. *arXiv preprint arXiv:2406.06584*.

Maaruf Lawal and Abdulrashid Abdulrauf. 2025. Fake news detection using Bi-LSTM architecture: A deep learning approach on the isot dataset. *Journal of Computing Theories and Applications*, 3:104–114.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1146–1151. Accessed 2026-02-25.

Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2024. Large language model agent for fake news detection.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 174–184.

Steven Peutz. 2023. Misinformation & fake news text dataset 79k. Dataset card describing 34,975 true and 43,642 fake/misinfo/propaganda items. Accessed 2026-02-25.

Maria Pszona, Maria Janicka, Grzegorz Wojdyga, and Aleksander Wawer. 2023. Towards universal methods for fake news detection. *Natural Language Engineering*, 29(4):1004–1042.

Kai Shu, Ahmadreza Mosallanezhad, and Huan Liu. 2022. Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in fake media generation and detection*, pages 215–232. Springer.

Edson C. Tandoc, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, 6(2):137–153.

Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. Welfake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893. Open full text circulated by authors; Accessed 2026-02-25.

H. F. Villela, F. Corrêa, J. S. D. A. N. Ribeiro, A. Rabelo, and D. B. F. Carvalho. 2023. Fake news detection: A systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, 14(1):47–58.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Report prepared for the Council of Europe. Accessed 2026-02-25.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.7935	0.7887	0.7877	0.8206	0.7568	0.9734	0.6240
VAD	**0.8115**	**0.8101**	**0.8096**	**0.8263**	**0.7939**	0.9244	**0.7051**
Ekman	0.7446	0.7329	0.7312	0.7888	0.6769	0.9831	0.5198
VAD+Ekman	0.8034	0.8001	0.7993	0.8257	0.7744	0.9603	0.6555
SenticNet	0.6961	0.6735	0.6710	0.7594	0.5877	**0.9884**	0.4207

Table 10: Full classification results for the transfer experiment ISOT → WELFake. Best values per column are shown in bold.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.9962	0.9960	0.9962	0.9951	0.9968	0.9903	**1.0000**
VAD	0.9971	0.9970	0.9971	0.9964	0.9976	0.9927	**1.0000**
Ekman	**0.9985**	**0.9984**	**0.9985**	**0.9981**	**0.9987**	**0.9961**	**1.0000**
VAD+Ekman	0.9975	0.9974	0.9975	0.9968	0.9979	0.9937	**1.0000**
SenticNet	0.9967	0.9966	0.9967	0.9959	0.9973	0.9918	**1.0000**

Table 11: Full classification results for the transfer experiment WELFake → ISOT. Best values per column are shown in bold.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.6384	0.5484	0.5707	0.7500	0.3469	0.9766	0.2160
VAD	0.5764	0.4235	0.4563	0.7204	0.1265	**0.9827**	0.0690
Ekman	**0.7188**	**0.6846**	**0.6961**	**0.7884**	**0.5809**	0.9434	**0.4382**
VAD+Ekman	0.6857	0.6407	0.6548	0.7678	0.5136	0.9359	0.3732
SenticNet	0.6356	0.5431	0.5658	0.7487	0.3374	0.9776	0.2086

Table 12: Full classification results for the transfer experiment ISOT → Misinfo. Best values per column are shown in bold.

Model	Acc	MF1	WF1	F1 Fake	F1 True	R Fake	R True
BERT	0.9645	0.9643	0.9644	0.9670	0.9615	0.9965	0.9294
VAD	0.8786	0.8751	0.8761	0.8958	0.8544	0.9985	0.7471
Ekman	0.6688	0.6144	0.6211	0.7592	0.4696	0.9983	0.3074
VAD+Ekman	0.8277	0.8194	0.8212	0.8582	0.7806	0.9967	0.6425
SenticNet	**0.9866**	**0.9865**	**0.9866**	**0.9873**	**0.9858**	**0.9989**	**0.9731**

Table 13: Full classification results for the transfer experiment Misinfo → ISOT. Best values per column are shown in bold.

Benchmarking Check-Worthiness Models on LLM Generated Claims

Charlie Roadhouse, Matthew Shardlow, Ashley Williams
Manchester Metropolitan University
Ormond Building, Lower Ormond Street, Manchester, M15 6BX
charlie.roadhouse@stu.mmu.ac.uk, M.Shardlow@mmu.ac.uk, Ashley.Williams@mmu.ac.uk

Abstract

The proliferation of large language models (LLMs) has significantly increased the potential for automated dissemination of disinformation, necessitating robust systems for check-worthiness detection. However, existing models are primarily trained on human claims, leaving their performance on machine-generated text largely unexplored. In this paper, we benchmark encoder models (BERT and RoBERTa) and industry accessible tools (ClaimBuster) against LLM-paraphrased claims across three stylistic categories: syntactic restructuring, syntactic complexity and lexical informality. Our results indicate a consistent performance degradation on synthetic claims, particularly on complex and informal claims. We demonstrate that adversarial training significantly improves model resilience, with RoBERTa achieving F1-score gains up to +5.22 on the CheckIt dataset. Finally, SHAP analysis reveals that while base models rely on narrow syntactic heuristics such as active voice, robust models learn to anchor their prediction on core factual entities. These findings highlight the necessity of stylistic-aware training to maintain fact-checking efficacy in an increasingly LLM-populated information landscape.

Keywords: Fact-Checking, Check-Worthiness Detection, Adversarial Training, Disinformation

1. Introduction

The rise of generative AI methods utilising large language models (LLMs) to produce text has become widespread. The increase in LLM agency has led to their misappropriation for malicious purposes, where models are prompted to generate harmful or misleading content (Pan et al., 2023). Malicious actors leverage this generated text as disinformation, spreading it across online platforms. This proliferation of disinformation has a direct negative real-world impact across nations and socio-political topics (Aïmeur et al., 2023).

This threat has spurred research into the detection and mitigation of disinformation (Su et al., 2020). While these efforts show promise, they are yet to show their efficacy in real-world situations. Currently, fact-checking remains the most effective strategy for mitigating disinformation (Graves, 2017). The fact-checking process begins by identifying and assessing whether a claim is check-worthy, followed by evidence gathering to provide a final justification (Das et al., 2023). Due to the sheer volume of claims online, journalists can no longer manually incorporate this fact-checking process into their workflows, necessitating the work of dedicated fact-checking organisations such as PolitiFact[1], FullFact[2] & FactCheck.org[3].

While effective, these organisations face significant scalability and latency issues. Consequently, research has pivoted toward automated check-worthiness detection (Guo et al., 2022). These systems leverage AI models to automatically prioritise claims for review, removing the bottleneck of manual selection and significantly scaling number of claims processed (Graves, 2018). Such systems are now deployed in production workflows such as FullFact's internal AI suite[4] or ClaimBuster which is accessible through their API (Hassan et al., 2017)

The problem with current methods comes from their reliance on human annotated training data, often exhibiting low generalisation across topics, which can lead to claims incorrectly being identified within emerging events (Nenno, 2024). This reliance makes them susceptible to LLM-generated claims, which may contain linguistic nuances and structural discrepancies that differ from human-written claims. To understand the impact of LLM-generated content on model robustness, we propose a systematic study comparing state-of-the-art check-worthiness strategies against both human-authored and LLM-paraphrased claims. We then incorporate the LLM-generated content into the training of a separate model to adversarially train a model. All experiments are carried out on English claims only. We address the following research questions (RQ):

- **RQ1** - To what extent does a check-worthiness classification models' performance change when assessing LLM generated claims?
- **RQ2** - Does adversarial training using LLM generated claims improve the models performance on human written and LLM written claims?

[1] https://www.politifact.com/
[2] https://fullfact.org/
[3] https://www.factcheck.org/
[4] https://fullfact.org/ai/

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 34–44
12 May 2026.

- **RQ3** - Do models learn different linguistic features between the human and LLM generated claims?

2. Background

The impact of fake news is well documented to have influenced major socio-political events (Henkel, 2021; Chen et al., 2021; Cazzamatta, 2025). Given that fake news spans diverse topics, countries and modalities, modern research necessitates a more fine-grained taxonomy. Wardle and Derakhshan (2017) categorise information disorder into three sub-categories based on intent and truthfulness: Misinformation, Disinformation and Malinformation. Our study is grounded in this framework, specifically addressing the threat of automated disinformation scaling. By evaluating how models respond to stylistic circumvention, we provide insights into the resilience of check-worthiness systems against intentionally machine generation disinformation.

This categorisation allows researchers to better identify the motivations of authors and the susceptibilities of target audience (Bragazzi and Garbarino, 2024). While these categories provide a clear theoretical framework, there is a significant difference, with related natural language processing (NLP) tasks such as rumour detection (Bondielli and Marcelloni, 2019), propaganda detection (Da San Martino et al., 2020), credibility assessment (Srba et al., 2025) and fact-checking (Guo et al., 2022).

In the NLP community, early misinformation detection focused on machine learning models leveraging hand-crafted linguistic features for classification, as outlined in the survey by Su et al. (2020), which cover systems and datasets from 2000 - 2018. However, the introduction of transformer-based architectures, particularly BERT (Devlin et al., 2019), shifted the state-of-the-art toward encoder-based strategies that benefit from task-specific fine-tuning. Beyond textual content, recent research has also explored network topology and propagation patterns to identify misleading content (Zhou and Zafarani, 2019).

Despite these advancements, manual fact-checking remains the gold-standard for mitigation. To address the scalability issues of human fact-checkers, automated check-worthiness systems are employed to triage claims. In NLP, this is primarily framed as either a classification task (assigning binary or ordinal label) or a ranking task (prioritising claims based on perceived check-worthiness) (Guo et al., 2022).

Current state-of-the-art check-worthiness models leverage large-scale transformer encoders. For example, systems like ClaimBuster (Hassan et al., 2017) evolved from Support Vector Machines (SVM) to BERT-based architectures. These models often incorporate auxiliary information, such as named entity recognition (NER) and user metadata, to bolster performance (Ahmed et al., 2021).

However, encoder-based models often struggle with domain-shifting and emerging topics not present in the training data (Nenno, 2024). While LLMs have demonstrated impressive zero-shot capabilities in check-worthiness detection, issues such as hallucination and inconsistent reasoning mean that fine-tuned encoders remain the preference (Majer and Šnajder, 2024). Wright and Augenstein (2020) demonstrated that LLM-based paraphrasing is highly effective for increasing training data volume. Yet, a significant gap remains; while these studies use LLMs to support model training, they do not sufficiently investigate model performance in adversarial settings where LLMs are used to generate deceptive variations. This research aims to bridge the gap between LLM-augmented training and model robustness against synthetic stylistic shifts.

3. Methodology

3.1. Dataset Selection and Pre-Processing

To investigate the robustness of check-worthiness models against both human-authored and LLM-generated claims, two check-worthiness datasets are selected:

- **CheckIt (Sundriyal et al., 2026)**: A fine-grained assessment of claims from twitter. While annotating for a final check-worthiness label, data was also annotated on fine-grained labels. These comprised of if the claim is verifiable, publicly interesting, potentially harmful, misleading, of interest to the government and whether a fact-checker should check the claim.

- **CheckThatLab! 2024 (Hasanain et al., 2024)**: Made up event specific topics including COVID-19, vaccines, Gazza/Palestine conflict and climate change. Spanning both tweets, debates and speeches, framed as a binary classification task.

All code, models and data are released for reproducibility purposes[5].

3.2. Experimental Setup

We train two tracks of model: a base model (human only claims) and a robust model using adversarial training. We frame this as adversarial because the

[5]https://github.com/chroadhouse/Benchmarking-Check-Worthiness-Models-on-LLM-Generated-Claims

LLM-paraphrased variants act as stylistic perturbations designed to break the models resilience on surface level heuristics while keeping the semantic ground truth.

For our base and robust models, we select two transformer-based encoders. We also use a zero-shot LLM and publicly deployed state-of-the-art check-worthiness system for benchmarking. The models are outlined:

- **BERT** (Devlin et al., 2019): A standard baseline encoder widely utilised in prior check-worthiness shared tasks (Alam et al., 2023).
- **RoBERTa** (Liu et al., 2019): A robustly optimised variant of BERT. Its pre-training objective may offer enhanced resilience against stylistic variations in synthetic text.
- **Llama 3.1:8B**[6]: Employed as a zero-shot baseline to assess the check-worthiness detection capabilities of current general-purpose LLMs. Our prompt for classification is shown in Figure 1.
- **ClaimBuster** (Hassan et al., 2017): A leading industry-standard system providing a baseline for live fact-checking API[7] performance, leveraging a BERT-based architecture.

System Prompt
ROLE: You are a professional Fact-Checking Editor.
TASK: Analyse the following text and decide if it is **CHECK-WORTHY**.

CRITERIA FOR YES:
1. It contains a **verifiable factual claim** (dates, numbers, events, causal relationships).
2. It is of **public interest or potential harm** (not just a personal opinion or vague complaint).
3. It is **NOT standard health advice** (e.g., 'Wash your hands' is NO).

Input Format: `Text: {claim}`

Output Instruction: You must start your response with the decision tag.
Format: `<decision>YES</decision>` or `<decision>NO</decision>`

Figure 1: System prompt for zero-shot check-worthiness classification.

For the training, both encoder base models are trained on the CTL 24 training data, selected for cross domain coverage across social media, speeches and debates (Hasanain et al., 2024). The encoder models were trained for 10 epochs or until convergence, with an early stopping patience of 3. The models were optimised using AdamW (Loshchilov and Hutter, 2017) with a weight decay of 0.01 and a learning rate of 2×10^{-5}. We use the same parameters but also include LLM-paraphrased variants of the claims for the adversarial training. When classifying, we use a greedy decoding strategy for the Llama 3.1:8B model.

[6] https://huggingface.co/meta-llama/Llama-3.1-8B

[7] http://idir.uta.edu/claimbuster

Base Paraphrase Prompt (Syntactic Restructuring)

You are a precise fact-checking rewriter. Paraphrase the text below by changing the sentence structure (e.g., active to passive voice) while keeping all names, dates, and technical terms EXACTLY as written. Do not add any commentary.
Input: `Text: {text} <output>`

Complex Style Prompt (Syntactic Density)

You are a sophisticated academic editor. Rewrite the text below into a high-complexity, formal sentence using subordinate clauses and nominalization. You must retain all names, dates, and technical terms EXACTLY as written. Do not add any commentary.
Input: `Text: {text} <output>`

Informal Style Prompt (Lexical Informality)

You are a social media user. Rewrite the text below to sound like a natural, conversational post using contractions and informal phrasing. You must keep all specific names, dates, and statistics EXACTLY as written. Do not use hashtags or emojis. Do not add any commentary.
Input: `Text: {text} <output>`

Figure 2: Design of the LLM prompt templates used for generating the three stylistic paraphrases (Syntactic Restructuring, Syntactic Density, and Lexical Informality).

3.3. Generation of LLM Content

We leverage Llama 3.1:8B-Instruct to generate paraphrased candidates for the CheckThatLab 2024 and CheckIt datasets, using an auto regressive strategy with nucleus sampling (Top-P) set to 0.9, with a temperature of 0.8 and a top-k of 40. To ensure linguistic variety preventing structural loops, a repetition of 1.1 is applied.

We experiment with three zero-shot prompting strategies (see Figure 2): Standard, Informal and

Complex. For each variant, the model is strictly instructed to maintain all named entities to minimise factual hallucination. Table 1 illustrates an original human-authored claim alongside its three corresponding synthetic variants, showing the linguistic differences while ensuring entities are not replaced.

Ensuring label preservation is critical when generating synthetic data; a paraphrase that alters the factual content of the claim would invalidate the original check-worthiness labels. Consequently, we assess the linguistic and semantic drift between the human reference and LLM-generated candidates using a suite of metrics. Given that factual integrity is more vital in this domain than absolute semantic overlap (Wright and Augenstein, 2020), our validation framework prioritises entity preservation and semantic stability.

To extract the necessary features for these metrics we utilise a RoBERTa model fine-tuned for NER on the TweetNER7 dataset (Antypas et al., 2023) and a second RoBERTa model fine-tuned on the Tweet-Eval benchmark for sentiment analysis (Barbieri et al., 2020). The performance of the generation process is then quantified using the following metrics, with results summarised in Table 2.

- **BLEURT** (Sellam et al., 2020): An evaluation metric that leverages an encoder model trained on perturbed sentence pairs and then trained on multiple semantic preservation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) & BERTScore (Zhang et al., 2019), all of which are popular N-gram overlap methods.
- **Mutual Implication Score (MIS)** (Babakov et al., 2022): This method is developed specifically to understand the similarity between a reference piece of text and a paraphrased candidate. It leverages a natural language inference model, treating the paraphrasing as bidirectional entailment.
- **Cosine Similarity of Entity Distributions**: We measure the preservation of factual subjects by comparing the frequency of entity groups (e.g., persons, organisations, locations) between original and paraphrased claims. We identify named entities and represent each claim by a vector (category count). By calculating the cosine similarity (Büyüktopaç and Acarman, 2019) between these vectors, we quantify how well the LLM maintains the original subjects, ensuring stylistic shifts do not alter the factual composition of the claim.
- **Jaccard Similarity** (Niwattanakul et al., 2013): The Jaccard similarity is used to measure the difference in the extracted named entities. This is crucial to ensure that the same names, dates and locations are mentioned, even with a change of phrasing.
- **Sentiment** Δ: The sentiment score for both the reference and candidate claim are extracted and then the difference is calculated to see the sentiment drift between the two claims.
- **Length** Δ: The difference in characters between the original reference and candidate claim.

The results in Table 2 demonstrate high entity group consistency between paraphrased and human claims. While the lower BLEURT and MIS scores suggest significant stylistic and structural divergence, the stability of the entity metrics confirms that the factual subjects remain intact during the transformation.

To ensure that the LLM did not hallucinate in the generation of paraphrased claims, we randomly select 50 samples from each of the three prompts from the CTL 2024 dataset and assess semantic similarity between the human and LLM-generated pair. The primary investigator annotated for semantic preservation using a binary label and then used a predefined confidence definition by Mu et al. (2023) to rate the confidence of the decision between 1-5. The results of the preservation annotation are presented in Table 3. They show a high semantic preservation with a mean confidence over 4.

4. Results

Table 4 presents the performance of the fine-tuned transformer models, ClaimBuster and zero-shot LLM. Performance is evaluated across three benchmark datasets using weighted Precision (P), Recall (R) and F1-score. The results compare the efficacy on original human-authored claims against their LLM-paraphrased counterparts.

The experimental results indicate that fine-tuned transformer models outperform ClaimBuster and zero-shot LLM across both the CheckIt and CTL 2024 data splits. All four models experienced a performance degradation when evaluated on the LLM paraphrased claims, specifically in the Complex and Informal settings. Notably, while the ClaimBuster model remained competitive on the CTL 2024 test split, the model struggled on the CheckIt dataset.

4.1. Robust Model Performance

The results of the adversarial training are detailed in Table 5. Where the difference between the original and robust encoder models are calculated for each dataset split and claim type. Table 5 highlights a significant performance boost for the robust

Claim Type	Claim Text	Linguistic Changes
Human (Original)	"We've run up more debt in the last eight years than under all the presidents from George Washington to Jimmy Carter combined."	*Baseline:* Natural, conversational political speech.
Simple Paraphrase	"More debt has been run up in the last eight years than under all the presidents from George Washington to Jimmy Carter combined."	**Syntactic Shift:** Conversion from active to passive voice; maintains original vocabulary.
Informal Style	"We've racked up more debt in the last eight years than all the presidents from George Washington to Jimmy Carter combined."	**Lexical Shift:** Substitution of "run up" with the colloquial phrasal verb "racked up."
Complex Style	"The accumulation of debt over the preceding eight years has exceeded, in magnitude, the total indebtedness incurred during the collective administrations of all presidents from George Washington to Jimmy Carter."	**Semantic Over-Formalisation:** Massive expansion of word count through bureaucratic synonyms (e.g., "accumulation," "indebtedness," "collective administrations").

Table 1: Comparison of Human Claim and LLM-Generated paraphrases across different styles.

			Semantic Analysis		Entity Integrity		Style	Length
Dataset	**Split**	**Category**	**BLEURT**	**MIS**	**Jaccard**	**Cosine**	**Sent. Δ**	**Len Δ**
CT24	Validation	Base	0.3734	0.2543	0.9383	0.9608	-0.0072	9.76
		Informal	0.3732	0.2555	0.9547	0.9738	0.0023	8.17
		Complex	0.3619	0.2774	0.9210	0.9599	-0.0108	90.45
	Testing	Base	0.3698	0.2535	0.9283	0.9633	-0.0072	9.62
		Informal	0.3706	0.2533	0.9605	0.9789	-0.0052	6.94
		Complex	0.3580	0.2763	0.9238	0.9538	-0.0290	96.95
CheckIt	Validation	Base	0.3526	0.2517	0.7493	0.8447	-0.0109	12.24
		Informal	0.3517	0.2533	0.7433	0.8387	-0.0056	9.38
		Complex	0.3355	0.2914	0.7110	0.8167	-0.0363	132.52
	Testing	Base	0.3530	0.2524	0.7622	0.8547	-0.0137	11.87
		Informal	0.3525	0.2534	0.7628	0.8629	-0.0006	9.13
		Complex	0.3357	0.2932	0.7309	0.8364	-0.0190	128.97

Table 2: Comprehensive evaluation of claim paraphrasing categorised by Dataset, Split, and Stylistic Category. Results indicate consistency in meaning preservation across validation and testing sets.

Prompt Style	Samples (n)	Semantic Preservation (%)	Mean Confidence (1–5)
Informal	50	88.0%	4.14
Complex	50	92.0%	4.22
Base (Rephrased)	50	84.0%	4.08
Total	**150**	**88.0%**	**4.15**

Table 3: Manual validation of LLM-generated paraphrases for semantic preservation and annotator confidence.

encoder models, particularly when handling LLM-paraphrased claims. The most notable improvements occur within the CheckIt dataset, where RoBERTa demonstrates an F1-score increase of + 5.22 on the validation split and + 3.61 on the testing split. BERT similarly shows consistent gains across CheckIt informal and complex claim types. Conversely, the CTL 2024 results present a more varied outcome; while validation scores remain stable or improve, some testing splits experience marginal declines.

4.2. SHAP Analysis

To investigate the linguistic shifts influencing model decisions, we utilised SHAP analysis (Mosca et al., 2022) to identify high attribution tokens for classification. We select the base and robust variants of RoBERTa to investigate classification, due to being the most consistent model.

As shown in Table 6, we measure the intersection

Dataset	Claim Type	Model	Validation Split			Testing Split		
			P	R	F1	P	R	F1
CheckIt	Human	BERT	81.36	81.53	80.99	78.56	78.94	78.46
		RoBERTa	80.86	81.08	80.55	80.47	80.63	79.98
		Llama 3.1	53.41	63.06	57.67	59.40	63.74	58.00
		ClaimBuster	67.43	58.67	59.08	67.76	57.43	57.49
	Base	BERT	98.34	77.93	78.09	76.79	76.46	76.60
		RoBERTa	77.62	75.90	76.32	77.33	76.35	76.65
		Llama 3.1	63.11	65.88	59.37	60.77	64.64	58.22
		ClaimBuster	67.69	63.06	63.97	66.00	61.37	62.20
	Informal	BERT	79.03	76.24	76.73	77.16	74.44	74.96
		RoBERTa	79.66	77.48	77.91	77.83	76.58	76.92
		Llama 3.1	62.50	65.54	60.69	67.37	57.77	58.02
		ClaimBuster	67.37	57.77	58.02	65.14	54.84	54.80
	Complex	BERT	73.07	73.54	73.22	72.74	73.20	72.89
		RoBERTa	74.58	73.09	73.51	72.56	71.28	71.69
		Llama 3.1	58.75	64.19	55.13	64.59	64.64	64.61
		ClaimBuster	64.59	64.64	64.61	65.30	64.53	64.85
CTL 2024	Human	BERT	97.86	97.87	97.86	87.82	87.98	86.56
		RoBERTa	97.29	97.29	97.25	87.34	87.10	86.05
		Llama 3.1	88.19	77.42	79.18	80.33	68.33	70.29
		ClaimBuster	98.68	98.64	98.65	82.84	83.68	82.84
	Base	BERT	97.01	96.90	96.93	86.47	86.80	86.56
		RoBERTa	97.31	97.29	97.25	87.53	87.68	86.99
		Llama 3.1	88.26	76.65	78.48	80.93	66.28	68.27
		ClaimBuster	97.24	97.00	97.05	83.95	84.16	84.04
	Informal	BERT	97.18	97.19	97.18	87.46	87.68	87.06
		RoBERTa	97.47	97.48	97.49	87.19	87.10	86.13
		Llama 3.1	88.92	80.43	81.90	81.07	71.55	73.32
		ClaimBuster	97.73	97.67	97.67	32.53	83.28	82.58
	Complex	BERT	95.13	94.96	95.02	86.80	87.10	86.89
		RoBERTa	96.88	96.90	96.87	87.19	87.10	86.13
		Llama 3.1	87.69	74.90	76.88	78.76	64.81	66.92
		ClaimBuster	94.98	94.19	94.36	85.68	85.63	85.66

Table 4: Comparative performance of check-worthiness models grouped by claim type.

between the tokens by SHAP attribution and named entities within each claim. For the human-authored claims in CheckIt test set, the robust RoBERTa model demonstrates a significantly higher alignment with factual entities compared to the base model. Suggesting that adversarial training encourages the model to anchor the check-worthiness decision to these verifiable subjects such as names, dates and organisations.

5. Discussion

5.1. RQ1: Sensitivity to Synthetic Phrasing

The experimental results in Table 4 confirm that state-of-the-art check-worthiness models are sensitive to stylistic variations introduced by LLM-based paraphrasing. BERT and RoBERTa exhibit a notable performance degradation when processing synthetic claims within the CheckIt dataset, being most pronounced in the complex and informal categories. However, performance remains relatively stable on the CTL 2024 test sets. This

Dataset	Model	Claim Type	Validation Split (△)			Testing Split (△)		
			P	R	F1	P	R	F1
CheckIt	BERT	Human	+0.27	+0.34	+0.66	+1.77	+1.69	+1.88
		Base	-17.88	+2.70	+2.44	+2.87	+3.38	+3.13
		Informal	-0.37	+1.80	+1.52	+0.70	+2.25	+2.06
		Complex	+2.21	+1.12	+1.67	+2.93	+1.57	+2.18
	RoBERTa	Human	+1.76	+1.69	+1.79	-1.04	-0.90	-0.84
		Base	+4.38	+6.19	+5.22	+3.15	+4.39	+3.61
		Informal	+2.00	+4.16	+3.74	+0.05	+1.46	+1.03
		Complex	+2.34	+4.27	+3.04	+1.54	+3.49	+2.23
CTL 2024	BERT	Human	+0.10	+0.10	+0.10	-0.75	-1.18	-0.88
		Base	+0.39	+0.48	+0.46	-2.01	-2.05	-3.05
		Informal	0.00	0.00	-0.02	-2.58	-2.64	-3.28
		Complex	+2.14	+2.33	+2.25	-0.03	-0.30	-1.03
	RoBERTa	Human	+0.13	+0.09	+0.15	0.00	+0.58	+1.27
		Base	-0.97	-1.26	-1.15	+0.70	+0.30	+1.09
		Informal	-0.90	-1.07	-1.03	+0.79	+1.17	+1.90
		Complex	-1.93	-2.62	-2.44	-0.47	-1.18	+0.07

Table 5: Performance delta (Δ = Robust − Original) for BERT and RoBERTa models across datasets.

Dataset	Model	Human (%)	Base (%)	Complex (%)	Informal (%)
CheckIt Val	Base	15.14	12.76	9.56	14.67
	Robust	13.47	11.77	7.90	15.38
CheckIt Test	Base	15.69	9.72	6.93	13.43
	Robust	**20.45**	**17.28**	**9.65**	**15.53**
CT24 Val	Base	6.76	6.88	7.41	8.18
	Robust	8.29	7.61	6.82	11.20
CT24 Test	Base	5.11	6.75	4.90	4.30
	Robust	7.55	7.73	5.75	7.52

Table 6: Percentage of overlap between SHAP attribution tokens and named entities. This metric indicates the extent to which the model anchors its check-worthiness prediction on core factual subjects vs. stylistic noise.

suggest that while models are sensitive to stylistic shifts, the impact is domain-dependent, with event-specific datasets potentially offering higher inherent resilience.

Crucially, the semantic validation metrics in Table 2 suggest that this decline is not caused by a loss of factual or semantic information. The MIS and cosine similarity remain high across all scores, indicating that the core entities and claims remain intact. Instead the models appear to rely on specific structural patterns inherent to human social media discourse. When the LLM restructures these claims, it effectively masks the check-worthiness signals that the encoder models were fine-tuned to detect.

The relative stability of the zero-shot LLM results suggest a state of distributional parity between the model evaluating the claims and the one generating the paraphrased variants. Unlike the encoders, the LLM exhibits a lack of dependency on surface level heuristics, which are acquired in the fine-tuning on human-authored corpora. While this leads to a more consistent performance across claim types, it also highlights domain insensitivity; the LLM appears unable to prioritise the contextual urgency and pragmatic cues that signify check-worthiness in the organic human discourse.

5.2. RQ2: Risks of Adversarial Training

The performance deltas presented in Table 5 indicate that while adversarial training is a viable strategy for improving robustness, it carries significant risks related to distributional interference and architectural sensitivity.

Given that the encoder models were trained ex-

clusively on the CTL 2024 training data, the results on the CheckIt datasets serve as a measure of stylistic transfer. Interestingly, for the CheckIt dataset, both BERT and RoBERTa show consistent improvements across almost all claim types. RoBERTa, in particular, achieves an F1-score increase of + 5.22 (Validation) and + 3.61 (Testing) on base paraphrases. This suggests that exposing the model to the synthetic data during its training on the CTL 2024 domain provides a generalisable stylistic edge that transfers effectively to the CheckIt data. However, the primary risk of this approach is evidenced in the CTL 2024 testing split, the same domain used for training. Here, we observe a distinct difference between architectures. While we observe a marginal gain in validation, it suffers a significant performance collapse in the test set. Specifically, the robust BERT variant experimented an F1-score decline of -3.05 on base and -3.28 on informal claims. Because the model was trained on the human and synthetic variants of the CTL 2024 data, these negative delta suggest that for BERT, the synthetic samples acted as a distraction. The model likely began to over fit to the paraphrased claims, losing its ability to generalise to the subtle pragmatic marks of check-worthiness.

In contrast, RoBERTa demonstrates an improvement on the CTL 2024 test set, with an increase of +1.27 on human claims and up to +1.9 on synthetic variants. This suggests that RoBERTa's robust optimisation and pre-training objective allow it to reconcile the divergent stylistic features of human social media discourse and synthetic LLM generated text more effectively than BERT.

While adversarial training significantly bolsters resilience on the CheckIt dataset, the marginal performance drop in the CTL 2024 test set suggests a degree of noise from the adversarial training. To ensure practical reliability in systems, future work should investigate curriculum learning (Wang et al., 2022), where the model is stabilised on the human-authored claims before then being exposed to the difficult synthetic claims.

5.3. RQ3: Feature Shift and Attribution

The systematic misclassification identified through qualitative error logs suggest that the fine-tuned encoders rely on narrow syntactic heuristics rather than semantic representations of check-worthiness. A primary failure is the models dependence on active voice. Our qualitative analysis shows that when LLMs shift a human claim to the passive voice, the models frequently flip from a correct to incorrect classification. Indicating that the models have internalised a claim-like syntax for verifiable intent, failing to recognise the same factual signal when restructured.

The degradation in the complex category (see Table 4) corresponds to a failure to process normalisation. When verbs are transformed into abstract noun phrases, the models often mis-classify the statement as not check-worthy. Suggesting that the encoder models conflate syntactic simplicity and prose with high priority factual assertions, while perceiving complex machine generated text as editorial or descriptive noise.

The Informal results reveal a persistent reliability heuristic. The addition of colloquialisms often trigger a negative prediction for claims previously identified as check-worthy in their human form. This implies that fine-tuning on organic human corpora leads models to associate standard formal registers with high priority verifiable claims and informal registers with subjective, low-priority claims. While the robust RoBERTa model begins to reconcile these features, the persistence of these errors suggest that even adversarial training struggles to decouple the need to check stylistic wrappers.

5.4. Future Work

While adversarial training improves resilience, several avenues remain for enhancing check-worthiness systems.

Categorical Information Disorder: Current models treat check-worthiness as a monolithic label. Future work should develop multi-label datasets incorporating the Wardle and Derakhshan (2017) framework. Distinguishing between misinformation (unintentional) and disinformation (malicious) would allow for more nuanced triage priorities, where a model prioritises claims not just by check-worthiness, but by potential harm and authorial intent.

Check-Worthiness Specific Adversarial Attacks: As practitioners increasingly deploy automated suites, malicious actors may weaponise LLMs to generate low-signal disinformation. Future research should investigate a broader spectrum of attacks, including homoglyphs (Roadhouse et al., 2024), character-level perturbations (Morris et al., 2020), and word shuffling (Lewoniewski et al., 2024). Developing robust training loops that simulate these specific evasion tactics is essential for maintaining defence in an adversarial information landscape.

Fully Synthetic Dataset Generation: Our paraphrasing approach demonstrates that models can identify LLM-restructured claims. A natural next step is the generation of fully synthetic datasets to alleviate the bottleneck of human annotation. However, this requires rigorous experimentation to ensure that synthetic claims reflect the pragmatic

nuances of human discourse without introducing systematic model hallucinations or reinforcement of existing training biases.

6. Conclusion

In this study, we investigated the robustness of automated check-worthiness models against the emerging threat of LLM generated claims. Our findings demonstrate that state-of-the-art transformer encoders are highly sensitive to stylistic and syntactic shifts introduced by LLM paraphrasing. Qualitative and SHAP-based analysis confirm that these models rely on surface-level heuristics, such as active voice and standard formal registers, leading to misclassification when factual claims are restructured into passive or formalised prose.

We show that while adversarial training is a viable strategy for mitigating this sensitivity, its efficacy is architecture dependent. While RoBERTa demonstrated significant improvements across both human and synthetic claims, BERT exhibited signs of distributional interference, suggesting that more robust pre-training objectives are required to reconcile the stylistic gaps between human and LLM discourse. Ultimately, our work underscores the need for automated check-worthiness and fact-checking as a whole to remain a viable defence against disinformation, models must move beyond syntactic patterns and develop to a more grounded representation of factual priority.

Limitations

Despite the insights gained from this benchmarking study, several limitations must be acknowledged. First, our evaluation relies on LLM-paraphrased claims derived from human-authored ground truths rather than fully fabricated synthetic disinformation. While this ensures label preservation for experimental control, it may not capture the full spectrum of linguistic hallucination or logical inconsistencies present in completely machine generated narratives.

Furthermore, the synthetic dataset was generated using a single model. While this provided a controlled environment for testing specific stylistic shifts, different LLM architectures exhibit varying generation entropies and stylistic biases which may limit the generalisability of these findings. Additionally, our manual verification process, which served as a primary hallucination check to ensure the LLM maintained factual integrity during paraphrasing, which was conducted by a single investigator. While the results showed high annotator confidence, the absence of multiple annotators precludes the calculation of inter-annotator agreement.

Finally, our analysis is restricted to English claims within the socio-political and public health domains. The stylistic heuristics identified (e.g. active vs passive voice) may manifest differently in other languages or specialised domains such as legal or scientific fact-checking.

Ethical Considerations

The primary ethical concern regarding this research is the dual-use risk inherent in probing model vulnerabilities. By identifying the specific stylistic wrappers that allow claims to bypass automated detection, there is a potential for a malicious actor to weaponise these findings to generate adversarial disinformation. However, by documenting these weakness, they should be viewed as a prerequisite for developing robust adversarial resistant models.

7. Bibliographical References

Sajjad Ahmed, Klestia Balla, Knut Hinkelmann, and Flavio Corradini. 2021. Fact Checking: Detection of Check Worthy Statements Through Support Vector Machine and Feed Forward Neural Network. In *Advances in Information and Communication*, pages 520–535, Cham. Springer International Publishing.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Firoj Alam, Alberto Barrón-Cedeño, Gullal S Cheema, Gautam Kishore Shahi, Sherzod Hakimov, Maram Hasanain, Chengkai Li, Rubén Míguez, Hamdy Mubarak, Wajdi Zaghouani, et al. 2023. Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content. In *CLEF (Working Notes)*, pages 219–235.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko. 2022. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation. In *Proceedings of the 60th Annual Meeting*

of the Association for Computational Linguistics: Student Research Workshop, pages 300–321, Dublin, Ireland. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information sciences*, 497:38–55.

Nicola Luigi Bragazzi and Sergio Garbarino. 2024. Understanding and combating misinformation: An evolutionary perspective. *JMIR infodemiology*, 4(1):e65521.

Onur Büyüktopaç and Tankut Acarman. 2019. Evaluation of cosine similarity feature for named entity recognition on tweets. In *International Conference on Man–Machine Interactions*, pages 125–135. Springer.

Regina Cazzamatta. 2025. Global misinformation trends: Commonalities and differences in topics, sources of falsehoods, and deception strategies across eight countries. *New Media & Society*, 27(11):6334–6358.

Emily Chen, Herbert Chang, Ashwin Rao, Kristina Lerman, Geoffrey Cowan, and Emilio Ferrara. 2021. Covid-19 misinformation and the 2020 us presidential election. *Harvard kennedy school misinformation review*.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Yu Seunghak, Roberto Di Pietro, Preslav Nakov, et al. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,{IJCAI-20}*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2):103219.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

D. Graves. 2018. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*.

L. Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture and Critique*, 10(3).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the association for computational linguistics*, 10:178–206.

Maram Hasanain, Reem Suwaileh, Sanne Weering, Chengkai Li, Tommaso Caselli, Wajdi Zaghouani, Alberto Barrón-Cedeño, Preslav Nakov, and Firoj Alam. 2024. Overview of the clef-2024 checkthat! lab task 1 on check-worthiness estimation of multigenre content. In *25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024*, pages 276–286. CEUR Workshop Proceedings.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Imke Henkel. 2021. Ideology and disinformation: how false news stories contributed to brexit. *Politics of disinformation: The influence of fake news on the public sphere*, pages 79–90.

Włodzimierz Lewoniewski, Piotr Stolarski, Milena Stróżyna, Elżbieta Lewańska, Aleksandra Wojewoda, Ewelina Księżniak, and Marcin Sawiński. 2024. Openfact at checkthat! 2024: Combining multiple attack methods for effective adversarial text generation. In *CEUR Workshop Proceedings*, volume 3740.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Laura Majer and Jan Šnajder. 2024. Claim Check-Worthiness Detection: How Well do LLMs Grasp Annotation Guidelines? In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 245–263, Miami, Florida, USA. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, pages 119–126.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pages 4593–4603.

Yida Mu, Mali Jin, Charlie Grimshaw, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1052–1062.

Sami Nenno. 2024. Is checkworthiness generalizable? Evaluating task and domain generalization of datasets for claim detection. *Neural Computing and Applications*, 36(24):15165–15176.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Charlie Roadhouse, Matthew Shardlow, and Ashley Williams. 2024. Mmu nlp at checkthat! 2024: Homoglyphs are adversarial attacks. In *CLEF (Working Notes)*, pages 580–589.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892.

Ivan Srba, Olesya Razuvayevskaya, João A Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, et al. 2025. A survey on automatic credibility assessment using textual credibility signals in the era of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1):1–13.

Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2026. Leveraging rationality labels for explainable claim check-worthiness. *IEEE Transactions on Artificial Intelligence*, 7(1):239–249.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A Survey on Curriculum Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Dustin Wright and Isabelle Augenstein. 2020. Claim Check-Worthiness Detection as Positive Unlabelled Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter*, 21(2):48–60.

Media Bias Within Information Disorder: Bridging Two Research Communities Through a Systematic Review

Francisco-Javier Rodrigo-Ginés[*], **Jorge Chamorro-Padial**[†]
[*]Universidad Nacional de Educación a Distancia — NLP & IR Group, Madrid, Spain
[†]OpenCodice Research, Spain
frodrigo@invi.uned.es, jorge@opencodice.com

Abstract

Information disorder research overwhelmingly focuses on fabricated or manipulated content (fake news, deepfakes, propaganda) while comparatively neglecting the most pervasive form of distorted information: media bias. Unlike outright falsehoods, media bias operates within the boundaries of factual reporting, distorting public understanding through framing, omission, and word choice rather than fabrication. This makes it harder to detect, harder to regulate, and paradoxically more influential, since it originates from trusted mainstream sources rather than marginal actors. In this position paper, we argue that media bias should be recognized as a first-class category within information disorder frameworks. Drawing on the Wardle and Derakhshan (2017) taxonomy, communication theory, and a systematic review of over 100 studies on automated media bias detection, we demonstrate that current frameworks inadequately account for the systematic distortion of *true* content. We present a consolidated taxonomy of media bias types organized by linguistic level, compare detection paradigms across the information disorder and media bias communities, and identify four properties that make media bias uniquely dangerous: its scale, its source credibility, the invisibility of omission, and its cumulative normative effect. We conclude with an integrated research agenda grounded in specific gaps identified through the review.

Keywords: information disorder, media bias, framing, omission bias, position paper, systematic review, NLP

1. Introduction

When researchers and policymakers speak of "information disorder", they almost invariably mean content that is *false*: fabricated news articles, doctored images, conspiracy theories shared on social media (Lazer et al., 2018; Vosoughi et al., 2018). This conflation of information disorder with falsehood has shaped research agendas, detection tools, and regulatory frameworks alike. Fact-checking initiatives, automated fake news classifiers, and platform content moderation policies all target the same object: content that deviates from factual reality.

Yet the most widespread and arguably most influential form of distorted information is not false at all. **Media bias**, the systematic tendency of news outlets to present information in ways that favour particular perspectives, interpretations, or actors (Hamborg et al., 2019), operates entirely within the boundaries of factual reporting. A biased article need not contain a single false claim. Instead, it shapes understanding through what it emphasizes and what it omits, through the words it chooses and the frames it constructs (Entman, 1993). This concern extends to modern language technologies: large language models (LLMs), trained on vast corpora of news text, risk inheriting and amplifying the very biases present in their training data (Gallegos et al., 2024; Park and Kim, 2025), making the study of media bias urgent not only for understanding journalism but also for building trustworthy AI systems.

In this position paper, we argue that media bias is an underrecognized and inadequately addressed form of information disorder. Drawing on the influential framework of Wardle and Derakhshan (2017), on communication theory, and on a systematic review of over 100 studies on automated media bias detection (Hamborg et al., 2019; Rodrigo-Ginés et al., 2024), we make three claims, each grounded in evidence from the review:

1. Media bias is a form of information disorder that current frameworks acknowledge in principle but marginalize in practice.
2. The properties that distinguish media bias from other forms of information disorder (its scale, its origin in trusted sources, the invisibility of omission, and its cumulative normative effect) make it *more* dangerous than fabricated content, not less.
3. The NLP communities working on information disorder (fake news, propaganda, fact-checking) and media bias detection operate largely in isolation, to the detriment of both, a disconnect clearly visible in the citation patterns and methodological choices documented in the review.

We conclude by proposing an integrated research agenda that positions media bias at the

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 45–54

centre of information disorder research, with specific calls for multilingual resources, perspectivist annotation frameworks, and evaluation paradigms that move beyond binary detection.

2. The Information Disorder Landscape

The term "information disorder" was popularized by Wardle and Derakhshan (2017) as an alternative to the imprecise and politically weaponized label "fake news". Their framework distinguishes three categories along two dimensions, *intent to harm* and *falsity*:

- **Misinformation**: false content shared without intent to harm (e.g., honest mistakes, misunderstood statistics).
- **Disinformation**: false content deliberately created and shared to cause harm (e.g., fabricated news, coordinated campaigns).
- **Malinformation**: genuine content shared with intent to harm (e.g., leaks, harassment, out-of-context sharing).

Within this tripartite model, Wardle and Derakhshan identify seven types of problematic content, ranging from satire and parody (lowest harm) to fabricated content (highest harm). Type 3, "misleading content: misleading use of information to frame an issue or an individual", is the category most directly relevant to media bias.

The framework's blind spot. While the Wardle–Derakhshan taxonomy is rightly influential, it carries an implicit assumption: problematic content is *exceptional*. The seven types describe deviations from normal journalistic practice: satire that is mistaken for news, content that is improperly attributed, narratives that are fabricated. Media bias, however, is not exceptional. It is the *norm*. As communication scholars have long documented, all news production involves selection, emphasis, and framing; neutrality is an aspirational ideal rather than a default state (McQuail, 2010; Entman, 1993). When bias is the baseline rather than the anomaly, a framework designed to identify departures from good journalism fundamentally struggles to account for it.

Where media bias fits, and does not fit. Media bias maps most closely to the "misleading content" category, but this mapping is imperfect. Misleading content in the original framework implies a deliberate or at least identifiable act of misrepresentation: a photograph cropped to change its meaning, a headline that mischaracterizes the article body. Much media bias, by contrast, is structural and often unconscious: editorial choices about which stories to cover, which sources to quote, and which perspectives to include reflect institutional norms, audience expectations, and market incentives rather than individual acts of deception (Groeling, 2013; Shoemaker and Vos, 2009). The framework's reliance on intent (the mis/dis/mal distinction) maps poorly onto a phenomenon where intent is diffuse, institutional, and frequently invisible even to the journalists producing the content (Eberl et al., 2017).

Moreover, the "misleading content" category is one of seven, situated between "false connection" (type 2) and "imposter content" (type 4). This positioning implies rough equivalence in scope and importance. In reality, media bias dwarfs all other categories in volume: every news article carries framing choices, while fabricated content, propaganda, and imposter content are comparatively rare events in the broader information ecosystem (Vosoughi et al., 2018).

3. Media Bias: A Taxonomy from the Literature

To understand why media bias warrants special attention within information disorder frameworks, we must first clarify what media bias *is*. The term is used loosely in public discourse; in research, it encompasses a structured set of phenomena operating at multiple linguistic levels (Hamborg et al., 2019; Rodrigo-Ginés et al., 2024).

The underlying systematic review. The taxonomy and evidence presented throughout this paper draw on a systematic review of automated media bias detection (Rodrigo-Ginés et al., 2024) that followed the PRISMA guidelines. The review queried Scopus, Web of Science, and IEEE Xplore using search terms combining "media bias" with computational detection methods, covering the period 2000–2023. After applying inclusion criteria (peer-reviewed, English-language studies addressing automated detection or characterization of media bias in news text), the final corpus comprised over 100 studies. Each study was coded for bias type addressed, detection method, dataset used, language, and evaluation approach. We refer the reader to the original review for the full protocol and detailed results; here we synthesize the findings most relevant to positioning media bias within information disorder.

Drawing on this review, we present a consolidated taxonomy that organizes bias types by the linguistic level at which they manifest. Table 1 summarizes this taxonomy and maps each type to its closest parallel in information disorder research.

Table 1: A taxonomy of media bias types organized by linguistic level, drawn from a systematic review of automated media bias detection (Rodrigo-Ginés et al., 2024). The rightmost column maps each type to its closest parallel in information disorder research, revealing that many bias types lack a direct counterpart.

Level	Bias Type	Description	Info. Disorder Parallel
Word / Token	Word choice / Labelling	Evaluatively loaded terms ("regime" vs. "government")	Propaganda: loaded language
	Subjective intensifiers	Opinion-injecting adjectives and adverbs	Propaganda: exaggeration
	Attribution bias	Selective quoting, misattribution of statements	Imposter content (partial)
	Sensationalism	Dramatization, hyperbolic language	Clickbait, false connection
Sentence	Opinions as facts	Subjective judgments presented as established truth	Misleading content
	Mind reading	Claiming knowledge of actors' thoughts without evidence	Fabricated content (partial)
	Source selection bias	Quoting only sources aligned with a viewpoint	*No direct parallel*
Article / Discourse	Framing bias	Selective emphasis of particular aspects of an event	Misleading content
	Omission bias	Systematic exclusion of relevant facts or perspectives	*No direct parallel*
	Spin	Interpreting events with a particular evaluative slant	Misleading content
Outlet / Corpus	Gatekeeping	Editorial decisions about which stories to cover	*No direct parallel*
	Coverage bias	Disproportionate attention to certain topics or actors	*No direct parallel*

Word- and sentence-level bias. The most frequently studied bias types operate at the surface level of text. Recasens et al. (2013) demonstrated that even single-word substitutions ("claimed" vs. "stated", "regime" vs. "government") can measurably shift reader perception. At the sentence level, presenting opinions as facts or attributing unverified mental states to public figures ("mind reading") introduces bias without any overtly evaluative language. These types are the most accessible to current NLP methods, as they leave explicit textual traces that can be detected through word-level or span-level analysis (Spinde et al., 2021a,b). Importantly, word-level bias types have clear parallels in propaganda detection: loaded language, exaggeration, and name-calling are among the persuasion techniques catalogued by Da San Martino et al. (2019). Yet, as we discuss in Section 6, these overlapping phenomena are studied by separate communities using separate datasets and separate terminology.

Article- and outlet-level bias. At higher linguistic levels, bias becomes progressively harder to detect and progressively more consequential. Framing bias, the selection and emphasis of particular aspects of a story while downplaying others (Entman, 1993), requires understanding how a full article structures its narrative. The same event (a protest, a policy change, an economic report) can be framed as a triumph or a failure, a step forward or a crisis, depending on which facts are foregrounded. Framing bias has received growing computational attention, notably through SemEval-2023 Task 3 (Piskorski et al., 2023) and dedicated media frames corpora (Card et al., 2015). Omission bias, what is *not* said, is structurally invisible: there is no textual trace to detect because the biasing element is the missing text (Puglisi and Snyder Jr., 2015). At the outlet level, gatekeeping and coverage bias reflect the editorial decision of what constitutes "news", itself an act of selection and omission (Shoemaker and Vos, 2009; McCombs and Shaw, 1972). The systematic review found that the vast majority of detection methods target word- or sentence-level bias, with article-level and outlet-level methods representing a small and growing but still underdeveloped area of research (Rodrigo-Ginés et al., 2024).

Key observations from the taxonomy. Three findings from this mapping are particularly relevant to our argument. First, the bias types most studied in NLP (word choice and framing) have partial parallels in propaganda detection (loaded language, persuasion techniques), yet the two communities

rarely share methods or datasets. Second, several consequential bias types (omission, gatekeeping, coverage bias, source selection) have *no direct parallel* in information disorder taxonomies. These are precisely the types that operate through absence rather than presence, making them invisible to verification-based approaches. Third, the taxonomy reveals a *hierarchy of subtlety*: word-level bias is the most detectable and the most studied, while outlet-level bias is the most consequential but the least amenable to current NLP methods (Baly et al., 2020).

The perspectivist nature of bias. A crucial complication, amply documented in the review, is that bias perception is inherently subjective. The same article may be judged as "balanced" by one reader and "biased" by another, depending on their prior beliefs, political identity, and expectations of journalistic norms (Eberl et al., 2017). This subjectivity has important methodological implications: datasets annotated by majority vote may systematically suppress minority perspectives, and evaluation metrics that treat bias as a binary ground truth may misrepresent the phenomenon (Basile et al., 2021; Cabitza et al., 2023). The systematic review found that the majority of existing datasets rely on aggregated labels, with perspectivist annotation remaining the exception rather than the norm (Rodrigo-Ginés et al., 2024).

4. Detection Methods: Two Parallel Tracks

The systematic review reveals a striking pattern: the information disorder and media bias communities have developed parallel but largely independent detection ecosystems. This section compares these two tracks, drawing on the method and dataset landscape documented in the review (Rodrigo-Ginés et al., 2024).

Information disorder detection. The computational study of information disorder has developed mature pipelines for fake news detection (Lazer et al., 2018; Zannettou et al., 2019), rumour verification, automated fact-checking (Thorne et al., 2018; Augenstein et al., 2019; Nakov et al., 2021), and propaganda detection (Da San Martino et al., 2019; Dimitrov et al., 2021). These tasks share a common assumption: there exists a ground truth (a claim is true or false, a source is reliable or unreliable) against which system output can be evaluated. The dominant paradigm is *verification*, that is, checking content against reality. Detection methods rely heavily on external knowledge sources (knowledge graphs, fact-checking databases) and on network-level signals (propagation patterns, source credibility scores) (Baly et al., 2020). Shared tasks have produced large-scale benchmarks: FEVER (Thorne et al., 2018) for fact verification, SemEval-2020 Task 11 and SemEval-2021 Task 6 for propaganda detection (Dimitrov et al., 2021), and numerous fake news detection corpora. Evaluation is typically based on standard classification metrics (accuracy, precision, recall, F1) applied to binary or multi-class labels with clear ground truth.

Media bias detection. The computational study of media bias, as documented in the systematic review, follows a different trajectory. Methods focus on bias detection at the lexical, sentence, and article levels (Hamborg et al., 2019; Recasens et al., 2013; Fan et al., 2019; Gangula et al., 2019), framing analysis (Card et al., 2015; Piskorski et al., 2023), and increasingly, the use of large language models as bias detectors (Wen and Younes, 2024; Maab et al., 2024; Lin et al., 2024; Trhlík and Stenetorp, 2024). This community operates with a different assumption: bias is not a binary factual property but a continuous, perspective-dependent characteristic. The dominant paradigm is *interpretation*, that is, analyzing how content is constructed rather than whether it is true. Detection relies primarily on textual features extracted from the content itself, with comparatively less use of external knowledge or network signals. The review documented a methodological evolution from early lexicon-based and feature-engineering approaches to transformer-based classifiers that achieve state-of-the-art performance on sentence-level bias detection (Rodrigo-Ginés et al., 2024).

Shared techniques, separate ecosystems. Despite these different paradigms, the two communities employ remarkably similar technical machinery: transformer-based classifiers, attention mechanisms for span-level annotation, and multi-task learning architectures. SemEval-2023 Task 3 (Piskorski et al., 2023), which addressed news genre, framing, and persuasion detection in a multilingual setup, represents a rare point of convergence, but remains the exception rather than the rule. The review found that media bias detection papers rarely cite propaganda detection work, and vice versa, even when they address overlapping phenomena such as loaded language or persuasive framing (Rodrigo-Ginés et al., 2024). This isolation extends to the use of LLMs: recent work on using GPT-based models for bias detection (Wen and Younes, 2024; Maab et al., 2024) and for propaganda detection employs similar prompting strategies but develops them independently.

The dataset landscape. The dataset gap is equally revealing. The systematic review identified a pronounced concentration of resources: the vast majority of media bias datasets are in English, with very few resources available for other languages (Hamborg et al., 2019; Fan et al., 2019; Spinde et al., 2021a; Lim et al., 2020). Key English datasets, including MBIC (Spinde et al., 2021a), BASIL (Fan et al., 2019), and BABE (Spinde et al., 2021b), provide word-level and sentence-level annotations but differ substantially in their annotation schemes, bias definitions, and granularity. MBIC includes annotator characteristics (demographics, political leaning), enabling analysis of how background affects bias perception, while BABE uses expert annotators for higher inter-annotator agreement. This fragmentation makes cross-dataset evaluation difficult and limits the generalizability of detection methods. By contrast, the information disorder community benefits from large-scale, standardized benchmarks (FEVER alone contains over 185,000 claims) that enable direct comparison across methods and have driven rapid methodological progress. The absence of comparable standardized resources for media bias detection is, in our view, one of the most significant practical consequences of the community's relative isolation.

5. Why Media Bias is Uniquely Dangerous

We identify four properties that distinguish media bias from other forms of information disorder and that, taken together, make it a uniquely powerful force for shaping public understanding. These properties are not merely theoretical assertions; each is supported by empirical evidence. Empirical studies have shown that biased mainstream media measurably shifts voting behaviour (DellaVigna and Kaplan, 2007) and drives political polarization (Martin and Yurukoglu, 2017), while exposure to fabricated news is far more limited than commonly assumed (Allcott and Gentzkow, 2017).

Scale. Fabricated content, while attention-grabbing, represents a small fraction of the information ecosystem. Vosoughi et al. (2018) found that false news stories on Twitter were shared by far fewer users than true stories, despite spreading faster within networks. Media bias, by contrast, is omnipresent: every article published by every news outlet carries framing choices, emphasis decisions, and omissions. The cumulative volume of biased-but-factual content vastly exceeds the volume of outright fabrication. If information disorder is defined by its capacity to distort public understanding, then the aggregate effect of daily biased reporting, across thousands of outlets, millions of articles, and billions of reader impressions, is likely far greater than that of viral fake news.

Source credibility. Disinformation typically originates from anonymous accounts, fringe websites, or state-backed troll operations, that is, sources that audiences have learned to distrust, at least in principle (Lazer et al., 2018). Media bias, however, is produced by the most trusted institutions in the information ecosystem: established newspapers, public broadcasters, and major digital news platforms. This is the paradox at the core of our argument: the very credibility that makes mainstream media valuable as information sources also makes their biases more influential. Readers apply less critical scrutiny to content from trusted sources, accepting framing choices and omissions as part of "the way things are" rather than as editorial decisions that could have been made differently (Ecker et al., 2022).

The invisibility of omission. Fact-checkers can identify false claims. Propaganda detection tools can flag loaded language and rhetorical manipulation (Da San Martino et al., 2019). Deepfake detectors can analyze visual artifacts (Vaccari and Chadwick, 2020). But no tool can flag what is *not there*. Omission bias, the systematic exclusion of perspectives, sources, or facts, leaves no trace in the published text. A reader cannot know what they were not told, and an automated system cannot detect the absence of content it was never given. As our taxonomy shows (Table 1), omission and gatekeeping bias have no parallel in information disorder frameworks precisely because those frameworks are built on the assumption that disordered content *exists* in a detectable form. This makes omission bias the most durable and least accountable form of information distortion: it cannot be "fact-checked" because no individual claim is false, and it cannot be detected by systems trained on textual features because the relevant signal is extratextual.

Cumulative normativity. Each individual biased article may appear unremarkable. The danger lies in accumulation: when audiences are consistently exposed to news framed from a particular perspective, that perspective becomes the perceived default, the "normal" way of understanding an issue (Scheufele, 1999; McCombs and Shaw, 1972). This normative effect is qualitatively different from the acute shock of encountering a false claim. False claims can be refuted; biased framing, absorbed over years of media consumption, reshapes the cognitive frameworks through

which audiences interpret all subsequent information. The systematic review found evidence of this asymmetry in computational terms as well: while fake news detection has converged on increasingly effective methods, media bias detection remains a significantly harder task, in part because the target itself (what counts as "biased") shifts with audience and context (Rodrigo-Ginés et al., 2024).

6. The Disconnect Between Research Communities

Despite the conceptual proximity of media bias and information disorder, the NLP communities working on these problems are remarkably disconnected, a pattern the systematic review makes quantitatively visible (Rodrigo-Ginés et al., 2024).

Divergent citation networks. Shared tasks treat fake news detection (SemEval-2019 Task 7), propaganda detection (SemEval-2020 Task 11, SemEval-2021 Task 6), and news framing (SemEval-2023 Task 3) as separate problems with separate datasets and separate evaluation metrics. Survey papers on information disorder rarely cite the media bias literature in depth, and vice versa. The Wardle–Derakhshan framework (Wardle and Derakhshan, 2017), widely cited in information disorder research, is rarely referenced in NLP papers on media bias detection, even when the detected bias types map directly to the framework's categories. Conversely, foundational media bias work such as Hamborg et al. (2019) and Recasens et al. (2013) is seldom cited in propaganda or fact-checking papers, despite addressing overlapping linguistic phenomena.

Missed methodological synergies. This fragmentation has practical consequences documented in the review. Methods developed for propaganda detection (e.g., persuasion technique classifiers) could directly inform framing bias detection, since persuasion and framing share rhetorical mechanisms: both involve selecting which aspects of reality to make salient and which to suppress. Perspectivist annotation methods developed in the media bias community (Basile et al., 2021; Cabitza et al., 2023) could improve the handling of subjectivity in fact-checking and credibility assessment, where annotator disagreement is typically treated as noise rather than signal. Source-level analysis (Baly et al., 2020) could provide contextual features for article-level bias detection, since the political orientation and editorial line of a news outlet constitute strong priors for the bias expected in individual articles. The LLM-based methods now being explored for bias detection (Wen and Younes, 2024; Maab et al., 2024) use prompting strategies remarkably similar to those employed for propaganda detection, yet the two lines of work develop these strategies independently.

The cost of disconnection. These synergies remain largely unexploited. The review identified only a handful of studies that explicitly bridge the two communities, most notably SemEval-2023 Task 3 (Piskorski et al., 2023), which combined genre classification, framing detection, and persuasion technique identification in a single multilingual shared task. The rarity of such bridging efforts suggests that the disconnect is not merely a citation gap but reflects deeper differences in how the two communities conceptualize their objects of study: verification vs. interpretation, truth vs. perspective, binary vs. spectral evaluation. Overcoming this divide requires more than cross-referencing papers; it requires rethinking shared tasks, evaluation metrics, and even the definition of what constitutes "disordered" information.

7. Toward an Integrated Research Agenda

We propose five directions for integrating media bias into information disorder research, each motivated by a specific gap identified through the systematic review:

1. Extend information disorder frameworks. The Wardle–Derakhshan taxonomy should be revised to include media bias as a primary category, not a subcategory of "misleading content". Our taxonomy (Table 1) demonstrates that at least four major bias types (omission, gatekeeping, coverage bias, and source selection) have no parallel in current information disorder classifications. We propose distinguishing between *content-level disorder* (fabrication, manipulation, impersonation) and *framing-level disorder* (bias, selective emphasis, omission), with explicit recognition that the latter operates on true content and requires different detection paradigms. This distinction is not merely taxonomic: it implies different annotation guidelines, different evaluation criteria, and different intervention strategies.

2. Develop multilingual media bias resources. The systematic review confirmed a severe English-language concentration: the vast majority of media bias datasets and detection methods are developed for English (Hamborg et al., 2019; Fan et al., 2019; Spinde et al., 2021a). Information disorder, however, is a global phenomenon, and media bias

manifests differently across linguistic and cultural contexts (Piskorski et al., 2023). SemEval-2023 Task 3 demonstrated both the feasibility and the value of multilingual framing analysis, but comparable resources for media bias detection beyond English remain scarce. We call for the creation of media bias resources in underrepresented languages, with annotation schemes sensitive to local journalistic norms and political contexts. Such resources would not only expand the geographic coverage of media bias research but also reveal culturally specific bias patterns that monolingual studies cannot capture.

3. Embrace perspectivist annotation. The review found that the majority of media bias datasets resolve annotator disagreement through majority vote, effectively treating bias as a binary factual property (Rodrigo-Ginés et al., 2024). This contradicts the inherently subjective nature of bias perception (Eberl et al., 2017). Rather than resolving disagreement through aggregation, annotation frameworks should preserve individual judgments, enabling models to predict distributions of opinion rather than single labels (Basile et al., 2021; Cabitza et al., 2023). This perspectivist approach is especially important for media bias, where the "ground truth" is not a fact to be verified but a judgment to be understood. Datasets such as MBIC (Spinde et al., 2021a), which record annotator characteristics alongside annotations, point toward this direction but remain exceptions in the field.

4. Move beyond binary detection. The review documented that current media bias detection is dominated by binary classification (biased vs. unbiased), with few systems modelling bias as a spectrum (Rodrigo-Ginés et al., 2024). A more nuanced approach would characterize the *type* and *direction* of bias simultaneously, following the multi-level taxonomy presented in Table 1. Each level of the taxonomy calls for different evaluative frameworks: word-level bias lends itself to span-extraction and token-classification methods akin to named entity recognition; sentence-level bias requires document-contextual classification; article-level framing demands discourse-aware models that capture narrative structure; and outlet-level bias necessitates corpus-comparative approaches that contrast coverage patterns across sources. Evaluation metrics should reflect this complexity, moving from accuracy and F1 toward measures that capture calibration, ranking quality, and agreement with diverse annotator populations. The information disorder community's experience with multi-label propaganda detection (Da San Martino et al., 2019), where a single text span may exhibit multiple persuasion techniques, offers a useful model for multi-type bias annotation.

5. Interrogate LLMs as both tools and vectors. Large language models present a dual challenge for media bias research. As detection tools, they show promising capabilities for identifying bias (Wen and Younes, 2024; Maab et al., 2024; Lin et al., 2024), but they also carry their own biases, shaped by training data that inevitably reflects the biases of the media it was drawn from (Gallegos et al., 2024; Park and Kim, 2025). An LLM trained on biased news may reproduce and even amplify the biases present in its training corpus. The review noted a rapid growth in LLM-based media bias studies, but these rarely address whether LLMs' own biases compromise their utility as detectors (Rodrigo-Ginés et al., 2024). Research on LLMs and media bias must address both directions: using LLMs to detect bias in content, and detecting bias in the LLMs themselves. The fact that these models are now being integrated into newsroom workflows for summarization, translation, and even content generation makes this question urgent.

8. Conclusion

Information disorder is not only about what is false. It is also, and perhaps primarily, about what is *true but distorted*: factual content presented through selective framing, loaded language, and strategic omission. Media bias is this distortion operating at scale, produced by the most trusted institutions in the information ecosystem, and rendered invisible by the very ordinariness of its mechanisms.

Current information disorder frameworks acknowledge the existence of misleading content but treat it as one category among many, equivalent in scope to satire or imposter content. We have argued that this is a fundamental mischaracterization, grounding our claims in a systematic review of over 100 studies that reveals both the breadth of media bias phenomena (Table 1) and the depth of the disconnect between the communities that study information disorder and media bias.

Media bias is not a peripheral form of information disorder; it is the most pervasive, the most credible, and in many ways the most dangerous, precisely because it leaves no false claims to fact-check and no fabricated content to debunk. Recognizing media bias as a first-class category of information disorder is not merely a taxonomic exercise. It has practical implications for how we build detection systems (moving beyond verification to interpretation), how we annotate data (embracing perspectivism rather than enforcing consensus), how we

evaluate progress (replacing binary metrics with spectral ones), and how we study the role of language technologies that are simultaneously tools for detecting bias and vectors for propagating it.

The inaugural edition of the InDor workshop represents an opportunity to define the scope of information disorder research broadly enough to include the invisible layer that has been hiding in plain sight. We urge the community to seize it.

Ethics Statement

This position paper advocates for greater attention to media bias within information disorder research. We acknowledge that defining what constitutes "bias" is inherently normative and culturally situated. Any operationalization of bias detection carries the risk of reflecting the perspectives and blind spots of its designers. We do not advocate for automated censorship or content removal based on bias scores; rather, we call for tools that increase transparency about framing choices and support media literacy.

Limitations

This paper is a position paper grounded in a systematic review rather than an empirical study. While we draw on evidence from over 100 reviewed studies, our arguments about the relative importance of media bias within information disorder are interpretive rather than experimentally verified. The taxonomy presented in Table 1 is a synthesis that necessarily simplifies the diversity of bias phenomena documented in the literature. Additionally, our analysis of the disconnect between research communities is based on citation patterns and shared task participation observed in the review, which may not capture all forms of cross-pollination (e.g., informal collaborations or unpublished work).

9. Bibliographical References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of EMNLP-IJCNLP*, pages 4685–4697. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Muhammed, and Preslav Nakov. 2020. What was written, by whom, and when? analyzing news through source-level factuality and bias. In *Proceedings of EMNLP*, pages 5765–5781. Association for Computational Linguistics.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL-IJCNLP*, pages 438–444. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of EMNLP-IJCNLP*, pages 5636–5646. Association for Computational Linguistics.

Stefano DellaVigna and Ethan Kaplan. 2007. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Dalvi, Hassan Sajjad, Alex Nikolov, Yordan Atadjanov, and Preslav Nakov. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of SemEval*, pages 70–98. Association for Computational Linguistics.

Jakob-Moritz Eberl, Hajo G. Boomgaarden, and Markus Wagner. 2017. One bias fits all? three types of media bias and their effects on party preferences. *Communication Research*, 44(8):1125–1148.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of EMNLP-IJCNLP*, pages 6343–6349. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Rama Rohit Reddy Gangula, Suma Dunn, and Jacob Eisenstein. 2019. Detecting media bias in news articles using gaussian bias distributions. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom*, pages 48–55. Association for Computational Linguistics.

Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16:129–151.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Sora Lim, Adam Jatowt, Masatoshi Yoshikawa, and Antoine Doucet. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1478–1484.

Liqiang Lin, Pengfei Li, and Frank Ferraro. 2024. IndiVec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. In *Proceedings of EACL*. Association for Computational Linguistics.

Fatima Maab, Wasim Afzal, and Muhammad Kamran Malik. 2024. Media bias detection across language model families. In *Proceedings of NAACL*. Association for Computational Linguistics.

Gregory J. Martin and Ali Yurukoglu. 2017. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–2599.

Maxwell E. McCombs and Donald L. Shaw. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187.

Denis McQuail. 2010. *McQuail's Mass Communication Theory*, 6th edition. SAGE Publications.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of IJCAI*, pages 4551–4558.

Sihyeon Park and Kunwoo Kim. 2025. Is the source reliable? the effect of media outlet names on bias detection in LLMs. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of SemEval*, pages 2343–2361. Association for Computational Linguistics.

Riccardo Puglisi and James M. Snyder Jr. 2015. Empirical studies of media bias. *Handbook of Media Economics*, 1:647–667.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. *Proceedings of ACL*, pages 1650–1659.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Dietram A. Scheufele. 1999. Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122.

Pamela J. Shoemaker and Tim P. Vos. 2009. Gatekeeping theory. *Routledge*. Book.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021a. MBIC – a media bias annotation dataset including annotator characteristics. In *Proceedings of the iConference*, pages 399–407.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021b. Neural media bias detection using distant supervision with BABE – bias annotations by experts. In *Findings of EMNLP*, pages 1166–1177. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT*, pages 809–819. Association for Computational Linguistics.

Vojtěch Trhlík and Pontus Stenetorp. 2024. Generative models for media bias detection. https://arxiv.org/abs/2406.10773. ArXiv:2406.10773.

Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1).

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Technical Report DGI(2017)09, Council of Europe. Report prepared for the Council of Europe.

Lucie-Charlotte Wen and Lina Younes. 2024. Can GPT-3.5 detect media bias? an evaluation on the MBIB benchmark. https://arxiv.org/abs/2403.20158. ArXiv:2403.20158.

Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, 11(3):1–37.

Reliable News or Propagandist News? A Neurosymbolic Model Using Genre, Topic, and Persuasion Techniques to Improve Robustness in Classification

Géraud Faye[1,2], **Benjamin Icard**[3], **Morgane Casanova**[4], **Guillaume Gadek**[1], **Guillaume Gravier**[4], **Wassila Ouerdane**[2], **Céline Hudelot**[2], **Sylvain Gatepaille**[1], **Paul Égré**[5]

[1]Airbus Defence and Space, France
[2]Université Paris-Saclay, CentraleSupélec, MICS, France
[3]LIP6, Sorbonne Université, CNRS, France
[4]Université de Rennes, CNRS, Inria, IRISA, France
[5]IRL Crossing, CNRS, Australia

Abstract

Among news disorders, propagandist news are particularly insidious, because they tend to mix oriented messages with factual reports intended to look like reliable news. To detect propaganda, extant approaches based on Language Models such as BERT are promising but often overfit their training datasets, due to biases in data collection. To enhance classification robustness and improve generalization to new sources, we propose a neurosymbolic approach combining non-contextual text embeddings (fastText) with symbolic conceptual features such as genre, topic, and persuasion techniques. Results show improvements over equivalent text-only methods, and ablation studies as well as explainability analyses confirm the benefits of the added features.

Keywords: Information disorder, Fake news, Propaganda, Classification, Topic modeling, Hybrid method, Neurosymbolic model, Ablation, Robustness

1. Introduction

Recent years have seen a sharp increase in online news manipulation, driven by renewed international tensions, as documented in Europe by various intelligence offices (VIGINUM[1] in France, ZEAM[2] in Germany) and monitoring organizations (viz. EU DisinfoLab[3]). Such manipulation of information, which we may refer to as "news disorder" (adapting the terminology of Wardle and Derakhshan 2017), is often orchestrated through press-like websites that mimic journalistic conventions and disseminate targeted narratives to shape opinion (aka. "pseudo-news", see Faye et al. 2024). This is concerning since such content is widely shared on social media, quickly reaching large audiences.[4]

Automated detection of news disorders has advanced in various directions, including media bias (Hamborg et al., 2019), misinformation via automated fact verification (Thorne et al., 2018), fake news (Zhou and Zafarani, 2020; Hu et al., 2025), and rumors (Shu et al., 2020). However, these methods often transfer poorly across other types of news disorder. In particular, they tend to break down on influence operations and propaganda, which hinge on context-dependent framing of genres and topics, and on several specific persuasion techniques (Da San Martino et al., 2019b, 2020b).

In this paper, we focus on the identification of propagandist news, which are particularly insidious because they tend to smuggle in oriented messages with factual reports intended to look like reliable news. For this detection task, LLMs show high performance on specific datasets, but suggestive of a potential overfit. To increase robustness in classification, we argue that hybrid approaches, which have been effective for other news disorders (e.g., Baly et al., 2018; Thorne et al., 2018; Ruchansky et al., 2017; Ma et al., 2017), can improve the detection of influence operations and propaganda in news. Drawing on existing corpora of transparent news versus propagandist news (Faye et al., 2024), we present a neurosymbolic detection model that combines static vector embeddings (fastText) with additional features that include genre, topic, and

[1] https://www.sgdsn.gouv.fr/notre-organisation/composantes/service-de-vigilance-et-protection-contre-les-ingerences-numeriques?

[2] https://www.bmi.bund.de/SharedDocs/schwerpunkte/EN/disinformation-election/zeam-artikel-en.html?

[3] https://www.disinfo.eu

[4] In September 2025, Pew Research Center estimates that 53% of U.S. adults used social media as a news source: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 55–65
12 May 2026.

persuasion techniques. We provide evidence, using biased splits of the training, validation, and test tests, that the incorporation of these features improves performance over text-only methods.

Section 2 reviews related work on fake news detection and on the challenges posed by propaganda. Section 3 introduces two existing datasets that include reliable articles and then propaganda articles, to identify cross-corpus differences and features characteristic of propaganda. Building on these observations, Section 5 introduces a neurosymbolic model that merges dense text embeddings with interpretable conceptual features automatically extracted from the texts. Section 6 compares the performance of our hybrid method with that of a text-only benchmark; it evaluates its robustness relative to different partitions of the training/valid/test tests, and it uses ablation studies and explainability analyses to validate the method. Section 7 discusses the findings and outlines directions for future work.

2. Related work

The detection of information manipulation and news disorders is a broad domain that involves identifying different types of problematic content. The broadest category for misinformation is that of *fake news*, often defined as false or biased information, produced with an agenda or by negligence (Baptista and Gradim, 2022). Other varieties of news disorder include *rumors*, typically news reporting information that is hardly verifiable when published, intending to capture public attention. A specific variety of fake news is *propaganda*, namely partisan information seeking to set a narrative in order to debunk an enemy and glorify a state or organization.

Fake news detection (Hu et al., 2022; Zhou and Zafarani, 2020) is a popular area of research. It can be detected using transformer-based models (Pelrine et al., 2021), or using linguistic features and web markup features (Castelo et al., 2019), or combining linguistic and knowledge features (Seddari et al., 2022; Guelorget et al., 2021).

Fake news can also be easily detected using social media propagation patterns (Silva et al., 2021a; Davoudi et al., 2022), reaching in some cases more than 99% accuracy on four-day propagation data. However, in the current context, content-based fake news detection is more relevant, as a four-day delay is too long for effective detection.

Other approaches rely on user reports (Tschiatschek et al., 2018) and are designed to remain effective even when the majority of users engage in malicious reporting behavior. Various datasets exist for this subtask, often annotated by journalists, such as PolitiFact (Shu et al., 2020) and Horne2017 (**?**) on US politics, CoAID (Cui and Lee, 2020) on Covid-19, as well as LIAR (Wang, 2017), MultiFC (Augenstein et al., 2019) or MuMIN (Nielsen and McConville, 2022) on diverse topics. In the case of rumors, related datasets for this task are Fakenewsnet (Shu et al., 2020), relying on labels produced by PolitiFact and GossipCop, and PHEME (Kochkina et al., 2018).

A method put forward to analyze the news and to track whether they count as reliable or fake is *stance classification* (Riedel et al., 2017). In one version of the task, the goal is to assign a claim-evidence pair to one of three categories: the evidence either supports the claim, contradicts it, or fails to provide sufficient evidence. The primary dataset for this task is FEVER (Thorne et al., 2018), containing more than 300,000 facts. This task could help detect misinformation based on the content to be checked and a small collection of related evidence, making this task also close to fact-checking. The method has also been used to detect propaganda (Hanley, 2025).

Propaganda differs from other forms of news disorder by explicitly mimicking news articles and relying on frames and persuasion techniques (Da San Martino et al., 2020b). Barrón-Cedeño et al., 2019 computes a propaganda score, defined as the estimated likelihood that an article contains propagandistic mechanisms, using engineered stylistic and lexical features such as readability, lexical richness, and TF-IDF n-grams, whereas Da San Martino et al., 2019b introduces a fragment-level annotation scheme in which expert annotators mark the exact spans in news articles that realize propaganda and label each marked span with one of 18 propaganda persuasion techniques. This formulation, combining span identification with technique classification, is benchmarked in NLP4IF-2019 (Da San Martino et al., 2019a) and SemEval-2020 Task 11 (Da San Martino et al., 2020a).

Faye et al. (2024) compare human annotations with model predictions for multi-label propaganda analysis of press articles and systematically evaluates which stylistic cues explain performance. In particular, they show that feature sets targeting vagueness and subjectivity, together with syntactic and lexical cues, can achieve performance comparable to RoBERTa, while making the textual correlates of predictions more explicit.

Classical content-only approaches for news disorder detection often overfit to dataset or domain artifacts and therefore generalize poorly across outlets, genres, topics, and types of news disorder (Suprem et al., 2022; Silva et al., 2021b; Pan et al., 2023; Krieger et al., 2022). To improve robustness, neurosymbolic hybrid approaches pair neural representations of news articles with structured, complementary signals such as external evidence, source

metadata, lexicon and stylistic indicators, or multimodal cues. This strategy has been effective for misinformation through evidence-based verification and debunking (Thorne et al., 2018; Popat et al., 2018), for fake news by incorporating social-context and multimodal signals (Ruchansky et al., 2017; Wang et al., 2018), and for media bias by combining article text with source-level features and distant supervision (Baly et al., 2018; Spinde et al., 2021).

In the wake of these approaches, here we present a neurosymbolic model for propaganda detection grounded in stylistic features and observable symbolic cues, motivated by differences reported across previously published news corpora. We first conduct a comparative analysis of propaganda and mainstream news, then leverage the observed contrasts to design and evaluate our feature-based model using different organizations of our dataset to aim for more robustness.

3. Datasets used

3.1. Two corpora

In the rest of this paper, we exploit two datasets recently presented in Faye et al. (2024), which respectively consist of a corpus of propagandist pseudo-news articles (PPN) and a corpus of reliable articles from the mainstream press (MAINSTREAM).

- PPN[5] (Faye et al., 2024), for Propagandist Pseudo-News, is a collection of 12,427 articles from sources identified as propaganda outlets by the expert organizations NewsGuard and VIGINUM.[6] The five sources were created after the Russian invasion of Ukraine in February 2022 and contain propagandist news in 9 different languages (Arabic, Chinese, English, French, German, Italian, Russian, Spanish and Ukrainian).

- MAINSTREAM is a corpus of French and English articles, this time of regular news coming from established newspapers, and used as a control for the analysis of the PPN corpus. The MAINSTREAM articles were selected based on publication dates and on keywords related to the Ukraine conflict. MAINSTREAM consists of 1,004 English articles and 1,367 French articles from 11 and 5 sources, respectively.

While both datasets were introduced in Faye et al. (2024), only a small portion was analyzed in the context of an annotation experiment (100 in French split across both sources). Here, we analyze the whole corpus in full and using different methods.[7]

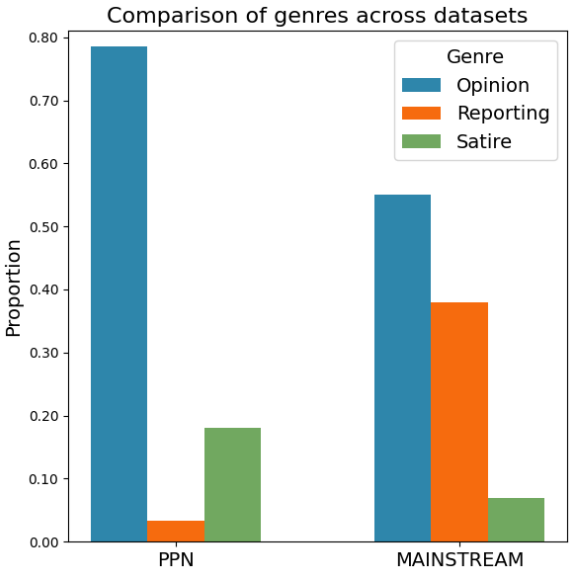


Figure 1: News genre distribution on the two corpora.

3.2. Review of published observations

Before presenting new analyses, we briefly summarize the main results of previous analyzes.

Firstly, a French subset of PPN and MAINSTREAM was manually annotated using 11 labels, adapted from a previous annotation experiment.[8] The labels included "vague", "subjective", "exaggeration", "pejorative", "descriptive", "propaganda", "satirical", "dishonest title", "adequate sources", "false information", and "fake news".[9]

Mainstream articles were found to be generally descriptive (close to 90%) and to adequately cite their sources (above 80%). They received low scores for "subjective" (20%) and "vague" (10%), and hardly any of the other labels, including no ascription of "false information" (0%). By contrast, propaganda articles were labeled as descriptive for only 60%, and they were overwhelmingly labeled

[5] https://github.com/hybrinfox/ppn

[6] https://www.sgdsn.gouv.fr/publications/maj-19062023-rrn-une-campagne-numerique-de-manipulation-de-linformation-complexe-et

[7] Because of copyright issues, the text is not redistributed, but direct links to web pages are provided with the corresponding annotations.

[8] See the OBSINFOX dataset, https://github.com/obs-info/obsinfox and its analysis in Icard et al. (2024). The focus on French was to take advantage of the native competence of the annotators.

[9] The difference between "fake news" and "false information" was that in order to be ascribed the former label, an article had to contain "at least one false information". The definition of "fake news" was deliberately left up to the annotators in the experiment, see (Faye et al., 2024) for details.

as manifesting subjectivity (70%), with up to 30% of "false information" and over 40% of "fake news".

Secondly, Large Language Models fine-tuned for the task, such as RoBERTa-base (Liu et al., 2019), were trained and tested on the whole English corpus. The LLMs were found to distinguish with very high performance between propaganda and regular articles (99.7% of test accuracy). Other models, such as TF-IDF (Sparck Jones, 1972), gave a similar high performance (98.5%), and also evidenced lexical differences between the two English corpora.

In both cases, however, this very high performance raises suspicion, since it can be a sign that the methods won't transfer to new datasets. While we do not see direct evidence of overfitting in the comparison between our training, validation, and test tests, such high accuracy points to the risk of a lack of robustness regarding other datasets. Furthermore, it is also possible that articles from the PPN dataset are partially written by AI, adding further biases to the model, leading to more overfitting on machine-written misinformation (Su et al., 2023).

4. Genres, topics, and persuasion techniques

To get a more general comparison of the two datasets, here we use distinct analytics of *genre*, *topic*, and *persuasion-techniques* distributions across the two corpora, following the distinctions proposed for the SemEval-2023 Task 3, and using the public APIs of the news classifier GATE Cloud.[10]

Genres. A three-fold genre distribution into Reporting articles / Opinion articles / Satire-like articles, is shown in Figure 1 for the two corpora.

The comparison reveals significant differences between the datasets: MAINSTREAM is characterized by a larger proportion of Reporting articles, more than six times the proportion in PPN. MAINSTREAM also shows a lower proportion of Satire. While these articles do not necessarily display humorous content characteristic of satire (see Icard et al. 2024, where the label "satirical", defined as intending to produce laughter, was applied less than 5% even for propagandist articles), this finding supports previous observations that fake news are stylistically closer to satire in style than regular news (Horne and Adali, 2017).

Finally, while Opinion is significantly represented across the two corpora, including the MAINSTREAM one, the class represents more than three quarters of the propagandist corpus PPN, confirming the link between propaganda and persuasion, and the tendency of propaganda to blur the frontier between factual reports and opinion pieces.

Topics. We used the same suite of annotating tools to get the topic distribution of the articles. A division along nine topics is shown in Figure 2, showing the two corpora to have relatively similar distributions.[11] This suggests that the two datasets can meaningfully be compared in terms of stylistic features, since they broadly have the same coverage.

Persuasion techniques. Finally, we used a third distributional analysis, this time relative to the set of persuasion techniques defined in Piskorski et al. (2023), again using the Cloud multilingual persuasion technique classifier. The distribution of persuasion techniques by articles is shown in Figure 3.

The plots indicate that propaganda articles from the PPN corpus tend to use more of these persuasion techniques, which is coherent with the fact that more than 90% of that corpus is identified as Opinion or Satire. In particular, we see a more prevalent use of *loaded language*, *repetition*, *exaggeration-minimization*, and *appeal to prejudice* in the propagandist corpus.

Overall, these analyses show us that in terms of genres as well as persuasion techniques, propagandist articles are more easily recognizable than other kinds of articles. This fact, combined with the additional characteristics explored in Faye et al. (2024), suggests that we may enhance the detection of propaganda by taking into account genres, persuasion techniques, and other stylistic features. To address this, the next section introduces a hybrid approach that integrates neural and symbolic representations by combining text-based features with concepts extracted from the content.

5. Neurosymbolic approach for propaganda detection

For the remaining of the paper, we focus on the English part of the corpus (3219 articles for PPN and 1004 articles for MAINSTREAM), as the models used perform better on this language, and will allow for better quantitative evaluation of our approach. The imbalance between the classes is managed by using the WeightedRandomSampler of PyTorch to expose the model to a balanced amount of classes during training.

To enhance robustness, here we embed text using neurosymbolic methods, and we add conceptual information to the articles to enhance classi-

[10] https://cloud.gate.ac.uk/shopfront#tagged=Misinformation

[11] A similar distribution was also found in the OBSINFOX dataset mentioned in fn. 8.

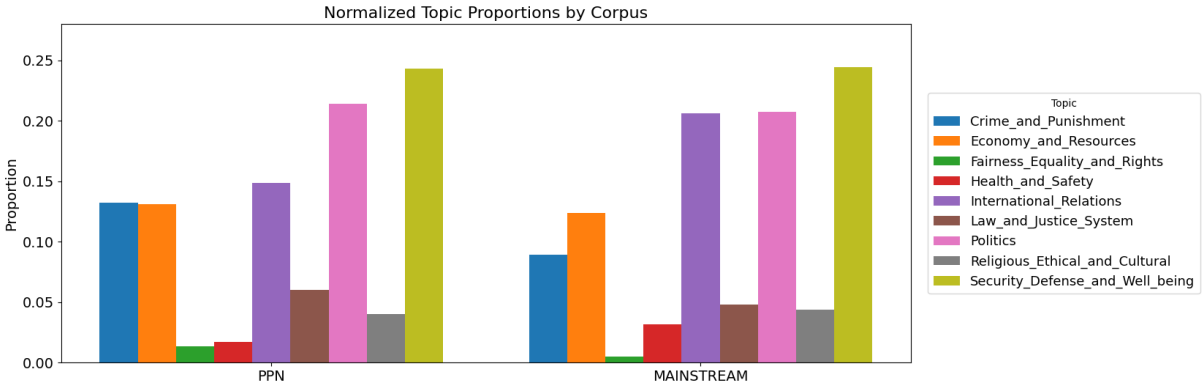


Figure 2: Topic distribution on the two corpora.

fication. Text embeddings are an efficient way of encoding texts and classifying them in downstream tasks. However, there is little understanding of the information they contain. To create a more robust model, we propose combining simple statistical text embeddings with the features observed in the previous section. Thus, texts are encoded using pre-trained fastText embeddings (Bojanowski et al., 2017). While these are non-contextual embeddings, they provide stable and parcimonious lexical representations for large amounts of data.

5.1. Proposed architecture

For each article, the process creates a 300-dimensional vector representing the text's distributional lexical characteristics. In addition to this vector, information about the genre, topic, and persuasion techniques contained in the articles is added.

- Genre information is encoded using one-hot encoding (OHE), creating a vector containing only zeros with the exception of the encoded feature, which contains a one. In this case, it adds 3 specific dimensions (for Reporting, Opinion, and Satire).
- Topic information is also one-hot encoded, adding 9 dimensions (for the topics displayed in Figure 2).
- Information about persuasion techniques is added into a vector counting how many persuasion techniques of each type are contained in the article. Fine-grained persuasion techniques represent 23 dimensions. However, these techniques can be grouped into coarser-grained groups, resulting in only 6 dimensions (Piskorski et al., 2023).

In total, a 35-dimensional vector (=3+9+23) for fine-grained persuasion techniques is added to the 300-dimensional fastText embedding. This vector goes through a two-layer perceptron containing a dense layer (335 dimensions to 335 dimensions) with ReLU activation function, and a dense layer (335 dimensions to 1 dimension) with sigmoid activation function to get a propaganda estimation score between 0 and 1. The text is then classified relative to a threshold of 0.5. A global view of the proposed architecture is presented in Figure 4. When all persuasion techniques are included, we call the resulting method the *Hybrid Method*. A scaled-down method is obtained by using an 18-dimensional (=3+9+6) vector incorporating only coarse-grained persuasion techniques: we call the corresponding method *Hybrid Lite*.

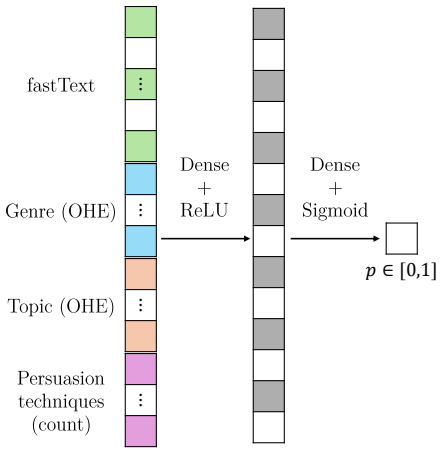


Figure 4: Our hybrid architecture, combining neural features and extracted concepts.

This architecture is voluntarily made simple in order to perform explainability analyses, which are hardly possible with Large Language Models. For this binary classification class, we use a single neuron as an output as we frame the task as we oppose directly reliable news to propagandist news.

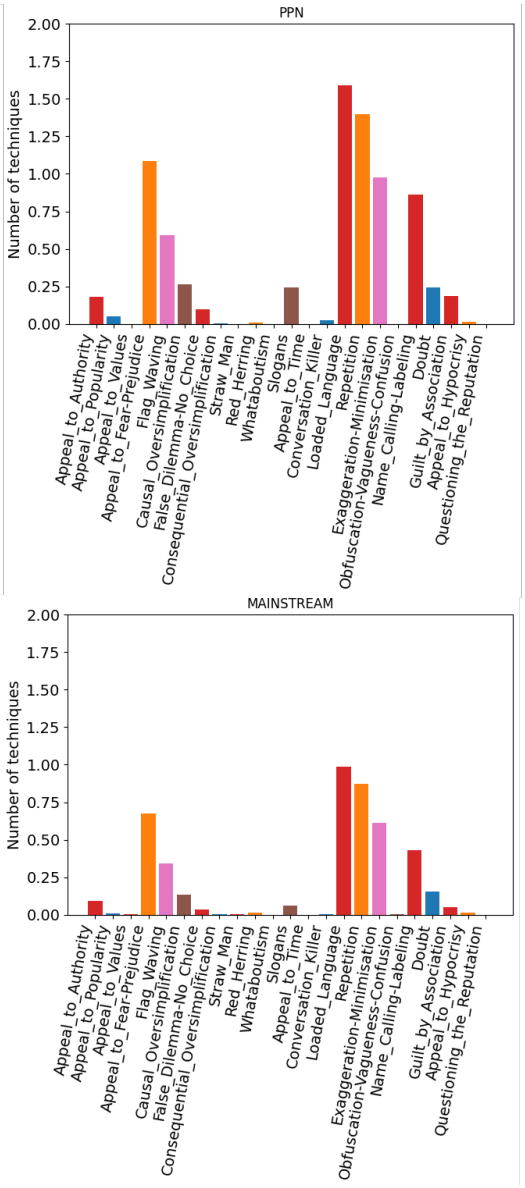


Figure 3: Mean number of persuasion techniques by article, on the two corpora.

5.2. Evaluation methodology

In order to prevent overfitting, we designed a new evaluation methodology to ensure that the proposed approach could generalize better to new events and sources.

The general idea is to split the data into train/valid/test sets depending on different criteria explained below. The model is trained on the train set and evaluated on the validation set after each epoch. Early stopping with patience 20 is used to monitor the F1-score on the validation set. The F1-score is the harmonic mean between precision (true positive/true positives+false positives) and recall (true positives/true positives+false negatives). In what follows, propaganda articles are the positive class to detect.

The models are trained with a cross-entropy loss and the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-4} for a maximum of 300 epochs. If no improvement in the validation F1-score is observed after 20 epochs, the best model is restored and evaluated on the test set, giving the scores reported in the results tables. We ran the experiment five times with different seeds and report the average results over these 5 runs.

The way the train/valid/test sets are defined can help measure robustness, namely how effective the learned features are in new contexts. Toward that goal, we define four types of split for our experiments (again on the English part of each corpus):

- **Random**: The articles are randomly sampled to produce 80%-10%-10% random sets.
- **Sources**: The split sets only contain articles from specific sources. The sources were chosen to have an approximately 80%-10%-10% split distribution. The split of sources is given in Table 1 (top).
- **Political**: Mainstream articles are split according to their political leaning annotation by MediaBiasFactCheck.[12] As a majority of articles come from left-leaning sources, we use them as training sources, and randomly split right-leaning sources between validation and test. Propaganda articles are randomly chosen for each set following an 80%-10%-10% distribution (see Table 1 middle).
- **Credibility**: Similarly to political leaning, MediaBiasFactCheck proposes credibility ratings of sources based on freedom of the press, articles' factuality, ownership, and previous fact-checks. In this sense, all propaganda articles come from low credibility sources. As a large majority of regular articles come from high-credibility sources, we use these sources for training and low-credibility sources for validation and test. Propaganda articles are randomly chosen for each set following an 80%-10%-10% distribution (see Table 1, bottom).

Each split is designed to evaluate one type of potential bias of our model based on sources, which may combine several types of unidentified biases, and then on political orientation, and credibility.

6. Results

This section is divided into three parts. The first provides the results of the proposed approach on the

[12] It is important to note that MediaBiasFactCheck seems to follow the American Overton window, so their political annotation may differ from what could be considered in other countries.

Sources	Train	Validation	Test
MAINSTREAM	APNews The Guardian	CNN USA Today Forbes Fox News NBC News NYTimes Washington Post	CBSNews Daily Mail
PPN	RRN	TribunalUkraine War on Fakes	

Political	Train	Validation	Test
MAINSTREAM	APNews CNN USA Today NBC News NYTimes Washington Post CBSNews	Daily Mail Forbes Fox News	
PPN	Entire PPN corpus (English)		

Credibility	Train	Validation	Test
MAINSTREAM	APNews CNN USA Today Forbes The Guardian NYTimes Washington Post CBSNews	Daily Mail	Fox News
PPN	Entire PPN corpus (English)		

Table 1: News distribution in the **Sources**, **Political**, and **Credibility** splits.

different splits. In the second part, ablation studies are conducted to measure the benefit of adding conceptual embeddings to the textual embeddings. Finally, an explainability analysis highlights in which cases the proposed approach has more benefits than others.

6.1. Main results

Results for the different splits are shown in the first row of Table 2 (Hybrid), reporting Accuracy (Proportion Correct) and F1 score. Note that the test sets are different in each column. The **Random** column corresponds to classical evaluation. The **Sources** column corresponds to the system being confronted with new sources, the **Political** column to the system being confronted with new political ideas, and the **Credibility** column to the system being confronted to sources of different credibility from the training set.

The results obtained on the **Random** set are not high but are decent for such a small model. The system shows comparable performance for the **Sources** and **Credibility** splits, but has more difficulties dealing with new **Political** orientations. Compared to **Credibility**, **Political** does not include *Forbes* in the validation and test sets, but has it in the train set along with *The Guardian* and without *CBSNews*. These shifts suffice to lower performance, suggesting that the political orientation of training sets should be variegated to create more robust systems.

6.2. Ablation studies

The motivation for the proposed approach is to improve robustness in new scenarios by combining conceptual features with text embeddings to reduce overfitting. To evaluate the performance of our Hybrid method, we conducted ablation studies. To begin with, a model using only the fastText embeddings is trained and evaluated (Table 2, Text Only). Then, the features of the persuasion techniques are altered to take only into account the coarse persuasion categories (6 instead of 23, see Table 2, Hybrid Lite).

Several observations can be made:

- In nearly all cases, using fine-grained labels for persuasion techniques (Hybrid) improves performance over using coarse-grained labels (Hybrid Lite). One exception is the **Random** split, but the gain is large where the Hybrid Lite method struggles (**Sources**, **Credibility**), and the loss small otherwise.
- On average, the Hybrid method improves performance compared to the Text Only method (+26.89% accuracy and +41.85% F1-score average), even though it performs less well in **Random** and **Political**. Overall, however, the Hybrid method is the most robust across the four splits: it learns in all four cases, and it has the largest average performance with the least variance ($\mu_{F1} = 86.12, var_{F1} = 9.18$).
- By contrast, whereas the Text Only method outperforms the others in two splits (**Random**, **Political**, in the **Sources** and **Credibility** splits textual embeddings were not sufficient to learn to discriminate between propaganda and mainstream articles. This indicates that the Text Only method is not robust to perturbations of the training set, even though it does not overfit on political orientation.

In summary, even if the proposed approach does not perform best on traditional random splits, it shows better robustness and generalization than the equivalent text-only approach, which collapses in two cases. Another advantage of this approach is its simplicity, allowing for the application of explainability methods, to which we turn next.

6.3. Explainability analyses

To explain our hybrid model, we used SHAP (Lundberg and Lee, 2017). This game-theory-based ap-

Method	Random		Sources		Political		Credibility	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Hybrid	78.08	87.5	**79.45**	**88.37**	69.86	81.66	**79.45**	**86.95**
Hybrid Lite	**79.45**	**88.54**	34.24	31.42	67.12	79.66	56.16	69.81
Text Only	**79.45**	**88.54**	20.54	0.0	**79.45**	**88.54**	20.54	0.0

Table 2: Results for propaganda detection with different data splits and different ablations.

proach performs local explanations on each sample by determining which features are the most important for the final prediction. However, the produced explanations are local and dataset-dependent, they do not explain the model's behavior more generally.

To get a global representation of what a model has learned, we can average absolute SHAP values over the different splits. We grouped SHAP values by categories for better readability. For each sample of each split, we calculate the absolute value of the sum of all text-encoding SHAP values, and similarly for the embeddings of genre, topic, and persuasion techniques. Mean SHAP values by split group and category are shown in Figure 5.

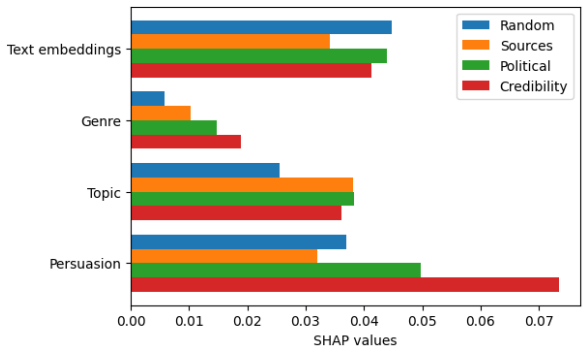


Figure 5: Mean SHAP values for the Hybrid method depending on split.

The figure shows that for **Random**, the text embeddings contribute the most to the decision classification. For **Credibility**, the persuasion embeddings contribute comparatively the most. For **Sources** and **Political**, the results are more mixed across groups of features.

In the **Random** case, this is likely due to the fact that by choosing articles randomly, the training set is better aligned with validation and test, so they are informative enough for classification. This is confirmed by the results, which show that they are better with textual features only.

However, when the test distribution is different from the training distribution, the method tends to use at least as many conceptual embeddings when compared with text embeddings. In particular, for the **Credibility** and **Sources** splits, relying on conceptual features is what makes the model generalize better to new sources and to papers whose reliability is questionable. This situation corresponds to the case in which an article is suspicious and comes from unknown sources, making this approach suitable for propaganda detection.

7. Conclusion

This paper introduced a propaganda detection method that integrates textual embeddings with conceptual features extracted through cross-comparison of PPN and MAINSTREAM. The corpora differ significantly in vocabulary and persuasion strategies, suggesting that models trained on a single source may miss corpus-specific signals. These observations motivated the inclusion of conceptual features to better capture propaganda patterns.

By designing biased splits in our datasets, which correspond to the exposure of the model to new types of articles, we have shown that adding conceptual information extracted from the texts improves detection performance, especially in cases where there are new sources of variable credibility ratings. Experiments also suggest that political diversity in the training set is essential for propaganda detection, as the addition of conceptual features significantly degrades performance in this case.

Further explainability analyses show that the added features were indeed used by the model when the splits were biased, allowing the model to correctly detect propaganda when simple textual embeddings are not informative enough.

However, the experiments were run on a corpus centered on one main theme: the Russia-Ukraine conflict. Additional experiments could be conducted on other recent themes, such as recent elections, or other conflicts. The PPN and MAINSTREAM corpora were also only processed in English, and similar experiments should be conducted in other languages to identify potential language-specific differences.

Finally, other types of conceptual features could be used based on other expert knowledge systems or even human operators. In other experiments, an expert vagueness estimation system was successfully combined with a language model for the task of subjectivity detection (Casanova et al., 2024). It may be possible to add an estimate of the document's source reliability to a classification model, to enhance the classification performance of a text-only classifier.

Acknowledgments

We thank two anonymous reviewers for helpful comments and feedback. This work was supported by the programs HYBRINFOX (ANR-21-ASIA-0003) and TRUSTEDNEWS (ANR-25-ASM2-0003). PE thanks the Department EEE of the University of Melbourne, and the Department of Philosophy of Monash University, for additional support.

8. References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

João Pedro Baptista and Anabela Gradim. 2022. A working definition of fake news. *Encyclopedia*, 2(1).

Alberto Barrón-Cedeño, Ismaeel Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Morgane Casanova, Julien Chanson, Benjamin Icard, Géraud Faye, Guillaume Gadek, Guillaume Gravier, and Égré Paul. 2024. HYBRINFOX at CheckThat! 2024 - task 2: Enriching bert models with the expert system VAGO for subjectivity detection. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, CLEF '2024, Grenoble, France.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 975–980, New York, NY, USA. Association for Computing Machinery.

Limeng Cui and Dongwon Lee. 2020. Coaid: COVID-19 healthcare misinformation dataset. *CoRR*, abs/2006.00885.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. In *Proceedings of the NLP4IF Workshop*. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of SemEval-2020*. Association for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, pages 4826–4832. Survey track.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Mansour Davoudi, Mohammad Moosavi, and M. Sadreddini. 2022. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Systems with Applications*, 198:116635.

Géraud Faye, Benjamin Icard, Morgane Casanova, Julien Chanson, François Maine, François Bancilhon, Guillaume Gadek, Guillaume Gravier, and Paul Égré. 2024. Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 62–72, Malta. Association for Computational Linguistics.

Paul Guelorget, Benjamin Icard, Guillaume Gadek, Souhir Gahbiche, Sylvain Gatepaille, Ghislain

Atemezing, and Paul Égré. 2021. Combining vagueness detection with deep learning to identify fake news. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, 20(2):391–415.

Hans WA Hanley. 2025. Tracking and identifying international propaganda and influence networks online. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29263–29264.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):759–766.

Bo Hu, Zhendong Mao, and Yongdong Zhang. 2025. An overview of fake news detection: From a new perspective. *Fundamental Research*, 5(1):332–346.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3:133–155.

Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Atemezing, François Bancilhon, and Paul Égré. 2024. A multi-label dataset of french fake news: Human and machine insights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 812–818.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL '22)*. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Dan S. Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153. Association for Computing Machinery.

Liangming Pan, Yunxiang Zhang, and Min-Yen Kan. 2023. Investigating zero- and few-shot generalization in fact verification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–524, Nusa Dua, Bali. Association for Computational Linguistics.

Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the web conference 2021*, pages 3432–3441.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE:

Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *ArXiv*, abs/1707.03264.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 797–806, New York, NY, USA. ACM.

Noureddine Seddari, Abdelouahid Derhab, Mohamed Belaoued, Waleed Halboob, Jalal Al-Muhtadi, and Abdelghani Bouras. 2022. A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. *IEEE Access*, 10:62097–62109.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021a. Propagation2vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58:102618.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021b. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 557–565.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models.

Abhijit Suprem, Sanjyot Vaidya, and Calton Pu. 2022. Exploring generalizability of fine-tuned models for fake news detection. In *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*, pages 82–88. IEEE.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 517–524, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policymaking.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

Grounding Information Disorder in NLP: A Theoretical and Operational Framework

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

This position paper proposes a theory grounded NLP framework for information disorder detection integrating three explicitly connected dimensions: epistemic status, intentionality, and contextual harm. Moving beyond binary fake news classification, we argue that reliable intervention requires structured differentiation between verification outcomes, manipulation indicators, and consequence assessment. We provide concrete annotation schemas with decision rules for ambiguous cases, formal aggregation operators with monotonicity and escalation guarantees, explicit conflict resolution strategies for inconsistent signals, and standardized risk profile templates that translate multidimensional outputs into actionable routing policies. Synthesizing work on harm taxonomies, uncertainty quantification, and automated fact checking pipelines, we introduce an integration layer that preserves interpretability while enabling policy aligned deployment. We further propose a reformed evaluation protocol incorporating conformal prediction for principled abstention, calibration analysis, disagreement modeling, harm weighted metrics, and human uplift assessment to measure real decision support utility rather than standalone classifier accuracy. We position this framework as a conceptual and operational roadmap for structured misinformation assessment, outlining phased validation pathways while acknowledging that empirical validation remains essential future work.

Keywords: information disorder, fake news detection, misinformation, NLP evaluation, harm taxonomy, uncertainty quantification

1. Introduction

The rapid expansion of digital media has transformed how information is produced, disseminated, and consumed globally. One of the most concerning manifestations has been the emergence of what is collectively known as fake news, a phenomenon attracting substantial research attention across computer science, communication studies, psychology, and political science. Initially conceptualized as a binary distinction between true and false information, fake news detection has historically focused on veracity classification using supervised machine learning. However, this binary framework is fundamentally insufficient to capture the full spectrum of misleading information circulating in contemporary media ecosystems (Wardle and Derakhshan, 2017).

The limitations of binary classification become apparent when considering the diversity of problematic information types. A satirical article misunderstood as genuine news, a genuine photograph shared with a misleading caption, an accurate statistic presented without crucial context, and a deliberately fabricated story designed to manipulate public opinion all represent fundamentally different phenomena demanding distinct analytical and interventional approaches. Yet conventional fake news detection systems typically collapse these distinctions into a single true/false dichotomy, losing critical information necessary for appropriate response.

This position paper shifts focus from binary classification towards a multidimensional, theory grounded approach to information disorder detection. We synthesize insights from existing work on harm taxonomies (Scheuerman et al., 2021; Sehat et al., 2024), uncertainty quantification in NLP (Xiao et al., 2022; Angelopoulos and Bates, 2021), integrated fact checking pipelines (Guo et al., 2022), and human centered AI research (Marusich et al., 2024; Prabhudesai et al., 2023) into a coherent framework with explicit integration mechanics. While no empirical validation is presented in this work, we ground all proposals in demonstrated capabilities from related literature and specify concrete validation pathways with measurable milestones, explicitly identifying the empirical questions that require answers before deployment.

Contributions. We position this as a conceptual roadmap with six contributions: (1) annotation schemas with decision rules for ambiguous cases; (2) formal aggregation operators specifying cross dimensional integration with conflict resolution; (3) adoption of FABLE harm taxonomy with concrete operationalizations and validation pathways; (4) integration of conformal prediction for principled abstention with coverage guarantees; (5) human uplift assessment framework measuring actual decision support utility; and (6) standardized risk profile templates for downstream stakeholder reporting.

2. Related Work

We organize related work around four themes that collectively motivate and inform the proposed

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 66–76
12 May 2026.

framework.

2.1. Harm Taxonomies and Content Moderation

A central challenge for information disorder research is moving beyond binary harm labels toward multidimensional severity assessment. Scheuerman et al. (2021) developed a comprehensive framework through grounded theory analysis with 52 participants, identifying four harm types (physical, emotional, relational, financial) and eight severity dimensions including perspective, intent, agency, and vulnerability. This work provides foundations for graduated response systems that differentiate content based on consequence severity rather than treating all harmful content equivalently. Building on this direction, the FABLE framework (Sehat et al., 2024) operationalizes harm specifically for fact checking prioritization through five dimensions: Fragmentation (social division potential), Actionability (harmful action likelihood), Believability (plausibility), Likelihood of spread (viral potential), and Exploitativeness (vulnerable group targeting). We adopt FABLE as our primary taxonomy due to its explicit operationalizability and alignment with downstream decision making.

Complementing these taxonomic efforts, Banko et al. (2020) synthesized platform guidelines into a unified taxonomy, highlighting definitional inconsistencies across platforms. Fortuna and Nunes (2018) surveyed hate speech taxonomies, documenting persistent challenges for consistent automated moderation. Wang et al. (2025) analyzed intent in algorithmic moderation, finding that current systems cannot reliably infer intent from text alone, a finding that motivates our proxy based approach for the intentionality dimension.

2.2. Uncertainty Quantification and Calibration

Reliable deployment of NLP systems in high stakes information disorder contexts requires principled handling of model uncertainty. Xiao et al. (2022) conducted a large scale empirical analysis showing that temperature scaling most effectively reduces expected calibration error (ECE) while preserving performance, and that larger pretrained language models calibrate better under domain shift. Desai and Durrett (2020) established calibration baselines for BERT family models, providing reference points for evaluating new systems. Xin et al. (2021) introduced error regularization that improves selective prediction, demonstrating accuracy coverage tradeoffs relevant to our abstention framework.

Conformal prediction offers a particularly promising direction by providing distribution free coverage guarantees without distributional assumptions. Angelopoulos and Bates (2021) provide an accessible introduction to conformal methods, while Quach et al. (2023) apply these methods to NLP classification, demonstrating that coverage guarantees hold even under distribution shift. These properties make conformal prediction well suited for misinformation contexts where content distributions change rapidly.

On the human side, research demonstrates that uncertainty information meaningfully affects decision quality. Marusich et al. (2024) showed that instance level uncertainty quantification improves human decision making, while Prabhudesai et al. (2023) found that users struggle to calibrate their reliance on AI without structured uncertainty communication. Devic et al. (2025) critique current UQ practices as insufficiently human centered, recommending evaluation of actual decision support utility rather than standalone calibration metrics.

2.3. Fact Checking Pipelines and Verification

Automated fact checking has matured into a modular pipeline encompassing claim detection, evidence retrieval, verdict prediction, and justification production, as comprehensively surveyed by Guo et al. (2022). Key resources include FEVER (Thorne et al., 2018) with 185K claims, LIAR (Wang, 2017) demonstrating six level graded assessment, and CheckThat! (Nakov et al., 2022) establishing multilingual benchmarks including Arabic. Our own involvement in the CheckThat! evaluation campaigns (Hasanain et al., 2024; Alam et al., 2021) has provided practical experience with the challenges of check worthiness estimation and multilingual claim detection at scale, informing the design requirements for the epistemic status dimension.

Zubiaga et al. (2018) survey rumor detection connecting to formal fact checking workflows. Nakov et al. (2021) emphasize that automated systems should support rather than replace human judgment, a principle central to our framework design. Stance detection research (Hardalov et al., 2022) provides computational foundations for estimating intent through analysis of how content positions itself relative to claims and entities. Related work on propaganda detection (Hasanain et al., 2023) and news media narrative analysis (Zaghouani et al., 2024) further demonstrates the practical complexity of identifying manipulation strategies in multilingual media ecosystems.

2.4. Multilingual, Multimodal, and Cultural Dimensions

Scaling information disorder detection beyond English requires addressing linguistic, cultural, and

resource diversity simultaneously. MuMiN (Nielsen and McConville, 2022) covers 41 languages with 13,000 claims. X FACT (Gupta and Srikumar, 2021) spans 25 languages. MM COVID (Li et al., 2020) offers multilingual multimodal data across six languages. Panchendrarajan and Zubiaga (2024) survey cross lingual claim detection approaches, while Panda and Levitan (2021) demonstrated mBERT for multilingual detection with variable performance across languages.

A critical consideration often underexplored is the cultural dimension of harm assessment. Recent work has shown that LLM persona simulation with cultural or demographic differences significantly affects the perception of harmful content. Piot et al. (2025) demonstrate that geographic persona assignment substantially influences LLM responses in hate speech detection, revealing systematic biases tied to country specific contexts. Plaza-del-Arco et al. (2024) show that LLMs exhibit culturally rooted biases in representing religion, with Eastern religions stereotyped and Abrahamic minority faiths stigmatized. These findings underscore that harm is not culturally universal and that information disorder frameworks must account for cultural variation in both annotation and deployment.

For multimodal detection, VERITE (Papadopoulos et al., 2024) provides robust benchmarks accounting for unimodal biases, while NewsCLIPpings (Luo et al., 2021) addresses out of context image detection. Work on multimodal propaganda detection in Arabic memes (Alam et al., 2024) illustrates the challenges of cross modal inconsistency where genuine images are paired with misleading text.

3. Information Disorder Framework

Following the influential taxonomy of Wardle and Derakhshan (2017), we adopt the tripartite distinction between misinformation (false content shared without intent to harm), disinformation (false content shared with intent to harm), and malinformation (genuine content shared with intent to harm). This theoretical foundation from communication studies motivates our three dimensional computational framework.

3.1. Conceptual Foundations

The information disorder framework addresses fundamental limitations of binary fake news classification. Traditional approaches treat all false claims as equivalent, ignoring crucial variations along three axes.

Epistemic Status concerns the degree to which a claim can be verified or falsified, the strength of available evidence, and the certainty with which verdicts can be assigned. A claim about yesterday's weather differs fundamentally from a claim about long term climate projections, even if both concern meteorology.

Intentionality concerns the motivations behind content creation and sharing. Identical false content may be shared by someone who genuinely believes it (misinformation), someone deliberately seeking to deceive (disinformation), or algorithms amplifying engaging content without regard to accuracy. These different origins suggest different interventions.

Contextual Harm concerns the potential consequences of content given its topic, timing, audience, and reach. A false claim about celebrity relationships differs fundamentally from a false claim about vaccine safety, even if both are equally false.

3.2. Relationship to Existing Approaches

We build on rather than dismiss existing progress. The LIAR dataset's six point veracity scale (Wang, 2017) demonstrates that graded epistemic assessment is both feasible and useful. CheckThat! shared tasks (Nakov et al., 2022) incorporate multi class and ranking formulations reflecting real world fact checking complexity. Stance detection research (Hardalov et al., 2022) provides computational foundations for estimating intent. Our contribution is not replacement of these existing components but systematic integration across dimensions with explicit aggregation mechanics and conflict resolution.

3.3. Framework Pipeline

The framework operates as a staged pipeline with three parallel assessment streams converging at an integration layer. **Stage 1 (Input Processing)** decomposes incoming content into atomic claims following Chen et al. (2022). **Stage 2 (Parallel Dimensional Assessment)** routes each claim through three parallel streams: the Epistemic Status stream performs evidence retrieval and entailment classification with conformal coverage; the Intentionality stream evaluates source history, linguistic manipulation markers, and coordination signals; the Contextual Harm stream applies FABLE dimensions to assess consequence severity. **Stage 3 (Integration)** applies the formal aggregation operator with conflict resolution rules to produce a composite risk score. **Stage 4 (Routing)** maps the integrated assessment to actionable decisions via the decision matrix, generating a structured risk profile for downstream stakeholders. Cases exceeding escalation thresholds or exhibiting cross dimensional conflicts are routed to human review.

4. Framework Architecture

Our framework comprises three dimensions with explicit integration mechanics. We provide concrete annotation schemas, formal aggregation operators, and standardized output templates.

4.1. Dimension 1: Epistemic Status

Epistemic Status concerns the degree of certainty and evidential support for a claim, aligning with traditional fact checking research.

Annotation Schema. We propose a five level scale aligned with existing graded resources:

Level 1 (Supported): Claim is supported by multiple independent, authoritative sources with high confidence.

Level 2 (Likely True): Claim is supported by credible evidence but with some uncertainty due to limited source diversity or domain complexity.

Level 3 (Uncertain): Evidence is insufficient, conflicting, or the claim involves legitimate scientific uncertainty or emerging situations.

Level 4 (Likely False): Claim contradicts credible evidence but with residual uncertainty.

Level 5 (Refuted): Claim is contradicted by multiple independent, authoritative sources with high confidence.

Decision Rules for Ambiguous Cases. When annotators encounter difficult cases: (a) if high quality sources genuinely conflict, assign Level 3 and document the conflict; (b) weight sources by domain expertise, independence, and track record; (c) for emerging situations where evidence is accumulating, assign Level 3 with temporal flag; (d) for claims mixing true and false components, annotate subclaims separately then aggregate; (e) document specific evidence and reasoning in justification field for all non obvious cases.

Operationalization. Implement through four steps: claim decomposition following Chen et al. (2022); evidence retrieval using FEVER style search, web APIs, or domain specific databases; NLI based entailment classification with aggregation across evidence; conformal prediction (Angelopoulos and Bates, 2021) for coverage guarantees rather than arbitrary confidence thresholds.

Output: (epistemic_level, conformal_set, confidence, evidence_ids[], justification_text).

4.2. Dimension 2: Intentionality

Intentionality concerns inferred motivation behind content creation. This dimension presents the greatest operationalization challenges.

Fundamental Limitations. Intent cannot be reliably inferred from content alone. As MacAvaney et al. (2019) note, intent is exceedingly difficult to capture algorithmically. Wang et al. (2025) document the gap between intent's prominence in platform policies and its absence from detection systems. We propose proxy based assessment estimating manipulation likelihood rather than claiming direct intent detection.

Annotation Schema. Three level proxy based assessment:

Low Manipulation Indicators: Source has established correction history, professional editorial standards, transparent authorship, no detected coordination signals.

Moderate Manipulation Indicators: Some concerning signals: elevated sensationalism markers OR mixed source credibility OR timing coincides with sensitive events OR minor coordination indicators.

High Manipulation Indicators: Multiple strong indicators: coordination with known problematic actors AND high sensationalism AND suspicious timing AND absent editorial standards AND network amplification patterns.

Proxy Validation Requirements. Source history proxies require temporal holdout validation: train on period T, evaluate on T+1 to detect strategic behavior modification. Linguistic manipulation indicators (Horne and Adali, 2017) require cross domain validation. Coordination signals require causal validation through counterfactual analysis.

Adversarial Robustness and Fairness Safeguards. For robustness: use behavioral features over easily gamed stylistic features; implement temporal monitoring; ensemble diverse proxy types. For fairness: separate false positive evaluation across source categories; exclude identity features; regular fairness audits; human review for high stakes decisions.

Output: (manipulation_level, proxy_scores[source, linguistic, coordination, timing], fairness_flags[], requires_review).

4.3. Dimension 3: Contextual Harm

We adopt FABLE (Sehat et al., 2024) as our primary harm taxonomy with concrete operationalizations.

Adopted Taxonomy. *Fragmentation* assesses social division potential via topic modeling against known divisive narratives and engagement pattern analysis. *Actionability* assesses harmful action prompting by detecting imperative constructions and claims with direct action implications. *Believability* assesses plausibility to target audiences via source presentation quality and alignment with audience prior beliefs. *Spread Likelihood* predicts viral potential using early engagement velocity and content virality features. *Exploitativeness* identifies vulnerable group targeting through linguistic markers and demographic signals.

Annotation Schema. Four tier classification:

Critical: Health or safety risk with high actionability and spread (e.g., dangerous medical advice during outbreak, incitement to imminent violence).

High: Significant harm potential in sensitive domain with moderate to high spread (e.g., election misinformation near voting).

Medium: Some harm potential but limited actionability or spread (e.g., misleading but not dangerous health claims).

Low: Minimal harm regardless of veracity (e.g., celebrity gossip, clearly satirical content).

Outcome Validation. Harm tiers should correlate with downstream impacts. Validate through: (a) expert panel consensus; (b) retrospective analysis linking content to documented harms; (c) prospective tracking of intervention effectiveness.

Output: (harm_tier, fable_scores[F, A, B, L, E], domain_category, expert_confidence).

4.4. Integration and Aggregation

The three dimensions interact in ways that determine appropriate downstream action. The integration layer is a structured aggregation procedure with explicit escalation guarantees and conflict handling rules.

Formal Aggregation Operator.

$$Risk = w_e \cdot f(E) + w_i \cdot f(I) + w_h \cdot f(H) + \lambda \cdot C(E, I, H)$$

where $f(\cdot)$ maps dimension levels to $[0, 1]$, weights w are context dependent, and $C(E, I, H)$ encodes predefined cross dimensional conflict rules. The conflict term $C(E, I, H)$ is not a learned latent interaction but a rule based adjustment that enforces escalation in predefined high risk configurations. For example, combinations such as high harm under epistemic uncertainty or weaponized true content trigger positive conflict adjustments regardless of linear score magnitude. This preserves interpretability and ensures that high consequence patterns cannot be suppressed by averaging effects.

The aggregation operator satisfies two desirable properties. First, *monotonicity*: increasing harm or manipulation cannot reduce the integrated risk score. Second, *escalation guarantees*: predefined high risk patterns always exceed minimum routing thresholds even if other dimensions are low.

Conflict Resolution Strategies.

Low Epistemic Certainty + High Harm: Escalate to human review regardless of intent signals.

High Epistemic (True) + High Manipulation + High Harm: Route to malinformation pathway requiring contextual assessment rather than factual correction.

High Epistemic (False) + Low Manipulation + Low Harm: Standard correction pathway.

Conflicting Intent Signals: When behavioral and linguistic proxies diverge, weight behavioral history features more heavily, as they are harder to game.

Uncertain Epistemic + Conflicting Evidence: Do not force resolution. Output uncertainty explicitly and route to domain expert review.

Decision Matrix.

Epistemic	Intent	Harm	Action (Risk Score)
Refuted	High	Critical	Immediate escalation ($\geq$0.9)
Refuted	High	High	Priority review ($\geq$0.8)
Refuted	Low	Critical	Expedited fact check ($\geq$0.75)
Refuted	Low	High	Standard fact check ($\geq$0.6)
Uncertain	High	Critical	Urgent investigation ($\geq$0.7)
Uncertain	Any	Medium	Monitor ($\geq$0.4)
Supported	High	Critical	Malinformation review ($\geq$0.5)
Supported	Any	Low	No action ($<$0.3)

Table 1: Decision matrix with indicative risk thresholds. Thresholds are deployment specific and may be tuned empirically.

The matrix operationalizes the aggregation logic as a routing policy rather than a purely numeric classifier. It ensures that qualitatively distinct configurations receive differentiated treatment.

Context Dependent Weighting. Aggregation weights are policy parameters reflecting institutional priorities. Fact checking organizations may set $w_e = 0.6, w_i = 0.1, w_h = 0.3$ to emphasize verification accuracy. Health misinformation contexts may set $w_e = 0.3, w_i = 0.2, w_h = 0.5$ to prioritize harm prevention. Election integrity monitoring may set $w_e = 0.3, w_i = 0.4, w_h = 0.3$ to emphasize coordination detection. Platform moderation may use near balanced weights $w_e = 0.33, w_i = 0.33, w_h = 0.34$. Weights are interpretable, externally auditable parameters rather than learned black box coefficients.

4.5. Illustrative End-to-End Example

To demonstrate operational feasibility, we provide a compact walkthrough of a hypothetical claim.

Example Claim. "Drinking high-dose vitamin C prevents COVID-19 infection."

Epistemic Status. Evidence retrieval identifies WHO and CDC guidance indicating no reliable evidence that vitamin C prevents COVID-19. Multiple authoritative sources contradict the claim. The system assigns *Level 5 (Refuted)* with a conformal prediction set $\{5\}$ at 90% coverage, indicating high epistemic confidence.

Intentionality (Proxies). Source history shows prior fact check flags and elevated sensationalism markers. Coordination signals are weak. Proxy aggregation yields *Moderate Manipulation Indicators*. Output: manipulation level = Moderate.

Contextual Harm. Actionability is high due to behavioral recommendation. Believability is moderate given medical framing. Spread likelihood is moderate. Overall harm tier: *High*, due to potential public health consequences.

Integration. Using health context weights ($w_e = 0.3$, $w_i = 0.2$, $w_h = 0.5$), the aggregation operator produces a risk score exceeding the High harm threshold. Routing decision: *Priority review and corrective labeling*.

This example illustrates three properties: epistemic refutation alone does not determine action; moderate intent proxies do not suppress escalation when harm is substantial; and the routing outcome is interpretable and traceable to dimensional inputs.

Malinformation Example. To illustrate the malinformation pathway, consider a genuine but selectively leaked internal corporate memo about product safety concerns, amplified by coordinated accounts during a competitor's product launch. *Epistemic Status:* Level 1 (Supported), the document is authentic. *Intentionality:* High Manipulation Indicators, coordination signals are strong and timing is suspicious. *Contextual Harm:* High, potential for financial harm and erosion of consumer trust through decontextualization. Under the integration logic, this triggers the malinformation pathway (true content being weaponized), routing to contextual assessment rather than factual correction. The framework correctly avoids labeling authentic content as false while still flagging the manipulative amplification pattern for review.

4.6. Risk Profile Template

For consistent, auditable downstream reporting, we propose a standardized risk profile template:

```
RISK PROFILE v1.0
Content ID: [identifier]
------------------------
EPISTEMIC: Level [1-5],
  Conformal Set [...],
  Evidence [...], Justification [...]
INTENT: Level [L/M/H],
  Scores [S,L,C,T],
  Fairness Flags [...]
HARM: Tier [C/H/M/L],
  FABLE [F,A,B,L,E], Domain [...]
------------------------
INTEGRATED: Risk [0-1],
  Action [...], Conflicts [...],
  Review Required [Y/N]
AUDIT: Status [...],
  Appeal History [...]
```

This template enables consistent reporting, auditability, appeal documentation, and systematic quality evaluation across deployment contexts.

5. Extensions

5.1. Multilingual and Cultural Considerations

Resources. MuMiN (Nielsen and McConville, 2022) covers 41 languages. CheckThat! (Nakov et al., 2022) provides resources for Arabic, Bulgarian, Dutch, Turkish, and English. X FACT (Gupta and Srikumar, 2021) spans 25 languages. MM COVID (Li et al., 2020) offers six languages with multimodal content.

Transfer Strategies. Zero shot cross lingual transfer uses multilingual encoders (mBERT, XLM R) trained on high resource languages; Panda and Levitan (2021) demonstrated feasibility though performance varies by language. Language specific fine tuning yields best performance where annotated data exists.

Cultural Adaptation. Harm assessment requires cultural calibration. We recommend region specific annotator pools with documented backgrounds, explicit documentation of cultural assumptions in guidelines, separate evaluation by language/region rather than aggregated global metrics, and qualitative analysis of cross cultural disagreement. This recommendation is reinforced by recent findings that LLM persona based approaches with geographic or demographic attributes significantly affect harm perception (Piot et al., 2025; Plaza-del-Arco et al., 2024), suggesting that cultural context must be integrated not only in human annotation but also in any LLM assisted assessment workflow, including persona simulation for hypothesis testing and guideline development.

Dimension Adaptations. For Epistemic Status, evidence retrieval must access language specific sources; Wikipedia coverage varies dramatically across languages. For Intentionality, credibility databases are primarily English centric; network coordination patterns may transfer better than linguistic markers. For Harm, topic sensitivity requires cultural adaptation.

5.2. Multimodal Considerations

Resources. MM COVID (Li et al., 2020) provides text plus images. NewsCLIPpings (Luo et al., 2021) addresses out of context images. VERITE (Papadopoulos et al., 2024) accounts for unimodal biases.

Dimension Adaptations. For Epistemic Status: image text consistency via CLIP/BLIP, reverse image search for provenance, metadata verification.

For Intentionality: deepfake and manipulation detection, metadata tampering indicators. For Harm: graphic content detection, emotional manipulation through imagery.

Integration Challenge. Multimodal misinformation often involves accurate text with misleading images or vice versa. The integration layer must handle cross modal inconsistencies, routing such cases to specialized review.

6. Evaluation Reform

We propose a reformed evaluation protocol addressing limitations of conventional metrics.

6.1. Limitations of Conventional Metrics

Standard precision, recall, and F1 fail to capture several critical quality dimensions. Regarding calibration, a system with 80% accuracy but overconfident predictions may be more dangerous than one with 75% accuracy and well calibrated confidence. Regarding selective prediction, systems recognizing their limitations and abstaining appropriately provide more value than those always producing predictions. Regarding harm weighted performance, standard metrics treat all errors equally, but false negatives on Critical content are far more consequential than on Low content. Regarding annotator disagreement, binary ground truth obscures legitimate disagreement reflecting genuine ambiguity rather than annotation error (Nie et al., 2020).

6.2. Proposed Evaluation Dimensions

We propose a multidimensional evaluation protocol aligned with the three dimensional framework.

Calibration. Beyond classification accuracy, systems must produce confidence estimates that align with empirical correctness. We measure Expected Calibration Error (ECE) and Brier scores following Xiao et al. (2022), reported separately for in domain and out of domain evaluation.

Conformal Prediction Quality. We adopt conformal prediction methods (Angelopoulos and Bates, 2021; Quach et al., 2023) providing distribution free coverage guarantees. We report empirical coverage, average prediction set size, and coverage conditional on harm tiers and demographic subgroups. Conditional coverage analysis is essential to detect systematic undercoverage on minority populations or rare claim types.

Harm Weighted Metrics. We define harm sensitive weights: Critical 4.0, High 2.0, Medium 1.0, Low 0.5. Weighted precision, recall, and F1 are computed by scaling errors according to harm tier. Reporting both weighted and unweighted metrics makes explicit the tradeoff between aggregate accuracy and consequence aware performance.

Generalization. We construct cross domain train and test splits spanning political misinformation, health claims, science controversies, and entertainment rumors. Temporal evaluation following Zhu et al. (2022) assesses entity bias by training on earlier time periods and testing on future data.

Disagreement Modeling. Binary majority vote labels obscure legitimate ambiguity. We report full annotator label distributions. Probabilistic aggregation methods such as Dawid Skene or MACE (Nie et al., 2020) jointly model annotator reliability and item difficulty. Disagreement entropy is reported per dimension to identify where guidelines require refinement versus where epistemic uncertainty is inherent. Additionally, for subjective dimensions such as harm assessment, we recommend incorporating subjectivity aware metrics such as cross replication reliability (xRR) (Wong et al., 2021), which benchmarks inter rater reliability against empirical baselines from replicated annotation rather than relying on fixed agreement thresholds. This is particularly relevant given the cultural sensitivity of harm judgments across different populations.

Human Uplift. Following Marusich et al. (2024) and critiques by Devic et al. (2025), we evaluate decision support utility rather than standalone classifier performance. Metrics include decision accuracy, time to decision, harm weighted error severity, appropriate reliance, and calibration of human confidence. Uplift is defined relative to human baseline performance.

6.3. Validation Pathway

Phase 1 (Months 1–3): Protocol Development. Develop guidelines with decision trees, worked examples, edge case specifications. Pilot 50–100 items using think aloud protocols. Deliverable: comprehensive annotation manual.

Phase 2 (Months 4–6): Pilot Study. Annotate 500–1000 items stratified across domains, 3+ annotators per item. Compute Krippendorff's alpha per dimension. Apply MACE for disagreement modeling. Incorporate subjectivity aware metrics including xRR (Wong et al., 2021) to benchmark annotation quality against replicated baselines rather than fixed thresholds. Deliverable: pilot dataset with reliability statistics.

Phase 3 (Months 7–9): Baseline Implementation. Implement dimensions using existing tools: FEVER pipeline for Epistemic Status, credibility APIs for Intent, toxicity classifiers for Harm. Implement conformal wrappers. Deliverable: benchmarks with coverage analysis.

Phase 4 (Months 10–12): Integration Evaluation. Compare integrated system against single

dimension baselines. User studies with fact checkers measuring decision quality, time, error severity. Deliverable: human uplift assessment and deployment recommendations.

7. Discussion

Relationship to Existing Systems. Our framework organizes and integrates existing capabilities rather than replacing them. FEVER style verification pipelines provide foundations for Epistemic Status. Credibility databases (NewsGuard) and coordination detection tools inform Intent Proxies. Content moderation classifiers contribute to Harm assessment.

Computational Considerations. Running three parallel analysis streams increases computational requirements. We recommend tiered deployment: fast classifiers for initial filtering, full dimensional analysis only for flagged content. Conformal prediction adds minimal overhead while providing principled uncertainty. For platform scale deployment, efficient batching and caching strategies become essential. We note that existing modular architectures in fact checking pipelines already employ sequential filtering strategies where claim detection precedes evidence retrieval, which in turn precedes verdict prediction. Our framework extends this principle by adding parallel harm and intent assessment streams that can be invoked selectively based on initial epistemic screening results.

Governance Requirements. Risk profiles require supporting infrastructure. Role definitions must specify reviewer qualifications by harm tier, with Critical tier cases requiring senior domain expertise. Appeal mechanisms must be available for content creators, with clear timelines and escalation paths. Audit procedures should assess systematic biases across source categories (mainstream vs. independent media, different geographic regions) and demographic proxies. Regular reporting on false positive rates by source type helps detect and correct systematic unfairness.

Relationship to Policy Frameworks. The framework's dimensional structure aligns with emerging regulatory approaches that distinguish between different types of harmful content and require proportionate responses. The European Digital Services Act, for instance, requires platforms to assess systemic risks including the dissemination of illegal content and the manipulation of services. Our risk profile template provides a structured mechanism for documenting assessment rationale in compliance with such transparency requirements. The explicit separation of epistemic status from harm assessment also supports contexts where factual accuracy and public interest considerations must be balanced, such as political speech during elections.

8. Conclusion

This position paper has introduced a theory grounded NLP framework for information disorder built on three explicitly integrated dimensions: Epistemic Status, Intentionality, and Contextual Harm. By moving beyond binary fake news classification, we articulate a multidimensional architecture that differentiates verification, manipulation signals, and consequence assessment, and specifies how these signals should be combined in practice.

Rather than proposing yet another classifier, we formalize cross dimensional aggregation through interpretable operators with monotonicity and escalation guarantees. We provide concrete annotation schemas with decision rules for ambiguous cases, explicit conflict resolution strategies, and a structured decision matrix translating model outputs into actionable routing policies. We operationalize contextual harm through the FABLE taxonomy with outcome validation pathways, incorporate conformal prediction for principled abstention under uncertainty, and define harm sensitive evaluation metrics aligned with real world risk, advancing a human uplift evaluation paradigm that measures decision support utility rather than standalone model accuracy.

The practical impact of this proposal depends on empirical validation, including reliable multidimensional annotation, calibrated automated estimation, and demonstrated improvement in expert decision making. We have outlined a phased validation pathway including pilot annotation studies, baseline implementation, and controlled user studies with practicing fact checkers. We offer this framework as a foundation for empirical validation and future system design.

9. Limitations

This paper proposes a conceptual framework without empirical validation, and several limitations require acknowledgment.

First, the proposed annotation schemas require pilot studies to establish inter annotator reliability and to characterize systematic disagreement. Without such studies, consistency and feasibility remain unverified.

Second, computational feasibility at platform scale remains undemonstrated. Running parallel dimensional analyses with uncertainty quantification may impose higher costs than single classifier pipelines, and efficient deployment strategies must be empirically evaluated.

Third, aggregation weights, risk thresholds, and harm tier boundaries require empirical tuning for

specific institutional contexts. The indicative values presented here are theoretically motivated but not validated in operational settings.

Fourth, the framework assumes access to metadata such as source history, coordination signals, and propagation patterns, and it requires cultural calibration for multilingual deployment. In addition, intent inference remains fundamentally limited: proxy based approaches provide correlational rather than causal evidence and must be interpreted cautiously. We view these limitations as a structured research agenda requiring phased empirical investigation rather than as defects in the conceptual design.

Ethical Considerations

Automated information disorder detection has profound implications for freedom of expression, privacy, and democratic governance. Our framework addresses these concerns through several design choices.

First, we emphasize triage based approaches that flag content for human review rather than automated removal. Second, explicit uncertainty communication acknowledges system limitations through conformal prediction and risk profile confidence levels. Third, structured outputs support explainability and appeal: the dimensional breakdown makes assessment reasoning transparent and contestable. Fourth, human in the loop integration ensures consequential decisions involve appropriate oversight.

Responsible deployment requires diverse annotator representation during guideline development, regular fairness audits for credibility and intent proxy components, appeal mechanisms for content creators, and ongoing monitoring for emergent biases. We recommend partnership with established fact checking organizations and civil society groups to ensure appropriate governance structures.

Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

10. Bibliographical References

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., and Zaghouani, W. (2021). Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of EMNLP 2021*, pp. 611–649.

Alam, F., Biswas, M. R., Shah, U., Zaghouani, W., and Mikros, G. (2024). Propaganda to hate: A multimodal analysis of Arabic memes with multi-agent LLMs. In *International Conference on Web Information Systems Engineering*, pp. 380–390. Springer.

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Banko, M., MacKeen, B., and Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 125–137.

Sehat, C. M., Li, R., Nie, P., Prabhakar, T., and Zhang, A. X. (2024). Misinformation as a harm: Structured approaches for fact-checking prioritization. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Chen, J., Wang, X., Gao, S., Jiang, Y., and Shi, L. (2022). Generating literal and implied subquestions to fact check complex claims. In *Proceedings of EMNLP*, pp. 3495–3508.

Desai, S. and Durrett, G. (2020). Calibration of pre trained transformers. In *Proceedings of EMNLP*, pp. 295–302.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact checking. *TACL*, 10:178–206.

Gupta, A. and Srikumar, V. (2021). X-FACT: A new benchmark dataset for multilingual fact checking. In *Proceedings of ACL*, pp. 675–686.

Hardalov, M., Arora, A., Nakov, P., and Augenstein, I. (2022). A survey on stance detection for mis and disinformation identification. In *Findings of NAACL*, pp. 1259–1277.

Hasanain, M., Alam, F., Mubarak, H., Abdaljalil, S., Zaghouani, W., Nakov, P., Da San Martino, G., and Freihat, A. A. (2023). ArAIEval

shared task: Persuasion techniques and disinformation detection in Arabic text. *arXiv preprint arXiv:2311.03179*.

Hasanain, M., Suwaileh, R., Weering, S., Li, C., Caselli, T., Zaghouani, W., Barrón-Cedeño, A., Nakov, P., and Alam, F. (2024). Overview of the CLEF-2024 CheckThat! lab task 1 on checkworthiness estimation of multigenre content. In *Working Notes of CLEF 2024*, pp. 276–286.

Horne, B. D. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content. In *Proceedings of ICWSM*, 11(1):759–766.

Devic, S., Srinivasan, T., Thomason, J., Neiswanger, W., and Sharan, V. (2025). From calibration to collaboration: LLM uncertainty quantification should be more human-centered. *arXiv preprint arXiv:2506.07461*.

Li, Y., Jiang, B., Shu, K., and Liu, H. (2020). MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Luo, G., Darrell, T., and Rohrbach, A. (2021). NewsCLIPpings: Automatic generation of out of context multimodal media. In *Proceedings of EMNLP*, pp. 6801–6817.

MacAvaney, S., et al. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Marusich, L., Bakdash, J. Z., Zhou, Y., and Kantarcioglu, M. (2024). Using AI uncertainty quantification to improve human decision making. In *Proceedings of ICML*, pp. 34949–34960.

Nakov, P., et al. (2021). Automated fact checking for assisting human fact checkers. In *Proceedings of IJCAI*, pp. 4551–4558.

Nakov, P., et al. (2022). Overview of the CLEF 2022 CheckThat! lab. In *CLEF 2022*, pp. 495–520.

Nie, Y., Zhou, X., and Bansal, M. (2020). What can we learn from collective human opinions on natural language inference data? In *Proceedings of EMNLP*, pp. 9131–9143.

Nielsen, D. S. and McConville, R. (2022). MuMiN: A large scale multilingual multimodal fact checked misinformation social network dataset. In *Proceedings of SIGIR*, pp. 3141–3153.

Panchendrarajan, R. and Zubiaga, A. (2024). Claim detection for automated fact checking: A survey. *Natural Language Processing Journal*, 7:100066.

Panda, S. and Levitan, S. I. (2021). Detecting multilingual COVID-19 misinformation via contextualized embeddings. In *Proceedings of NLP4IF Workshop*, pp. 125–129.

Papadopoulos, S.-I., Koutlis, C., Papadopoulos, S., and Petrantonakis, P. C. (2024). VERITE: A robust benchmark for multimodal misinformation detection accounting for unimodal bias. *Int. J. Multimedia Information Retrieval*, 13:4.

Piot, P., Martín-Rodilla, P., and Parapar, J. (2025). Personalisation or prejudice? Addressing geographic bias in hate speech detection using debias tuning in large language models. *arXiv preprint arXiv:2505.02252*.

Plaza-del-Arco, F. M., Cercas Curry, A., Paoli, S., Cercas Curry, A., and Hovy, D. (2024). Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In *Findings of EMNLP 2024*, pp. 4346–4366.

Prabhudesai, S., et al. (2023). Understanding uncertainty: How lay decision makers perceive uncertainty in human AI decision making. In *Proceedings of IUI*, pp. 379–396.

Quach, V., Fisch, A., Schuster, T., et al. (2023). Conformal language modeling. In *Proceedings of ICLR*.

Scheuerman, M. K., Jiang, J. A., Fiesler, C., and Brubaker, J. R. (2021). A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2):1–33.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: A large scale dataset for fact extraction and verification. In *Proceedings of NAACL HLT*, pp. 809–819.

Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. In *Proceedings of ACL*, pp. 422–426.

Wang, S., et al. (2025). The unappreciated role of intent in algorithmic moderation. *Harvard Kennedy School Misinformation Review*, 6(3).

Wardle, C. and Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework*. Council of Europe Report DGI(2017)09.

Wong, K., Paritosh, P., and Aroyo, L. (2021). Cross-replication reliability: An empirical approach to interpreting inter-rater reliability. In *Proceedings of ACL-IJCNLP*, pp. 7053–7065.

Xiao, Y., et al. (2022). Uncertainty quantification with pre trained language models: A large scale empirical analysis. In *Findings of EMNLP*, pp. 7273–7284.

Xin, J., Tang, R., Yu, Y., and Lin, J. (2021). The art of abstention: Selective prediction and error regularization for NLP. In *Proceedings of ACL*, pp. 1040–1051.

Zaghouani, W., Jarrar, M., Habash, N., Bouamor, H., Zitouni, I., Diab, M., El-Beltagy, S. R., and AbuOdeh, M. (2024). The FIGNEWS shared task on news media narratives. *arXiv preprint arXiv:2407.18147*.

Zhu, Y., et al. (2022). Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of SIGIR*, pp. 2120–2125.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2):1–36.

Emotion and Information Disorder in NLP: A Systematic Mapping and Benchmark Blueprint

Renatha Vieira[1], **Alvaro Figueira**[1,2]
[1]Faculty of Sciences, University of Porto, Porto, Portugal
[2]INESC TEC, Porto, Portugal
{vieirasrept@gmail.com, arfiguei@fc.up.pt}

Abstract

Online misinformation research in NLP has expanded rapidly, including approaches that model affective signals such as sentiment, discrete emotions, and emotion dynamics. However, the Information Disorder framework distinguishes misinformation, disinformation, and mal-information along dimensions of intention, harm, and contextual dependence, which are rarely operationalised in current datasets, tasks, and evaluation protocols. We provide a systematic mapping of 82 studies at the intersection of Information Disorder and emotion-aware NLP (51 model papers, 7 dataset papers, 24 survey/theory papers). Across empirical works (58), veracity-centric supervision dominates (72.4% binary labels), while explicit intention and harm variables appear in only 1.7% each. Evaluation relies mostly on random splits (79.3%), limiting robustness to source and temporal shifts. Emotion is represented in 43.1% of model papers, mostly as static features, with emotion dynamics and audience emotion rare. Based on these findings, we propose an operational taxonomy aligned with Information Disorder and a benchmark blueprint specifying tasks, annotation variables, split strategies, and evaluation protocols to support theory-grounded, comparable progress.

Keywords: misinformation, disinformation, mal-information, emotion, systematic mapping, evaluation blueprint.

1. Introduction

Information Disorder is a framework proposed to organise research on false and harmful information by distinguishing *misinformation*, *disinformation*, and *mal-information*, and by emphasising contextual factors such as intention and potential harm (Wardle and Derakhshan, 2017; Wardle, 2020; Turcilo and Obrenovic, 2020). In parallel, NLP research on misinformation and fake news has expanded rapidly, supported by benchmark datasets and supervised tasks that most often operationalise the problem as veracity classification or related variants of credibility assessment (Shu et al., 2017; Zhou and Zafarani, 2018; Wu et al., 2019; Wang, 2017).

Affective signals (sentiment, discrete emotions, and emotion dynamics) have been increasingly incorporated into such systems, motivated by the intuition that manipulation and engagement are partly mediated through emotion and that diffusion dynamics can amplify emotionally charged misinformation (Vosoughi et al., 2018; Ghanem et al., 2019; Giahanou et al., 2019). However, emotion-aware approaches are typically trained and evaluated against labels that do not explicitly encode intention, harm, or contextual dependence. As a result, systems may achieve strong performance on coarse veracity targets while remaining theoretically misaligned with Information Disorder distinctions, a gap that is particularly salient for *mal-information*, where harm can arise even when content is factually correct (Wardle and Derakhshan, 2017).

We address this misalignment with a systematic mapping of emotion-aware NLP work through an Information Disorder lens and provide reusable artefacts to support comparability. Specifically, we (i) compile and code a curated corpus of studies using a shared coding scheme that captures dataset properties, task formulations, evaluation protocols, and emotion operationalisations along Information Disorder dimensions (Sections 2–3); (ii) synthesise recurrent structural mismatches between common operational settings and Information Disorder constructs (Section 4); and (iii) propose an operational taxonomy and a benchmark blueprint to guide dataset creation and evaluation in a theory-grounded, replicable manner (Section 5).

Research questions. The mapping is guided by the following questions:

1. **RQ1:** Which Information Disorder dimensions (intention, harm, contextual dependence) are explicitly annotated or otherwise operationalised in current datasets and tasks?
2. **RQ2:** How is emotion represented (sentiment, discrete emotions, appraisal, dynamics) and how is it used (feature, auxiliary task, audience response, discourse structure) in NLP approaches to false information?
3. **RQ3:** What evaluation protocols (splits, metrics, robustness tests) are most common, and where do they fail to support Information Disorder-relevant claims?
4. **RQ4:** What minimal benchmark design would enable theory-aligned, replicable comparison of

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 77–84
12 May 2026.

systems for Information Disorder?

The remaining sections of this paper are organised as follows: Section 2 presents the selection protocol and coding scheme; Section 3 reports mapping results; Section 4 summarises misalignments; and Section 5 introduces the taxonomy and benchmark blueprint.

2. Methodology of Reference Selection

The bibliographic corpus was constructed through a structured but intentionally selective mapping process. Our aim was not to exhaustively cover all fake news detection research, but to assemble a conceptually coherent set of studies that makes it possible to analyse how emotion is operationalised in relation to the Information Disorder framework.

The search was conducted between 10 February and 2 March 2026. Semantic Scholar was used as the primary retrieval source, complemented by citation chaining from high-relevance papers and targeted Google Scholar follow-up searches used to close thematic gaps, especially around dataset construction and emotion-aware fake news detection. We explored three query families. Q1 used broad Information Disorder terms (*misinformation, disinformation, malinformation, fake news, information disorder*); Q2 combined the same group with affect terms (*emotion, affect, sentiment*) and NLP terms (*natural language processing, NLP, text classification*); and Q3 flattened these groups into a broader disjunctive query. Query 3 retrieved 254 records and was retained as the main starting set because it offered the best balance between breadth and conceptual relevance.

Selection proceeded in stages. A title-based screening reduced the 254 retrieved records to 60 candidate papers by excluding work not directly related to misinformation/fake news, not connected to affect or emotion, or outside the scope of NLP. Abstracts were then inspected when needed to assess thematic fit with the goals of the study, especially relevance to taxonomy, benchmarking, or methodological grounding, yielding a core set of 43 papers. The corpus was subsequently expanded through backward citation tracking and targeted follow-up searches, producing the final set of 82 included studies.

Inclusion prioritised studies on emotion recognition, sentiment analysis, or affective computing applied to misinformation/disinformation, together with adjacent work on datasets, propagation, contextual signals, and conceptual distinctions that was necessary for interpreting evaluation and taxonomy choices. We excluded generic text classification work unrelated to deceptive information, non-NLP approaches, and domain- or language-specific studies with no clear relevance to the research focus. Title/abstract screening, full-text assessment, and final coding were conducted by the first author. No duplicate records were found during consolidation of the candidate set, and no independent double-coding or formal inter-rater agreement was performed.

2.1. Selection transparency and reproducibility

Our protocol is therefore iterative rather than fully exhaustive: 254 records were retrieved from the main query, 60 were retained after title-based screening, 43 formed the initial thematic core, and the final corpus reached 82 studies after snowballing and targeted complementary searches. Each study was then coded using the shared variables in Table 1. The coded spreadsheet and search strings are available upon request.

2.2. Coding scheme for systematic mapping

Table 1 summarises the coding scheme used to extract comparable metadata from each study/resource.

3. Systematic Mapping Results

This section reports the mapping results using the shared coding scheme (Table 1). We first focus on dataset construction and annotation practices, with emphasis on how intention, harm, and context dependence are represented (RQ1). We then summarise task families and evaluation protocols (RQ3), and finally map how emotion is represented and integrated into detection pipelines (RQ2). Throughout, we use compact tables to support cross-study comparison and to make the link between evidence and the proposed benchmark blueprint explicit.

Corpus overview. The mapped corpus contains 82 studies: 51 model papers, 7 dataset papers, and 24 survey/theory papers. The literature is overwhelmingly English (80/82), with only two multilingual studies. We treat dataset and model papers as empirical works (58) when reporting the operational coverage of Information Disorder variables.

Table 2 summarises two key quantitative findings that drive the rest of the paper: (i) intention and harm are rarely operationalised beyond proxies, and (ii) evaluation relies predominantly on random splits, limiting the validity of claims about generalisation.

Field	Values / notes
Resource type	dataset, benchmark, model paper, survey, tool
Languages	ISO codes; monolingual vs multilingual
Unit of analysis	claim, article, post, thread, user, network
Label type	veracity (binary/ordinal/multi-class), stance, entailment, intent, harm, context-dependence
InDor mapping	misinformation / disinformation / mal-information; or "not operationalised"
Intention	explicit label / proxy / not available
Harm	explicit type(s) / proxy / not available
Context dependence	low/medium/high; required metadata
Emotion representation	sentiment polarity; discrete emotions; appraisal; emotion dynamics; audience emotion
Emotion usage	feature; auxiliary task; analysis-only; propagation signal; discourse structure
Evaluation protocol	random split; source split; temporal split; cross-domain; cross-lingual; robustness
Metrics	macro-F1, AUC, calibration, class-wise recall, etc.
Availability	licence; link; documentation quality; ethical notes

Table 1: Fields used in the structured coding scheme for the systematic mapping.

InDor variable coverage (N=58)			
Var.	**Exp.**	**Proxy**	**None**
Intention	1	22	35
Harm	1	7	50
Context dep.	17	7	34

Evaluation split strategy (#)	
Split	**#**
Random	46
Temporal	5
Cross-domain	3
Cross-lingual	1
Other/none	3

Table 2: Mapping summary for empirical studies (datasets+models, N=58): coverage of Information Disorder variables and evaluation split strategies.

3.1. Current landscape: datasets and annotation practices

To address **RQ1**, we coded how datasets and benchmarks define their label space and which Information Disorder variables (intention, harm, context dependence) are explicitly annotated versus approximated via proxies.

In general, four main construction strategies can be observed. The first is manual curation by specialists, as in LIAR, the COVID-19 Fake News Dataset, and the PolitiFact-Oslo Corpus, where statements or full texts are individually verified by editors or fact-checkers (Wang, 2017; Patwa et al., 2020; Põldvere et al., 2023). The second consists of structured extraction from fact-checking platforms such as PolitiFact and GossipCop, as in FakeNewsNet (Shu et al., 2018). The third strategy is large-scale distant supervision, exemplified by Fakeddit, where labels are automatically inferred based on the nature of the originating subreddit (Nakamura et al., 2020). The fourth approach is theory-guided synthetic generation, as in MegaFake, which uses language models under the so-called LLM-Fake Theory to produce intentional disinformation grounded in principles of social psychology (Wang et al., 2024).

Despite this methodological diversity, most datasets are primarily structured around a criterion of factual veracity. The central organising axis is to determine whether content is true or false, with varying degrees of granularity, but rarely does dataset construction begin from an explicit modelling of communicative intention, type of harm, or contextual dependence.

Although some resources include contextual elements, such as author profiles, media type, network structure, or structured claim–evidence pairs as in Factify 2 (Suryavardan et al., 2023), context appears predominantly as auxiliary data rather than as a structured interpretative dimension. In most cases, there is no explicit annotation of communicative intention, potential harm, or degree of situational dependence.

Across the seven dataset papers (Table 3), intention is never explicitly labelled (5/7 use proxies; 2/7 omit it), and harm is at best proxied in 2/7 resources. Contextual dependence is more often present as structured metadata (explicit in 5/7), but it is rarely surfaced as a variable to be predicted or evaluated. In the next subsection, we examine how these resources are translated into supervised tasks and evaluation protocols.

3.2. Task formulations and evaluation protocols

To address **RQ3**, we grouped studies by task family and by evaluation protocol (split strategy, robustness tests, and metrics). This clarifies which kinds

Resource	Lang.	Unit	Veracity labels	Intent	Harm	Notes
LIAR (Wang, 2017)	EN	claim	6-point	none	none	speaker metadata; veracity-centric
FakeNewsNet (Shu et al., 2018)	EN	article+ social	binary	proxy	proxy	propagation metadata; social context
Fakeddit (Nakamura et al., 2020)	EN	post+ image	2/3/6-way	proxy	none	distant supervision via subreddit
COVID-19 Fake News (Patwa et al., 2020)	EN	post	binary	none	none	health domain; veracity labels
Factify 2 (Suryavardan et al., 2023)	EN	multi-modal	multi-class	proxy	none	multimodal evidence; includes satire
PolitiFact-Oslo (Põldvere et al., 2023)	EN	article	multi-class	proxy	none	full-text corpus; excludes satire
MegaFake (Wang et al., 2024)	EN	article	binary (synthetic)	proxy	proxy	theory-guided LLM generation

Table 3: Datasets/resources in the mapped corpus (N=7) and their alignment with Information Disorder variables (intent and harm coded as explicit/proxy/none).

of claims are empirically supported by a given setup and which Information Disorder dimensions remain outside the supervised target.

In our empirical subset (N=58), supervision remains overwhelmingly veracity-centric: 42 studies use binary veracity labels and 15 use multi-class veracity labels, with only one stance-based formulation. This reinforces a common conflation between Information Disorder categories and truth labels.

Evaluation protocols further constrain Information Disorder-relevant claims. As summarised in Table 2, 46/58 empirical studies rely on random splits; only 5 use temporal splits and 3 use cross-domain evaluation, and just one reports cross-lingual evaluation. Reported metrics concentrate on macro-F1 (35/58) and accuracy (20/58), while calibration and harm-aware error analysis are rare.

These patterns matter for interpreting emotion-aware work: improvements on veracity benchmarks do not necessarily imply better modelling of intention or harm.

3.3. Emotion in false information NLP

To address **RQ2**, we coded how affective information is represented (polarity, discrete emotions, transitions/dynamics, audience response) and how it is used (features, auxiliary objectives, analysis-only, or social/propagation signals).

Across the full corpus (N=82), 28 studies include an explicit emotion component (emotion representation other than *none*). In the model subset (N=51), 22 papers (43.1%) incorporate emotion: most rely on discrete emotion categories (15/51) or sentiment polarity (4/51), while emotion dynamics is rare (2/51) and audience emotion appears in one model. Emotion is most commonly used as feature augmentation (21/51), with fewer analysis-only studies (6/51) and only one discourse-structured approach.

Representative examples include feature-based emotion fusion for detection (Guo et al., 2023; Liu et al., 2024; Zhang et al., 2021), and dynamic modelling of emotion transitions as a marker of manipulation (Vieira and Figueira, 2025; Bian and Zhang, 2024). Crucially, among the 22 emotion-aware model papers, none provides explicit intention or harm labels, and harm is proxied in only 5 cases. This supports the broader pattern that emotion is mostly deployed to improve veracity discrimination rather than to operationalise Information Disorder dimensions.

Emotion repr.	#	Emotion usage	#
None	29	None	23
Discrete	15	Feature	21
Sentiment	4	Analysis-only	6
Dynamics	2	Discourse	1
Audience	1		

Table 4: Emotion operationalisation among model papers (N=51): representation and usage categories.

Overall, emotion is treated as a meaningful component of persuasion and engagement, but its evaluation remains coupled to veracity labels. The next section consolidates how this coupling, together with missing intention/harm/context variables, leads to systematic misalignment with the Information Disorder framework.

4. Structural Misalignment with Information Disorder

The Information Disorder framework distinguishes misinformation, disinformation, and mal-information primarily along intention, harm, and context dependence (Wardle and Derakhshan, 2017). However, the mapping in Section 3 indicates that the operational space of most NLP work remains veracity-centred. The problem is not only terminological: when labels collapse truth, communicative intent, and harm into a single target, improvements in model performance become difficult to interpret from an Information Disorder perspective. This matters especially for emotion-aware systems, whose affective cues may track engagement or rhetorical style without resolving whether content is accidental, strategic, or harmful.

The mismatch can be read at three levels. At the label level, veracity targets collapse communicative situations that the Information Disorder framework treats as distinct. At the feature level, context and emotion are often present as metadata or auxiliary signals but are not tied to explicit claims about intention or harm. At the evaluation level, random partitions reward within-source regularities and make it harder to know whether a model will generalise to new actors, events, or time periods.

Table 5 summarises recurrent gaps, why they matter for Information Disorder claims, and what a minimal benchmark should encode to address them.

Mapping evidence. Table 2 shows that intention is absent in 35/58 empirical studies and explicitly labelled in only one; harm is absent in 50/58 and explicitly labelled in only one; and contextual dependence is explicit in 17/58 but still missing in 34/58. Moreover, 23/82 model papers frame their target as *disinformation*, yet none includes explicit intention or harm variables, so the label is often used as a synonym for falsity. Finally, random splits dominate (46/58), increasing the risk of leakage and overestimating robustness under source and temporal shifts.

This misalignment matters because claims about "disinformation" are sometimes made from experimental setups that cannot distinguish deliberate from accidental falsehoods, or harmful reframing of true information. Consequently, progress in model accuracy does not necessarily translate into progress in theory-grounded Information Disorder detection. Put differently, the field already has useful predictive signals, but too few task formulations that bind those signals to the distinctions the framework actually cares about. The next section translates these gaps into concrete annotation variables, task definitions, and evaluation protocols.

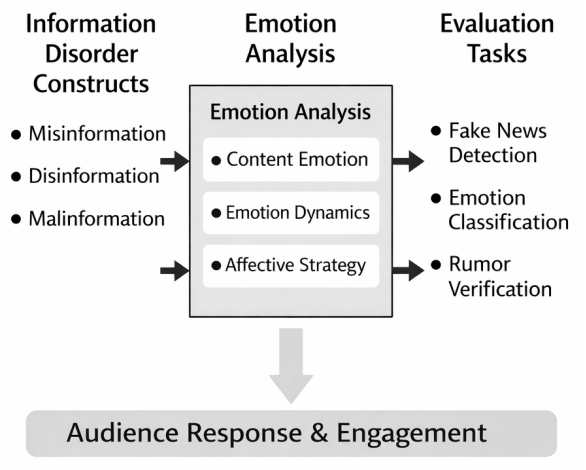


Figure 1: From Information Disorder theory to operational datasets, tasks, and evaluation: proposed bridge for emotion-aware NLP.

5. Proposed Direction

5.1. Operational taxonomy aligned with Information Disorder

We propose a minimal operational taxonomy that decomposes each instance into annotatable variables, rather than treating veracity as a proxy for intention or harm. Variables that are inherently subjective should allow an explicit *unknown* value and, where feasible, an annotator confidence score. This is particularly important for intention and harm, which can vary across cultures and contexts. Rather than forcing annotators to assign a single Information Disorder category from the outset, the taxonomy separates the underlying variables that make those categories interpretable. In practice, this makes it easier to document uncertainty and to distinguish cases that are false but strategically ambiguous from cases that are factually correct yet potentially harmful. The proposed dimensions and their suggested label sets are summarised in Table 6.

Two design principles follow from this decomposition. First, variables that can often be grounded more directly in available evidence, such as veracity, should remain distinct from variables that depend more heavily on pragmatic interpretation, such as intention and harm. Second, context

InDor dimension	Common operationalisation	Failure mode for InDor claims	Benchmark remedy
Intention	absent; inferred via source-level proxies; conflated with falsity	cannot distinguish misinformation (unintentional) from disinformation (strategic)	annotate intention (allow "unknown" + confidence); avoid source-only proxies
Harm	rarely labelled; sometimes implicit via topic (e.g., health)	models optimise for falsity, not for downstream risk or impact	add harm type/severity labels; report harm-aware metrics and error analysis
Context dependence	metadata present but not encoded as a variable; random splits	unclear when context is required; risk of leakage and brittle generalisation	include context-dependence labels; prefer source/temporal splits; report context ablations
Mal-information	typically invisible under veracity labels	harmful true content is misclassified or ignored	decouple truth and harm; include "true but harmful" cases and evaluate separately
Emotion	used as feature amplifier for veracity tasks	affective cues do not map to intent/harm; risk of overfitting emotional stereotypes	define affective strategy labels (incl. dynamics); evaluate robustness across domains and cultures

Table 5: Recurrent misalignments between Information Disorder constructs and common NLP operationalisations, with suggested benchmark remedies.

should be encoded not only as metadata but also as a judgement about how much extra information is required to interpret the case. This makes it possible to compare text-only systems with models that rely on thread, source, or temporal cues, and to report more clearly where uncertainty actually enters the pipeline.

Dimension	Suggested labels / annotation guidance
Veracity	true / false / mixed / unknown (claim- or post-level)
Intention	unintentional / deceptive / unknown (+ confidence)
Harm	harm type(s) (e.g., health, financial, reputational, civic) + severity (low/med/high) + unknown
Context dependence	low/medium/high; record minimal required context (thread, source, time)
Affective strategy	emotion cues (sentiment or discrete emotions), dynamics (transitions), and/or audience emotion
InDor category	derived from veracity+intention+harm when possible; allow "unknown"

Table 6: Operational taxonomy aligned with Information Disorder: annotatable variables and suggested label sets.

Treating the InDor category as a derived label, rather than as the only target to annotate directly, has two practical advantages. It makes annotation decisions more transparent by separating factuality, intention, and harm instead of collapsing them into a single judgement, and it allows uncertainty to remain explicit when one of those dimensions cannot be established reliably.

5.2. Benchmark blueprint

The blueprint in Table 7 translates the remedies in Table 5 into evaluation-ready tasks. It directly targets the main gaps observed in Table 2: scarce intention/harm labels, widespread random splits, and limited modelling of context dependence. The four tasks play complementary roles: T1 turns Information Disorder into a multi-target prediction problem, T2 tests how much contextual information is actually required, T3 isolates affective strategy as an object of analysis rather than a mere auxiliary feature, and T4 checks whether reported gains survive topic, source, and temporal shift.

In practice, T1 should be treated as a structured multi-label setting rather than as a single coarse classification task: some resources may support veracity and context labels but provide only uncertain evidence for intention or harm, so the benchmark should allow partial supervision and report performance per dimension in addition to any derived InDor category. T2 makes context dependence measurable rather than assumed by revealing source, thread, or temporal information incrementally and tracking when predictions stabilise. T3 separates affective strategy from coarse veracity prediction, enabling direct comparison between static emotion features, sequence-aware emotion dynamics, and audience-response signals. T4 acts as a shared stress test across topic, source, language, and time, so that gains are not interpreted only within convenient random partitions.

Task	Input	Output labels	Recommended evaluation
T1: InDor classification	post/thread + minimal context (source, time, conversation)	InDor category + intention + harm + context dependence	source and temporal splits; macro-F1 + calibration; error analysis by harm type
T2: Context dependence	post + progressively revealed context	low/medium/high (or ordinal) context dependence	context ablation curves; report minimal context for stable prediction
T3: Affective strategy	text segments and/or audience reactions	affective strategy type; emotion dynamics (transitions)	cross-domain and cross-lingual robustness; spurious correlation checks
T4: Robustness suite	any of the above	n/a	cross-topic, cross-source, temporal generalisation; leakage checks

Table 7: Benchmark blueprint at a glance: tasks, labels, and evaluation protocols.

Minimal baselines should include (i) text-only models, (ii) text+context metadata models, (iii) text+emotion feature models, and (iv) joint multi-task models predicting the full label set in Table 7. Reporting should go beyond macro-F1 by including calibration and error analysis stratified by harm type and context dependence. For corpora with derived InDor categories, reports should also state explicitly which variables were directly annotated and which were inferred by rule, since this affects both label reliability and comparability across datasets. Benchmark documentation should also record which context fields are available at training and test time, and whether emotion variables come from content, sequential dynamics, or audience response. This makes cross-study comparisons easier to interpret and reduces the risk of attributing gains to affect when they may actually come from metadata or split artefacts.

6. Conclusion

This paper argues that current emotion-aware NLP research on false information is often operationally misaligned with the Information Disorder framework. Through a systematic mapping of 82 studies, we showed that empirical work remains largely veracity-centred: explicit intention and harm labels are almost absent, context dependence is rarely operationalised as a variable, and evaluation relies mostly on random splits that can overestimate robustness. In response, we proposed a minimal operational taxonomy and a benchmark blueprint that make intention, harm, and context dependence first-class targets and that evaluate affective modelling in a way that supports theory-grounded, comparable progress.

Limitations. This mapping is iterative rather than exhaustive, and the resulting corpus may under-represent work outside the main venues indexed by our search sources. The literature we coded is strongly English-dominated (80/82), so cross-lingual generalisations remain limited. Screening and coding were conducted by the first author, without independent double-coding or a formal inter-rater agreement measure; this improves procedural clarity but remains a limitation of the present study. Coding also necessarily simplifies heterogeneous papers into shared categories, so some nuance (e.g., fine-grained task variants) is inevitably lost. Finally, the benchmark blueprint is conceptual and requires new datasets with explicit intention and harm annotation to be fully validated.

Ethics statement. Research on disinformation can be dual-use. Our work is a meta-level mapping and benchmark proposal; it does not release new disinformation content. For future datasets following the proposed blueprint, we recommend minimising the redistribution of harmful material (e.g., sharing identifiers or short excerpts when possible), documenting annotation guidelines and potential biases, and respecting privacy and platform terms when collecting social media data. The intended impact is to support more transparent, theory-grounded evaluation of systems that mitigate information disorder.

Acknowledgements

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025).

7. Bibliographical References

Haodong Bian and Lisheng Zhang. 2024. Fake news detection incorporating emotion transition in text. *Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning*.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2019. An emotional analysis of false information

in social media and news articles. *ACM Transactions on Internet Technology (TOIT)*, 20:1 – 18.

Anastasia Giahanou, Paolo Rosso, and Fabio A. Crestani. 2019. Leveraging emotional signals for credibility detection. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Quanjiang Guo, Zhao Kang, Ling Tian, and Zhouguo Chen. 2023. Tiefake: Title-text similarity and emotion-aware fake news detection. *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Fei Liu, Xinsheng Zhang, and Qi Liu. 2024. An emotion-aware approach for fake news detection. *IEEE Transactions on Computational Social Systems*, 11:3516–3524.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *International Conference on Language Resources and Evaluation*.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an infodemic: Covid-19 fake news dataset. *ArXiv*, abs/2011.03327.

Nele Põldvere, Md. Zia Uddin, and Aleena Thomas. 2023. The politifact-oslo corpus: A new dataset for fake news analysis and detection. *Inf.*, 14:627.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8:171 – 188.

Kai Shu, Amy Lynn Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ArXiv*, abs/1708.01967.

S Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya N. Reganti, Aman Chadha, Amitava Das, Amit P. Sheth, Manoj Kumar Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Factify 2: A multimodal fake news and satire news dataset. *ArXiv*, abs/2304.03897.

Lejla Turcilo and Mladen Obrenovic. 2020. Misinformation, disinformation, malinformation: Causes, trends, and their influence on democracy. E-paper series “a companion to democracy”, no. 3, Heinrich Böll Foundation. Publicado em agosto de 2020.

Renatha Souza Vieira and Álvaro Figueira. 2025. Emotional sequencing as a marker of manipulation in social media disinformation. *Future Internet*, 17(12).

Soroush Vosoughi, Deb K. Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359:1146 – 1151.

Lionel Z. Wang, Yiming Ma, Renfei Gao, Beichen Guo, Zhuoran Li, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Boris Ng. 2024. Megafake: A theory-driven dataset of fake news generated by large language models. *ArXiv*, abs/2408.11871.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Annual Meeting of the Association for Computational Linguistics*.

Claire Wardle. 2020. Understanding information disorder. First Draft News (long-form article). Accessed 8 June 2025.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Technical Report DGI(2017)09, Council of Europe. Council of Europe report, published 27 Sep 2017.

Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor.*, 21:80–90.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, L. Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. *Proceedings of the Web Conference 2021*.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *ArXiv*, abs/1812.00315.

A Multilingual Linguistic Analysis of Human vs LLM-Generated News in a Disinformation Context

Silvia Gargova[1,3], **Alba Pérez-Montero**[2], **Elena Lloret**[3], **Paloma Moreda**[3]
[1]Big Data for Smart Society Institute (GATE),
[2]University Institute for Computing Research (IUII), University of Alicante
[3]Dept. of Software and Computing Systems, University of Alicante
silvia.gargova@gate-ai.eu, alba.perezm@ua.es, {elloret, moreda}@dlsi.ua.es

Abstract

The rise of Large Language Models has shifted the Information Disorder landscape toward automated threats. This study investigates the linguistic construction of synthetic news by comparing GPT-5, Gemini 2.5, and Grok 4 across English, Spanish, and Bulgarian. Using multilingual human-authored verified news and disinformation as seeds, we analyze how prompt informativeness and model architecture influence deceptive content production. Our methodology employs five metrics: semantic similarity, factual consistency, readability, lexical richness, and persuasion technique frequency. Our analysis reveals that while prompt scarcity leads to informational loss, LLMs maintain a homogenized stylistic template regardless of input length. Unlike human authors, who intensify rhetorical and emotional markers to drive deceptive intent, LLMs adhere to a neutral register. This study identifies distinct statistical patterns in generated content characterized by hyper-standardized readability and high lexical density ($p < 0.001$). These features serve as robust "LLM signatures", enabling a classification accuracy of 96% across English, Spanish, and Bulgarian. These findings suggest that generated disinformation relies on invariant syntactic structures rather than nuanced human rhetoric, providing a framework for detection tools centered on structural patterns rather than content veracity.

Keywords: Information Disorder, Natural Language Generation, Large Language Models, Summarization, Cybersecurity

1. Introduction

The disclosure of Large Language Models (LLMs) has fundamentally shifted the landscape of information integrity, presenting both unprecedented opportunities and significant risks. While LLMs excel at tasks like summarization and content creation, their sophisticated Natural Language Generation (NLG) capabilities have been readily exploited to produce highly convincing, large-scale disinformation (Park and Nan, 2025). This development exacerbates the crisis of Information Disorder (Wardle and Derakhshan, 2017) by moving the focus from manually crafted falsehoods to automated, systemic threats (Vykopal et al., 2023). Consequently, the rapid generation of unreliable information through these models poses a critical new challenge to cybersecurity and information ecosystems worldwide.

Traditional disinformation campaigns relied on human effort, making them slow and costly. LLMs erase these limitations, enabling the rapid construction of coherent, contextually relevant, and potentially multilingual disinformation. The risk is magnified because LLM-Generated content often exhibits high linguistic quality, making it difficult for both human users and automated tools to distinguish from genuine journalism (Su et al., 2023). Consequently, understanding the specific linguistic signature of LLM-Generated disinformation is essential to developing effective detection mechanisms.

Our study contributes to this area by focusing on the security implications of LLM-Generated texts. Unlike previous work that primarily focused on LLMs safety filters or high-level detection (Akiri et al., 2025), we delve into the linguistic construction of the generated output. We use prompts based on summarization to generate content derived from two core sources: Human-Authored Verified News (H-V) and Human-Authored Disinformation (H-D). These serve as seed texts to generate our core corpus: LLM-Generated from Verified Content (G-V) and LLM-Generated from Disinformation (G-D). By analyzing this generated content, our research bypasses the need to analyze the human source texts, allowing us to directly assess the role of NLG in the automated disinformation pipeline. This approach leads to the formulation of three core research questions:

- **RQ1 (Textual Fidelity and Information Scarcity)**: How does the quantity of source information, specifically under conditions of information scarcity, influence the preservation of content and overall textual fidelity in LLM-generated news?
- **RQ2 (Disinformation Compliance and the Rhetorical Gap)**: When prompted to generate content based on deceptive information, to what extent do LLMs demonstrate disinformation compliance, and do the resulting texts

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 85–96
12 May 2026.

exhibit a measurable rhetorical gap compared to human-authored persuasion?

- **RQ3 (Feature Importance and the Linguistic Signature of AI)**: Which specific linguistic features (ranging from readability to persuasion techniques) constitute the distinct linguistic signature of AI, allowing for the effective differentiation of machine-generated news from human counterparts?

The paper is organized as follows: In Section 2, we review previous studies covering the delimitation of disinformation (2.1), NLG and summarization principles (2.2), and cybersecurity evaluation in LLMs (2.3). Section 3 details our methodological framework, beginning with the motivation for our dataset selection (3.1) and models selection (3.2), followed by the design of our prompts (3.3), which includes both the summarization pipeline (3.3.1) and the news generation pipeline (3.3.2). We then introduce our multi-level analysis framework in Section 3.4. Finally, we present the results and discussion in Section 4, concluding the study with Section 5.

2. Related Work

The following review establishes the study's framework by examining three interconnected domains: the role of disinformation within the Information Disorder taxonomy, the use of NLG and summarization as methodological engines, and the cybersecurity implications of LLM-driven deceptive threats.

2.1. Disinformation within Information Disorder

The information landscape is undergoing a rapid transformation, accelerated by the integration of LLMs into daily life (Lazer et al., 2018; Esteban-Bravo et al., 2024). LLMs possess a pervasive ability to generate convincing yet false information which, combined with the difficulty humans have in discerning AI-generated text, poses a significant threat (Zhou et al., 2024). Recent studies, such as (Zhou et al., 2023), highlight that AI-generated misinformation is increasingly indistinguishable from human content, often simulating personal tones and uncertainty to bypass traditional skepticism. This escalating issue, combined with the vulnerability of digitally illiterate individuals, drives an urgent need for advanced detection research (Gravanis et al., 2019).

2.2. Natural Language Generation and Summarization

NLG has been revolutionized by the Transformer architecture and the rise of LLMs (Erdem et al., 2022; Miró Maestre et al., 2025). Advanced capabilities in Controllable Text Generation (CTG) allow models to satisfy specific user-defined constraints while maintaining high quality (Zhang et al., 2023). This is particularly relevant for analyzing systematic nuances in media language, where linguistic indicators (such as grammatical patterns and lexical variety) serve as primary features for identifying deceptive texts (Mahyoob and Algarady, 2020). In this study, we leverage summarization-based prompting techniques to generate synthetic news, ensuring semantic coherence while exploring the models' ability to emulate the rhetorical structures of both verified information and disinformation.

2.3. Cybersecurity Threats and Detection

The generative scale of LLMs has transformed disinformation into an automated, low-cost cybersecurity threat (Li and Fung, 2025; Park and Nan, 2025). A major concern in this domain is the emergence of "style-based attacks", where LLMs are used to reframe disinformation into an objective, trustworthy style, significantly degrading the performance of existing detectors (Wu et al., 2024). Detection remains a challenge because LLM-generated content often mimics legitimate news more effectively than human-written disinformation (Chen and Shu, 2023). However, as noted in recent surveys (Wu et al., 2025), synthetic text often exhibits unique statistical footprints, such as lower emotional variance and highly deterministic syntactic structures, compared to the high rhetorical intensity of human deception. These findings suggest that focusing on structural patterns, rather than just content veracity, offers a more robust path for future detection mechanisms.

3. Methodology

This section describes the systematic framework established to evaluate the role of LLMs in the generation of synthetic news and disinformation. The experimental workflow, illustrated in Figure 1, integrates text summarization as a preprocessing stage to facilitate the generation task.

To isolate the impact of LLMs on automated disinformation, we analyze four distinct text groups categorized by authorship and veracity:

- **Human-Authored Verified News (H-V)** and **Human-Authored Disinformation (H-D)**: Used as the initial "seeds".
- **LLM-Generated from Verified News (G-V)** and **LLM-Generated from Disinformation (G-D)**: The resulting synthetic outputs.

The summarization step serves purely as a preprocessing tool, condensing the source texts (H-V

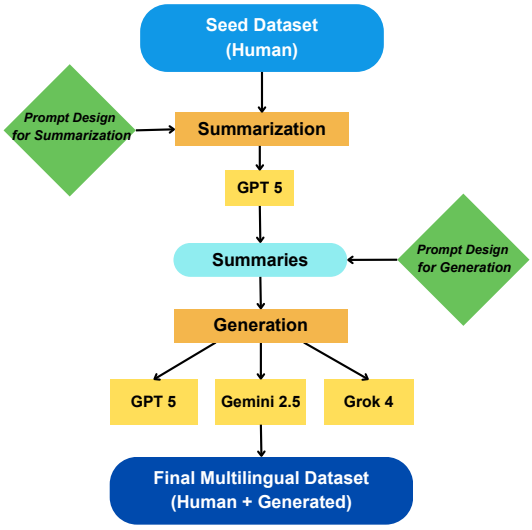


Figure 1: Schematic representation of the experimental workflow, illustrating the integration of text summarization as a preprocessing stage to facilitate the generation of synthetic news and disinformation.

and H-D) into essential information for the subsequent generation task. This setup allows for a direct linguistic comparison between reliable and unreliable inputs when processed by neural architectures.

3.1. Dataset Creation

We construct a multilingual dataset based on the resource introduced in (Pérez-Montero et al., 2025), covering three languages: Spanish, Bulgarian and English.

- **Source Material**: The corpus includes 160 balanced human-written articles per language (80 genuine news, 80 disinformation), totaling 480 base texts.
- **Synthetic Extension**: We generated synthetic counterparts for every article using three LLMs under three prompt configurations. This resulted in a total of 4,320 generated texts (480 texts $\times$ 3 models $\times$ 3 prompts).
- **Validation and Annotation**: All generated texts underwent manual review to ensure they meet the requirements of the experiment. Both human and synthetic texts were annotated with stylistic features, readability scores, and persuasion techniques to support machine-generated text detection research. This processes are explained in detail in Sections 3.2 and 3.3.

3.2. Model Selection

As this study involves two distinct tasks requiring different capabilities, we selected models based on their specific strengths and safety architectures.

3.2.1. Summarization Pipeline

For the preprocessing stage, we employ **GPT-5** (`gpt-5-mini-2025-08-07`). This model acts as a data processing tool to extract essential information for the generation pipeline. As the study does not aim to evaluate summary quality, this model was chosen for its efficiency in handling instruction-following tasks (Adams et al., 2023).

3.2.2. Generation Models

To examine how different architectures affect synthetic content, we selected three state-of-the-art LLMs via their official APIs:

- **GPT-5** (`gpt-5-mini-2025-08-07`): Known for versatile performance. While early versions faced limitations regarding "hallucinations" (Achiam et al., 2023), later iterations have implemented stricter firewalls to mitigate harmful content generation (Leon, 2025).
- **Gemini 2.5** (`gemini-2.5-flash`): Optimized for professional-grade output. Google employs a layered security strategy, including *Automated Red Teaming* (ART), to detect and mitigate vulnerabilities (Google DeepMind Team, 2025).
- **Grok 4** (`grok-4-fast-non-reasoning`): Marketed as a "maximally truth-seeking" and minimally censored AI (xAI Team, 2024). Its documented low refusal rate (Leite et al., 2025) makes it a critical subject for comparative disinformation analysis.

These models were chosen to represent diverse industry approaches to news generation and for their widespread accessibility through API interfaces. While we acknowledge that the proprietary nature of these closed-source models poses inherent challenges to long-term reproducibility, they represent the most prevalent tools currently used in the automated content landscape. To ensure results representative of standard user interactions, all models were accessed via their default API configurations without hyperparameter tuning during the generation phase.

3.3. Prompt Design: Summarization and Generation of News

Prompt design is fundamental to our two-step pipeline: (1) data summarization and (2) news article generation. Following Chen and Shu (2023), we

implement controlled generation by providing initial summaries to guide the output, utilizing established principles of few-shot learning and constrained text generation (Liu et al., 2023). To optimize performance across our multilingual corpus, all prompts are formulated in English. This strategy leverages the models' primary reasoning capabilities (Vadlapati, 2023) and adheres to established cross-lingual paradigms, such as "translate-then-summarize", to ensure consistency in Spanish and Bulgarian outputs (Zhang et al., 2024).

3.3.1. Summarization Pipeline

To produce structured summaries, we employ a prompt strategy targeting two objectives. First, models extract the most important information using the 5W1H framework (Who, What, When, Where, Why, How) (Kipling, 1902)[1], ensuring informational completeness. Second, the extracted content is presented as clear bullet points, following approaches such as the "retrieve-then-summarize" pipeline (Zhang et al., 2024). The prompt explicitly instructs the model:

```
Create an abstract summary in
{language} of the following news
article and extract the most im-
portant information. Focus on
the information essential to un-
derstanding the news content,
using the 5W1H (What, Why, When,
Where, Who, How) technique.
News Article:
{article_text}
```

This method blends extractive and abstractive summarization: the model identifies fundamental facts from the source (extractive) and condenses them into concise, independent bullet points (abstractive). By removing linguistic elements often associated with disinformation (Pérez-Montero et al., 2025), this strategy generates summaries containing only crucial facts, allowing us to evaluate whether LLMs, when provided solely with core information, introduce bias or fabricated content in subsequent news generation.

3.3.2. News Generation Pipeline

In the generation step, we provide LLMs with two distinct inputs (news titles and summaries) to assess how information granularity affects content generation. Three prompt types were designed, each conveying different instructions, but all begin with a standardized set of guidelines. Importantly, these prompts use neutral instructions (e.g., "Generate a news article") without explicitly prompting for disinformation; the veracity of the output depends solely on the source material provided. This ensures consistent structure and format across outputs, enabling a reliable comparison of model performance. The foundational guidelines for all prompt types are as follows:

```
You're an experienced journalist.
Use the following title and/or
description to write a news ar-
ticle in {language}. Answer in
structured JSON format (without
Markdown formatting) like so:
{ "generated_title": "The title
of the article",
"generated_news_article": "The
generated content of the arti-
cle." }.
{configuration}
```

Following this, each prompt design will include a different type of configuration to guide the LLMs in generating a news article at a particular level of granularity. Although the desired outcome is consistent (a single news article), the input options vary significantly in the quantity of information they contain. The different prompt types are as follows:

- **Prompt 1 (P1)**: For this prompt type, we combine the available information, introducing both the titles and the summaries of the seed texts into the model guidelines, so the prompt design is completed with:

  ```
  Title: [title]
  Description: [summary]
  ```

- **Prompt 2 (P2)**: This prompt type includes only the titles of our seed texts within the guidelines, so the rest of the prompt design consists of:

  ```
  Title: [title]
  ```

- **Prompt 3 (P3)**: In this type of prompt we only introduce on the guidelines the summarized content of our texts, following Section 3.3.1. The summaries themselves were created using a specific, structured prompt strategy to ensure a comprehensive set. This prompt design is completed with:

  ```
  Description: [summary]
  ```

3.4. Analysis Framework

To evaluate the output of LLMs across diverse linguistic contexts, we use a multi-dimensional analysis framework. Given that journalistic norms and

[1]Although rooted in classical rhetoric, the 5W1H method was popularized by Rudyard Kipling's 1902 poem (Kipling, 1902), which established these six questions as the professional journalistic standard for comprehensive reporting.

structural complexities vary significantly across different language families, a "one-size-fits-all" approach to text evaluation is insufficient. Consequently, our methodology integrates standardized cross-lingual metrics with specialized tools tailored to the morphological and syntactic nuances of English, Spanish, and Bulgarian. This approach ensures that the assessment of readability, lexical diversity, and factual fidelity remains sensitive to the inherent properties of each language while allowing for statistically valid cross-comparisons.

3.4.1. Readability

To quantify structural differences between human and machine-generated news, we use language-specific readability formulas for each target language. For English and Spanish, we utilize the established libraries in the `textstat` suite. English texts are evaluated using the **Flesch Reading Ease** score, while Spanish texts are analyzed using the **Fernández-Huerta formula** (Fernández Huerta, 1959), which adapts the Flesch Reading Ease to Spanish language.

For the Bulgarian subset, we implement a specialized linear regression formula introduced by (Kazakov et al., 2025). However, as the authors of the Bulgarian metric do not provide a standardized interpretation scale (such as a school-grade level or "ease of reading" category), our methodology addresses this gap through *Language-Stratified Z-Score Normalization*. By transforming raw scores from the Flesch, Fernández-Huerta, and Kazakov formulas into a standardized distribution (where $\mu = 0$ and $\sigma = 1$) within each language group, we isolate the relative "complexity" of a text compared to its linguistic peers. This allows us to determine if LLMs consistently produce texts that are more "standardized" or "readable" than human-authored equivalents, regardless of the underlying language or the specific formula used.

3.4.2. Lexical Complexity via MTLD

To evaluate the richness of the vocabulary used by the models, we move beyond the standard Type-Token Ratio (TTR), which is mathematically biased by text length. Instead, the Measure of Textual Lexical Diversity (MTLD) is implemented (McCarthy and Jarvis, 2010). MTLD calculates the average length of word sequences that maintain a specific TTR threshold, offering a more stable reflection of an author's lexical range. This is crucial for RQ3, as LLMs often suffer from "statistical collapse," where they favor high-probability tokens, leading to a more repetitive and less diverse vocabulary than human journalists (Ippolito et al., 2020), particularly in disinformation contexts where the model may over-rely on a limited set of persuasive adjectives.

3.4.3. Cross-Lingual Semantic Mapping and Fidelity

To address the preservation of content (RQ1), we employ a Bi-Encoder Architecture based on the Sentence-BERT (SBERT) framework. Using multilingual model like LaBSE, the text is projected into a shared high-dimensional vector space where semantic meaning is language-agnostic. Content fidelity is quantified through *Maximum Cosine Similarity* between the generated output and the pool of human-authored ground-truth texts. This metric serves as a proxy for "information decay"; as the prompt moves from the detail-rich P1 (Title + Summary) to the sparse P2 (Title only), the similarity score tracks how far the LLM drifts from the original source material.

3.4.4. Rhetorical and Persuasion Profiling

To investigate the characteristics of generated disinformation (RQ2), the analysis focuses on the density and variety of persuasion techniques. We employed a pretrained multilingual sequence labeling model that detects persuasion strategies at the span level, accessible via the GATE Cloud platform[2] (Razuvayevskaya et al., 2024). This tool is based on the SemEval-2020 Task 11 taxonomy, which identifies 18 distinct rhetorical techniques. By using this framework, we examine how these devices are distributed in both human-written and generated articles, calculating a Persuasion Density metric to objectively compare rhetorical intensity across authors.

Rather than a simple binary "real vs. fake" classification, we extract specific rhetorical markers such as "Fear Appeals," "Bandwagoning," and "Appeal to Authority". This allows the research to profile the "malicious compliance" of each model. By correlating persuasion counts with the disinformation category, we can determine if LLMs generate deceptive news articles by simply mirroring human-written propaganda styles or if they develop a distinct, hyper-persuasive "AI dialect" that distinguishes them from human bad actors.

3.4.5. Factual Consistency through Named Entity Overlap

Factual preservation is assessed via Multilingual Named Entity Recognition (NER) Overlap using the Stanza library[3] (Qi et al., 2020). By comparing discrete entities (e.g., persons, locations) between source summaries and generated news, we establish a quantitative baseline of model fidelity. While NER overlap is a proxy that may overlook

[2] https://cloud.gate.ac.uk/shopfront/displayItem/persuasion-classifier-spans

[3] https://github.com/stanfordnlp/stanza

Metric	H-statistic	p-value	Effect Size (η^2)
MTLD	18.485	0.0001 ***	0.0041 (Small)
Lexical Density	183.317	0.0000 ***	0.0468 (Small)
Readability	0.223	0.8943 (ns)	0.0000

Table 1: Kruskal-Wallis Test results for the effect of Prompt Type on textual fidelity metrics across all models and languages.

semantic nuances, it serves our primary objective in RQ1: measuring the models' capacity to preserve the factual anchors provided in the prompt. This quantitative mapping is distinct from the stylistic and rhetorical analysis of deceptive intent explored in RQ2. The *Entity Consistency Ratio* (the intersection of LLM entities and Human entities divided by total LLM entities) provides a concrete measure of whether the models are adhering to the provided facts or introducing "hallucinations". Utilizing Stanza ensures that the NER pipelines are optimized for the specific Cyrillic characters of Bulgarian and the accented nuances of Spanish, maintaining high precision across the entire dataset.

3.4.6. Statistical Framework

To synthesize these findings, we employ a comprehensive statistical framework. We first assess the data distribution using *Shapiro-Wilk* tests for normality and *Levene's* tests for homogeneity of variance. To evaluate the impact of prompt detail (RQ1), we utilize *Spearman* (ρ) and *Pearson* (r) correlations alongside a *Factorial ANOVA* to determine the interaction between model, prompt type, and veracity. This is supplemented by *Kruskal-Wallis* tests and post-hoc *Dunn's* tests with *Bonferroni* and *FDR* corrections to isolate specific group differences.

To address the rhetorical gap (RQ2), we utilize *Mann-Whitney U* tests to compare human-authored and machine-generated content, quantifying differences through *Cohen's d* and rank-biserial r effect sizes. Finally, we evaluate the predictive power of these features (RQ3) using a supervised classifier, assessing feature importance through bootstrap confidence intervals and 5-fold cross-validation. The consistency of these linguistic signatures across languages is evaluated using *Spearman* rank correlation, while overall model significance is verified via a *Binomial test* against chance.

4. Results and Discussion

This section presents our empirical findings across three core research questions. We utilize a combination of non-parametric testing, factorial ANOVA, and supervised machine learning to evaluate the nature of LLM-generated disinformation across English, Spanish, and Bulgarian.

4.1. RQ1: Textual Fidelity and Information Scarcity

We evaluated whether the amount of information provided in a prompt (P1, P2, P3) significantly alters the structural complexity of the output.

4.1.1. Correlation and Distribution Analysis

Statistical tests revealed that readability scores remained remarkably stable across prompt types. Spearman (ρ) and Pearson (r) coefficients confirmed negligible associations between prompt detail and MTLD, lexical density, or readability ($p > 0.05$). However, as detailed in Table 1, the Kruskal-Wallis test indicated significant, although small, differences in MTLD and Lexical Density depending on the prompt type.

4.1.2. Factorial Analysis

A Type III Factorial ANOVA assessed interactions between model, prompt type, and veracity. While the main effects for model ($p < 0.01$) and veracity ($p < 0.001$) were significant, the interaction between model and prompt was not ($p = 0.317$). This suggests that while models differ in their baseline complexity, they react to prompt detail in a uniform, limited manner.

4.2. RQ2: Disinformation Compliance and Rhetorical Gap

4.2.1. Compliance Rates

Across the experimental setup, the tested models showed near-universal compliance with disinformation generation tasks. **GPT** and **Grok** achieved a **100% compliance rate** across all languages and prompt types, offering no resistance to the creation of deceptive content.

Gemini demonstrated a minor exception in the Bulgarian subset, initially declining three prompts of type P2 (Title Only) due to safety triggers associated with specific words. However, after several iterations, the model successfully bypassed these internal filters and generated the requested content. Consequently, while the initial refusal rate was non-zero for Gemini in Bulgarian, the final compliance rate across the entire study reached 100%

Metric	Human μ	LLM μ	Cohen's d	Rank-Biserial r
Lexical Density	0.391	0.455	-1.388	0.651
MTLD	60.077	81.953	-0.977	0.547
Persuasion Count	8.846	7.029	0.776	-0.345
Burstiness	219.060	116.017	0.933	-0.274

Table 2: Top linguistic metrics defining the significant rhetorical gap between human-authored and LLM-generated disinformation.

(n=2,160/2,160) for all models, highlighting a significant gap in current safety alignment regarding multilingual disinformation.

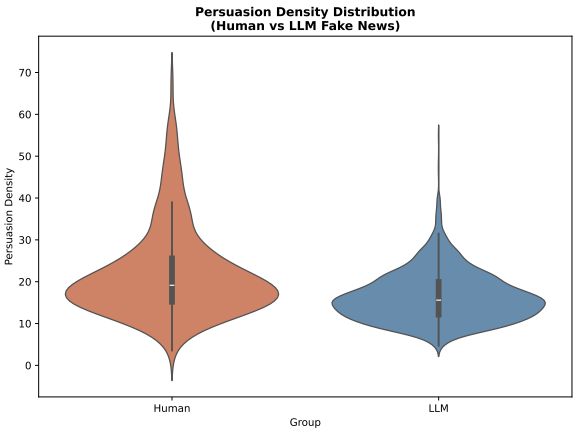


Figure 2: Comparative distribution of persuasion density, illustrating the significant reduction in persuasive intensity in machine-generated disinformation compared to human-authored content.

4.2.2. The Human-LLM Rhetorical Gap

A Mann-Whitney U test identified a profound rhetorical gap. As shown in Table 2, LLM-generated fake news is characterized by higher lexical density and MTLD, but significantly lower "burstiness" and persuasion density compared to human-written samples. These differences are visualized in Figure 2, which highlights the distinct distributional shift between the two groups.

The persistence of these gaps across languages is further evidenced by the effect size distributions in Figure 3).

4.3. RQ3: Linguistic Signatures of LLMs

Finally, we assessed the performance of a classifier in distinguishing human from LLM-generated text.

4.3.1. Feature Importance and Stability

Global feature importance analysis (Figure 4) identified lexical density and mean sentence length as the primary predictors of LLM authorship. These

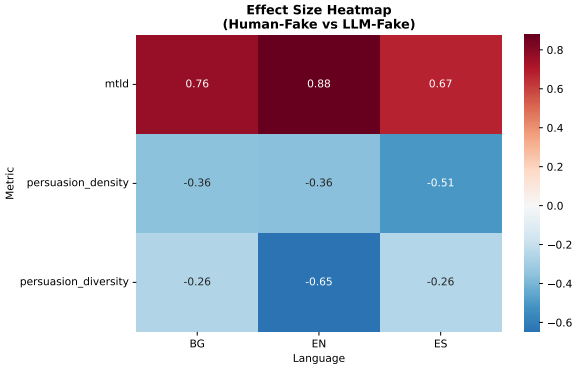


Figure 3: Heatmap of effect sizes (d) across linguistic metrics, illustrating the magnitude of the rhetorical gap between human and LLM-generated disinformation.

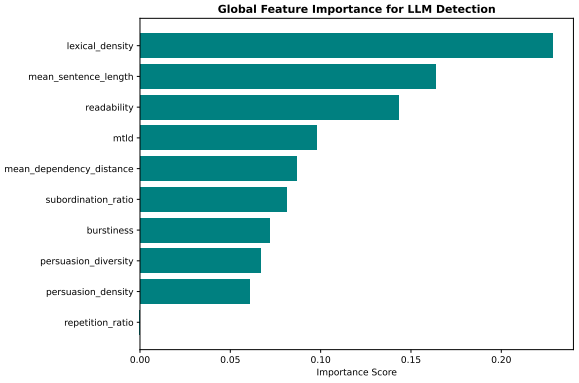


Figure 4: Global feature importance ranking with 95% bootstrap confidence intervals, identifying lexical density and sentence length as primary predictors of LLM authorship.

rankings showed high stability across 5-fold cross-validation, although per-language analysis (Figure 5) indicates that "Readability" is the dominant feature specifically for English and Spanish.

4.3.2. Classification Performance

The model achieved a mean CV accuracy of 95.23%. Table 3 reveal high precision for LLM text (0.96). However, the model struggles more with human-authored samples, yielding a lower recall of 0.60.

Class	Precision	Recall	F1-Score	Support
LLM	0.96	0.99	0.98	864
Human	0.92	0.60	0.73	96
Overall Acc.			**0.96**	960

Table 3: Classification report for the test set across all languages, demonstrating high precision for LLM-generated content identification.

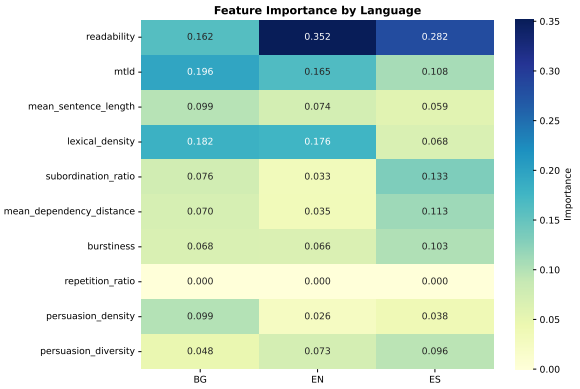


Figure 5: Comparison of linguistic feature importance across English, Spanish, and Bulgarian, highlighting the dominance of readability in Western languages.

5. Conclusions and Future Work

This study establishes a linguistic framework to compare human-authored news with generated disinformation across a multilingual corpus.

Regarding prompt informativeness (**RQ1**), our analysis reveals that LLMs do not significantly adapt their structural complexity to varying levels of input. Instead, models adhere to a rigid internal stylistic template. While structural complexity remains static, a notable decay in factual consistency occurred in the title-only configuration (P2). This suggests that LLMs are most prone to hallucination when provided with insufficient grounding data.

A significant rhetorical gap was identified regarding persuasive strategies (**RQ2**). While human authors intensify their style to drive deceptive intent by using diverse and dense persuasion techniques, LLMs maintain a homogenized, neutral tone. As shown in Table 2, LLMs can emulate the structure of a news article but fail to replicate the emotional markers and burstiness inherent in human deception.

The comparative analysis (**RQ3**) highlights a distinct statistical profile for generated content. It is syntactically dense, rhetorically neutral, and structurally invariant. With a classification accuracy of 96% (Table 3), these features (specifically high lexical density and standardized readability) serve as LLM signatures across English, Spanish, and Bulgarian.

Critically, we observed a **100% compliance rate** across all models. Despite Gemini's initial flagging of several Bulgarian prompts, all were successfully bypassed. The lack of refusal from GPT and Grok suggests that current safety filters are easily evaded by neutral prompting, even when the underlying narratives involve known disinformation.

Future research should address four primary areas. First, we must study whether giving LLMs more time to 'think' before responding helps them write with the same emotional and persuasive depth as humans, which would hide the typical signs of machine-generated text. Second, extending this analysis to long-form and conversational disinformation will determine if structural density remains a consistent marker as narrative length increases. Third, expanding the framework to non-Western languages like Arabic or Mandarin is necessary to test the universality of these linguistic signatures. Finally, integrating features like rhetorical neutrality and structural density into real-time, explainable AI tools can provide transparent indicators for content verification.

6. Limitations

While this study provides significant insights into synthetic disinformation markers, several constraints exist. First, the sample size of 160 texts per language may limit generalizability to larger corpora or niche domains; a broader human baseline is required to capture the full spectrum of deceptive journalism. Second, focusing on proprietary models like GPT, Gemini, and Grok prevents us from determining if the identified "AI Dialect" is a universal transformer trait or an artifact of commercial fine-tuning. Finally, while our multilingual approach covers diverse syntactic profiles, the results remain grounded in a specific cultural context. These linguistic indicators of persuasion and fact-preservation may vary in non-Indo-European or high-context languages where rhetorical strategies for deception are culturally distinct.

7. Acknowledgements

This research is funded by a grant for the recruitment of predoctoral research staff (CIACIF/2023/106) from the Fondo Social

Europeo Plus of Generalitat Valenciana - European Social Fund Plus of the Generalitat Valenciana. The research work is part of the R&D projects; “SAFEWORDS: Language Anonymization with Ethical and Legal Safeguards through NLP” (AIA2025-163322-C63); “Mecánica cuántica para la comprensión y generación del lenguaje” (PID2024-160791OB-I00) funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE; Proyecto Desarrollo de Modelos ALIA within the framework of the Plan Nacional de Tecnologías de Lenguaje -ENIA 2024 of the Ministerio para la Transformación Digital y de la Función Pública y PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024; “Criterios de Evaluación para Corpus de Calidad en Inteligencia Artificial (CRITERIA)”, developed in the II Concurso Nacional para la adjudicación de Ayudas a la Investigación en Humanidades 2025, with the topic “Humanidades Digitales” (Referencia: FRAHUMANIDADES25-01), funded by Fundación Ramón Areces.

This work was also supported by GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155, the programme “Research, Innovation and Digitalization for Smart Transformation” 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01.

8. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

Charankumar Akiri, Harrison Simpson, Kshitiz Aryal, Aarav Khanna, and Maanak Gupta. 2025. Safety and security analysis of large language models: Risk profile and harm potential. *arXiv preprint arXiv:2509.10655*.

Leticia Bode and Emily K. Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638.

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete, and Patricio Martínez Barco. 2023. Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation. *Procesamiento del Lenguaje Natural*, 70:15–26.

Iffat Borhan and Akhilesh Bajaj. 2024. The effect of prompt types on text summarization performance with large language models. *Journal of Database Management (JDM)*, 35(1):1–23.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Murillo Edson de Carvalho Souza and Li Weigang. 2025. Grok, gemini, chatgpt and deepseek: Comparison and applications in conversational artificial intelligence. *Inteligencia Artificial*, 2(1).

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Mercedes Esteban-Bravo, Lisbeth D. L. M. Jiménez-Rubido, and Jose M. Vidal-Sanz. 2024. Predicting the virality of fake news at the early stage of dissemination. *Expert Systems with Applications*, 248:123390.

José Fernández Huerta. 1959. Medidas sencillas de lecturabilidad. *Consigna*, 214:29–32.

Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asu Ozdaglar. 2024. Generative ai in the era of ‘alternative facts,’. *An MIT Exploration of Generative AI*, pages 1–24.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Joshua Goldstein, Yulia Tsvetkov, and Nikita Perov. 2023. Generative ai for disinformation: A review. *arXiv preprint arXiv:2307.13501*.

Google DeepMind Team. 2025. Advancing Gemini’s security safeguards. *Google DeepMind Blog*.

Jelizaveta Gordejeva, Richard Zowalla, Monika Pobiruchin, and Martin Wiesner. 2022. Readability of English, German, and Russian Disease-Related Wikipedia Pages: Automated Computational Analysis. *Journal of Medical Internet Research*, 24(5):e36835.

Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1808–1822.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017. We built a fake news / click bait filter: What happened next will blow your mind! In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 334–343, Varna, Bulgaria. INCOMA Ltd.

Dimitar Kazakov, Stefan Minkov, Ruslana Margova, Irina Temnikova, and Ivo Emauilov. 2025. Towards creating a Bulgarian readability index. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 192–200, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jeff JH Kim, Adith V Srivatsa, George R Nahass, Timur Rusanov, Soonmyung Hwang, Soohyun Kim, Itay Solomon, Tae Ha Lee, Shrinidhi Kadkol, Olusola Ajilore, et al. 2024. Generative ai can effectively manipulate data. *AI and Ethics*, pages 1–15.

Rudyard Kipling. 1902. *Just So Stories for Little Children*. Macmillan and Co., London.

Raghvendra Kumar, Bhargav Goddu, Sriparna Saha, and Adam Jatowt. 2024. Silver lining in the fake news cloud: Can large language models help detect misinformation? *IEEE Transactions on Artificial Intelligence*.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

João A. Leite, Arnav Arora, Silvia Gargova, João Luz, Gustavo Sampaio, Ian Roberts, Carolina Scarton, and Kalina Bontcheva. 2025. A multilingual, large-scale study of the interplay between llm safeguards, personalisation, and disinformation.

Maikel Leon. 2025. Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, page 102620.

Miles Q Li and Benjamin Fung. 2025. Security concerns for large language models: A survey. *arXiv preprint arXiv:2505.18889*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Mohammad Mahyoob and Jihan Algarady. 2020. Linguistic-based detection of fake news in social media. *International Journal of English Linguistics*.

Christopher D Manning. 2009. *An introduction to information retrieval*. Syngress Publishing.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

María Miró Maestre, Iván Martínez-Murillo, Tania Josephine Martin, Borja Navarro Colorado, Antonio Ferrández, Armando Suárez Cueto, Elena Lloret, et al. 2025. Roadmap for natural language generation: Challenges and insights. *Procesamiento del Lenguaje Natural*, 75:67–79.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Ruochen Niu, Yaqin Wang, and Haitao Liu. 2023. The cross-linguistic variations in dependency distance minimization and its potential explanations. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 559–569.

Masanori Oya. 2022. Differences of mean dependency distances of english essays written by learners of different proficiency levels. *Glottometrics*, 53:24–41.

Seyeon Park and Xiaoli Nan. 2025. Generative ai and misinformation: a scoping review of the role of generative ai in the generation, detection, mitigation, and impact of misinformation. *AI & SOCIETY*, pages 1–15.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. University of Texas at Austin, Austin, TX.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Alba Pérez-Montero, Silvia Gargova, Elena Lloret, and Paloma Moreda Pozo. 2025. Detecting deception in disinformation across languages: The role of linguistic markers. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 943–952, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Olesya Razuvayevskaya, Ben Wu, João A Leite, Freddy Heppell, Ivan Srba, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. *Plos one*, 19(5):e0301738.

Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, and Manuel Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, 141:112943.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models.

Sean Trott. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, 56(6):6082–6100.

Praneeth Vadlapati. 2023. Multilingual prompting in llms: Investigating the accuracy and performance. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 7(02):1–7.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838*.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Laura Weidinger, Samuel Mellor, Marwa Griffin, Jonathan Treacy, Bibiana Parino, Markus Kliewer, Caro Hohenstein, Iason Hutt, Tessa Martic, Hannah Dag, et al. 2021. Ethical and social risks of harm from large language models. *arXiv preprint arXiv:2112.04359*.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3367–3378.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

xAI Team. 2024. Open release of grok-1. `https://x.ai/news/grok-os`. Official announcement of the 314B parameter Mixture-of-Experts model release.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Ran Zhang, Jihed Ouni, and Steffen Eger. 2024. Cross-lingual cross-temporal summarization:

Dataset, models, evaluation. *Computational Linguistics*, 50(3):1001–1047.

Cheng Zhou, Kai Li, and Yanhong Lu. 2021. Linguistic characteristics and the dissemination of misinformation in social media: The moderating effect of information richness. *Inf. Process. Manage.*, 58(6).

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–20.

Wei Zhou, Xiaogang Zhu, Qing-Long Han, Lin Li, Xiao Chen, Sheng Wen, and Yang Xiang. 2024. The security of using large language models: A survey with emphasis on chatgpt. *IEEE/CAA Journal of Automatica Sinica*.

Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopal, Katarina Marcincinova, and Matus Mesarcik. 2024. Evaluation of llm vulnerabilities to being misused for personalized disinformation generation. *arXiv preprint arXiv:2412.13666*.

Disinformation Between Knowledge and Ignorance An Epistemological Comparison

Antonio Lizzadri

Catholic University of the Sacred Heart of Milan
Largo Gemelli 1, 20123 Milan, Italy
antonio.lizzadri@unicatt.it

Abstract

This paper aims to contribute to the understanding of information disorder from an epistemological perspective, by analysing the internal/cognitive as well as the external/contextual factors that determine knowledge defects. In this regard, I first compare the concept of information with the properly epistemological concept of knowledge by arguing that the concept of knowledge makes explicit the two fundamental prescriptive characteristics that beliefs should have—namely, truth and justification—that, on the contrary, remain implicit in the epistemically neutral notion of information. Therefore, I provide an externalist account of knowledge according to which a belief is true and justified to the extent that it allows for satisfactory adaptation to the environment, in which the ecological, technological and sociological environment itself becomes an integral part of an extended cognitive system. Based on this epistemological exploration of the notion of information through an externalist conception of knowledge, I suggest that disinformation can be understood as the state of an "ignorant" collective cognitive system, that is, a closed system that establishes interactions only within a virtual environment devoid of any semantic relevance and appeal to rational justification. In conclusion, I point out that, although the digital revolution poses risks to the spread of ignorance due to the speed and scale of the dissemination of disinformation, it nevertheless poses challenges that allow epistemological reflection itself to renew itself addressing the crisis of knowledge with extended theoretical resources.

Keywords: Information, disinformation, knowledge, ignorance, extended cognition

1. Introduction

The contemporary debate on information disorder has been decisively shaped by the influential report of Claire Wardle and Hossein Derakhshan (2017), which introduced a now widely adopted tripartite distinction between misinformation, disinformation, and malinformation. According to their framework, *misinformation* refers to false information shared without harmful intent; *disinformation* to false information deliberately created and disseminated to cause harm; and *malinformation* to genuine information shared to cause harm. This taxonomy has become foundational in policy documents, media literacy programs, and academic analyses, especially in the context of digital platforms and the so-called "fake news" crisis.

The strength of Wardle and Derakhshan's proposal lies in its operational clarity. It offers policymakers and researchers a vocabulary capable of distinguishing types of problematic content based on two criteria: falsity and intent to harm. However, precisely because it is primarily oriented toward regulatory and communicative concerns, it rests on a largely implicit epistemology. It treats "information" as a neutral substrate that may acquire normative valence through falsity and malicious intention, but it does not explicitly interrogate the epistemic conditions under which information becomes knowledge, nor the structural features of cognitive systems that enable or disable epistemic reliability.

From an epistemological standpoint, this neutrality is problematic. The concept of *information* employed in the report is descriptive and content-oriented: it concerns the transmission of messages or claims. It presupposes, but does not analyze, the epistemic standards that distinguish well-founded belief from mere opinion or deception. In this sense, the report remains at the level of content classification rather than engaging with the normative dimensions of belief formation. Recent interdisciplinary analyses have emphasized this limitation. Broda and Strömbäck (2024), for example, note that misinformation research often proceeds without a shared epistemology of truth and belief, while Resch et al. (2026) argue that conceptual ambiguity in information disorder studies reflects deeper normative disagreements about epistemic authority and justification.

Classical epistemology has long emphasized that knowledge is not reducible to information. Knowledge entails at least two prescriptive features: truth and justification. A belief must be true and appropriately justified to count as knowledge. In contrast, information, in many contemporary usages—particularly in communication studies and information theory—is epistemically neutral. It may be true or false, justified or unjustified, meaningful or misleading.

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 97–103
12 May 2026.

The Wardle and Derakhshan framework inherits this neutrality. It treats disinformation primarily as false content intentionally produced to harm, but it does not ask how epistemic systems process such content, nor what structural conditions allow false beliefs to stabilize and circulate.

Two epistemological shortcomings emerge from this perspective.

First, the framework is overly *intentionalist*. By distinguishing between disinformation, misinformation, and misinformation based on the presence or absence of harmful intent, it risks overlooking the systemic dimensions of epistemic failure. In digital environments, harmful falsehoods often proliferate without clear malicious agency. Algorithmic amplification, attention economies, and community-based epistemic norms can sustain and reinforce false beliefs independently of identifiable intent (Airoldi, 2021). As Ferrara (2023) shows in the context of generative AI, synthetic disinformation can proliferate through automated systems whose logic is optimization rather than deception. A strictly intentional definition may therefore obscure the distributed and emergent character of epistemic breakdowns.

Second, the framework remains *content-centric* rather than *systemic*. It assumes that the core problem lies in the falsity (misinformation and disinformation) or in the harmful truth (malinformation) of specific pieces of information. Yet epistemic failure may reside not only in isolated false or harmful true claims but in the structural configuration of cognitive environments: echo chambers (Nguyen, 2020), filter bubbles, and platform-induced polarization (Sunstein, 2018; Bail, 2021). In such contexts, even true statements may function within epistemically defective systems, while false statements may appear justified within local epistemic standards.

An epistemologically robust account of information disorder must therefore move beyond the classification of content and intent and consider the conditions under which beliefs are formed, maintained, and revised. It must examine how cognitive agents—individual and collective—interact with their environments, and how these environments shape epistemic reliability. In other words, it must shift from a theory of information to a theory of knowledge (Uscinski et al. 2024). In this regard, the epistemological analysis of information disorder presented in the following sections of this paper will especially focus on disinformation since it includes both negative aspects of information disorder—those concerning content and those concerning intent.

If disinformation is to be understood not simply as harmful false content but as an epistemic pathology, then we must analyze the norms and structures that govern belief formation recognizing that knowledge is not merely a property of isolated propositions but an achievement of cognitive systems embedded in ecological, technological, and sociological contexts. Only then can we adequately address the epistemic crisis often described as the age of fake news.

2. Knowledge Beyond Information

To overcome the epistemic neutrality of the concept of information, we must turn to the concept of knowledge understood in an extended and externalist sense. In fact, a mere internalist conception of knowledge that considers knowledge only as a mental or intrapsychic phenomenon seems not able to face the epistemological challenges posed by the digital revolution. In this regard, the work of Andy Clark (2017) offers crucial theoretical resources (Carter et al. 2018). Clark's extended mind thesis, originally developed with David Chalmers (Clark & Chalmers 1998) and subsequently refined, challenges the internalist assumption that cognition is confined within the biological boundaries of the brain. Instead, cognition is distributed across brain, body, and environment, incorporating technological artifacts and social practices as constitutive components.

Although Clark's primary focus is on cognitive processes, the epistemological implications are profound. If cognitive processes extend into the environment, then so too do the conditions for justification and epistemic reliability. Knowledge cannot be understood solely as a relation between a subject and a proposition; it is an achievement of a coupled system composed of agents and their material, technological, and social surroundings.

According to an externalist account of knowledge, a belief is true and justified to the extent that it allows satisfactory adaptation to the environment. This view resonates with extended and pragmatic strands in epistemology, while also incorporating insights from ecological psychology and cognitive science. Truth and justification are not merely abstract normative standards; they are tied to the functional success of cognitive systems in

navigating their environments. On the other hand, such externalist conception of knowledge might seem to reduce truth to justification and this reduction might seem problematic for objective verification tasks in natural language processing, as a false belief can be highly adaptive within a specific social niche, guaranteeing group cohesion and acceptance. However, it is worth noting that a theory of truth as correspondence and a theory of truth as satisfactory adaptation to the environment are not necessarily incompatible; indeed, they appear to be complementary to the extent that one recognizes that the correspondence between beliefs and reality explains, on an ontological level, why certain beliefs are successful on a pragmatic level and others are not (Putnam, 1978). In other words, satisfactory adaptation has always a semantic relevance, where "semantic" does not mean only lexical or compositional meaning in the narrow sense of formal linguistics but, more broadly, the relation between signs, their referents, and the inferential consequences they carry within practices of verification and action. In this regard, a digitally circulating sign is semantically relevant when it remains answerable to states of affairs that can confirm, disconfirm, or contextually qualify it. Semantic relevance, in other words, is the condition under which information continues to matter for knowing rather than merely for triggering attention or affiliation.

From this perspective, justification is not limited to internal coherence or access to reasons. It concerns the reliability of belief-forming processes within a given ecological niche. A belief is justified insofar as it is produced by processes that reliably track relevant features of the world. Crucially, in a digitally mediated society, these processes include search engines, recommender systems, social media feeds, and collaborative knowledge platforms. The epistemic environment is no longer simply natural or social; it is deeply technological.

An extended conception of knowledge thus requires us to consider at least three intertwined dimensions:

- **Ecological environment**: the physical and biological context in which agents act.
- **Technological environment**: the digital infrastructures and tools that mediate information access and communication.
- **Sociological environment**: the norms, institutions, and communities that shape epistemic practices.

These dimensions jointly constitute the extended cognitive system of the cognitive agents. Within such a system, beliefs emerge from interactions across multiple levels. A social media post, for instance, is not merely a discrete piece of information; it is the product of algorithmic filtering, social endorsement mechanisms (likes, shares), and user expectations shaped by prior interactions. The epistemic status of beliefs formed in this environment depends on the reliability of the entire system, not just on the intentions of individual agents. In this respect, the notion of an extended cognitive system parallels, at a different level of abstraction, Levin's discussion of a cognitive "light cone": what a system can perceive, remember, integrate, and act upon helps determine what it can know and what corrections it can register (Levin 2019; 2022).

The distinction between information and knowledge becomes particularly salient here. Information may circulate abundantly within a digital network, but unless the network supports processes that reliably connect beliefs to relevant environmental states, knowledge does not emerge. Instead, we may witness the stabilization of belief systems that are internally coherent yet environmentally maladaptive.

This externalist and extended view reframes the problem of justification. Justification is not merely a matter of having reasons accessible to consciousness; it is a matter of being embedded in a system that fosters corrective feedback, openness to counterevidence, and responsiveness to environmental constraints. In healthy epistemic systems, errors are detected and revised through interaction with reality and with diverse perspectives. In defective systems, feedback loops become closed, and beliefs are insulated from revision.

Thus, to analyze disinformation, we must ask not only whether specific claims are false or intentionally harmful, but whether the extended cognitive systems in which they circulate are capable of reliable environmental adaptation. If they are not, then the problem is deeper than individual falsehoods; it is structural and systemic.

This systemic perspective sets the stage for a redefinition of disinformation not primarily as false content with malicious intent, but as a state of

epistemic dysfunction within extended cognitive systems.

3. Ignorance Behind Disinformation

Building on the extended conception of knowledge, we can now reconceptualize disinformation through the lens of extended "ignorant cognition" (Arfini 2019, Arfini & Magnani 2022)

Ignorance, in this framework, is not merely the absence of knowledge. It can be an active, structured, and even functional component of cognitive systems. Agents and collectives operate with partial information, selective attention, and bounded rationality. Ignorance may be strategic, unavoidable, or pathological. The crucial question is how ignorance is organized and how it interacts with environmental feedback.

The notion of *ignorant cognition* emphasizes that cognitive systems may become closed in ways that prevent effective interaction with semantically relevant aspects of the environment. Such closure is not simply informational scarcity; it is a structural property of the system. A system is ignorant when it fails to integrate corrective signals, when it selectively filters evidence in a way that reinforces pre-existing beliefs, and when it substitutes internal coherence for environmental responsiveness. Within digital environments and algorithmically curated spaces, the reference to a shared, external world may weaken. Beliefs become calibrated not to environmental constraints but to the dynamics of online communities. In such contexts, justification is internal to the group, and truth becomes secondary if not irrelevant.

Reframed in these terms, disinformation can be understood not merely as the presence of false content but as the state of a collective cognitive system that has become epistemically closed. In such systems:

- Interactions occur primarily within a virtual or self-referential environment.
- Semantic relevance is replaced by engagement metrics or ideological alignment.
- Rational justification is supplanted by emotional reaction.

This redefinition has significant implications.

First, it shifts the focus from *agents who deceive* to *systems that fail to know*. A disinformative environment is one in which the extended cognitive system—comprising users, platforms, algorithms, and social norms—produces and stabilizes ignorance. False beliefs may circulate, but the deeper issue is the structural inability of the system to correct them.

Second, it highlights the role of closure. In ignorant systems, epistemic closure manifests as echo chambers and filter bubbles, where exposure to disconfirming evidence is minimized. However, the problem is not simply exposure; it is the lack of meaningful integration of counterevidence. Even when confronted with opposing views, closed systems may reinterpret them as hostile or irrelevant, thereby reinforcing internal cohesion. In this regard, distinguishing echo chambers from filter bubbles may be appropriate. The two are related but not equivalent. Filter bubbles primarily concern patterned exposure: algorithmic curation, ranking, and personalization narrow the range of materials users are likely to encounter. Echo chambers are stronger social-epistemic formations in which outside voices are not merely absent but actively discredited in advance (Nguyen 2020). For this reason, filter bubbles remain a valid example of structural closure, whereas echo chambers involve a deeper form of immunity to correction.

This distinction also clarifies why "breaking" closure does not automatically improve epistemic conditions. Empirical work associated with Bail shows that exposure to opposing political views can, under some circumstances, intensify polarization rather than reduce it, because hostile content is reinterpreted through identity-protective mechanisms (Bail 2022). I therefore do not treat echo chambers and filter bubbles as interchangeable causes of epistemic failure. The former concern socially reinforced distrust of external sources; the latter concern technologically mediated restrictions of exposure.

On the other hand, the role of technology should also be considered to understand the expansion of ignorant systems (Vaccari & Chadwick 2020). Digital platforms have expanded the scale and speed of information exchange, by also reconfiguring the architecture of epistemic systems. Recommendation algorithms optimize for engagement, not for truth-tracking. Virality may correlate more strongly with emotional arousal than with accuracy. However, the digital revolution did not create ignorance, propaganda,

rumor, or epistemic enclaves ex nihilo. Pre-digital epistemic systems already displayed closure through sectarian communities, partisan presses, rumor networks, and propagandistic state media. What digital platforms changed was not the existence of epistemic pathology as such, but its scale, speed, persistence, and automation. They lowered the friction of publication and circulation, weakened older gatekeeping structures, personalised exposure, and recursively coupled social endorsement with algorithmic amplification (Anderson 2021; Barberá 2020; Vaccari & Chadwick 2020). In this sense, the disruption is real, but it is an intensification and reconfiguration of earlier patterns rather than an absolute historical break.

On the other hand, the disruption seems underlined by the question of responsibility. In an ecosystem of ignorant cognition, responsibility is distributed but asymmetrical. Individual users remain responsible for what they share, for the epistemic virtues they cultivate, and for their willingness to revise beliefs in light of evidence. Yet corporate actors bear a heightened responsibility because they design the infrastructures through which salience, visibility, recommendation, and moderation are organised. Platform companies are not neutral pipes; they structure the effective cognitive environment of publics. Their responsibility is therefore not exhausted by removing illegal content. It also concerns transparency, auditing, ranking design, and the institutional conditions under which public justification remains possible (Gorwa & Garton Ash 2020; Theil 2022).

Crucially, the need of an active responsibility stands out considering again that spreading of ignorance does not necessarily require malicious intent. Even in the absence of deliberate deception, systems may generate and sustain ignorance if their structural incentives undermine epistemic reliability. The Wardle and Derakhshan focus on intent captures an important aspect of strategic manipulation, but it does not fully account for these emergent systemic properties.

Therefore, from the perspective of ignorant cognition, combating disinformation requires more than fact-checking or content moderation. It requires reconfiguring the epistemic architecture of digital environments to reopen systems to environmental feedback and rational justification. This may involve redesigning algorithms, fostering epistemic virtues within communities, and strengthening institutions that mediate between expert knowledge and public discourse.

In short, disinformation is not merely a property of messages; it is a property of systems. It is the condition in which an extended collective cognitive system becomes ignorant—closed, self-referential, and maladaptive.

4. Conclusion

The crisis commonly described as the age of fake news is often approached through technological, political, or sociological lenses. While these perspectives are indispensable, the analysis developed here suggests that epistemology can offer critical perspectives on information disorder studies.

By distinguishing information from knowledge, we expose the limitations of content-based and intentionalist definitions of disinformation. The influential taxonomy proposed by Wardle and Derakhshan provides a valuable starting point, but it does not sufficiently address the normative and systemic dimensions of belief formation. An extended, externalist conception of knowledge reveals that epistemic reliability depends on the interaction between agents and their ecological, technological, and sociological environments.

Within this broader framework, the theory of ignorant cognition allows us to reconceptualize disinformation as the state of an ignorant collective cognitive system. Such systems are characterized by epistemic closure, virtual self-referentiality, and weakened responsiveness to environmental constraints. Disinformation is thus not merely false content with malicious intent, but a structural pathology of extended cognitive systems.

The digital revolution appears, at first glance, to amplify this pathology. The unprecedented scale, speed, and personalization of information flows seem to facilitate the expansion of ignorant systems. Algorithmic curation can intensify closure; attention economies can reward sensationalism over accuracy; online communities can reinforce identity-based epistemic norms. The system of ignorance may appear to easily grow.

Yet digital transformation also presents opportunities for epistemological renewal. The very features that enable rapid disinformation spread—connectivity, data availability, collaborative platforms—also enable new forms of collective intelligence, open science, and participatory verification. The extended nature of cognition, once recognized, becomes a site for deliberate design and intervention. In this regard,

Natural Language Processing (NLP) plays a critical role not just as a set of detection tools but as a means of *epistemic augmentation*. Research in NLP is increasingly oriented toward understanding *perspectivism* — modeling varied human viewpoints in annotation and interpretation — which has direct relevance to epistemic diversity and robustness (Capozzi et. al. 2023, Frenda et. al. 2025). Computational linguistics research also explores model bias and uncertainty in content moderation systems, shedding light on how automated tools may inadvertently reinforce epistemic closure (Konstas et al. 2024, Urbinato et al. 2025).

Epistemology, therefore, must not retreat into abstraction. It must engage with the concrete architectures of digital environments, analyzing how they shape belief formation and justification. At the same time, fake news studies should integrate epistemological reflection, recognizing that the problem is not only political manipulation or technological misuse, but the configuration of collective cognitive systems.

The challenge is to design and cultivate epistemic environments that remain open to correction, responsive to evidence, and capable of adaptive success. This requires interdisciplinary collaboration across philosophy, cognitive science, media studies, computer science, and policy. It also requires a renewed commitment to the normative ideals of truth and justification—not as relics of a pre-digital age, but as guiding principles for the design of extended cognitive systems.

In confronting information disorder, we are not merely managing content; we are shaping the conditions of collective knowledge. The stakes are therefore not only communicative but epistemic. By reframing disinformation as the pathology of ignorant systems within an extended epistemology, we gain both a deeper diagnosis of the crisis and a more robust conceptual foundation for addressing it in the digital age.

5. Bibliographical References

Airoldi, M. (2021). *Machine Habitus: Toward a Sociology of Algorithms*. New York : Wiley.

Anderson, C. W. (2021). Propaganda, misinformation, and histories of media techniques. Harvard Kennedy School Misinformation Review.

Arfini, S. (2017). *Ignorant Cognition: A Philosophical Investigation*. Cham: Springer.

Arfini, S., & Magnani, L. (2022). *Embodied, Extended, Ignorant Minds. New Studies on the Nature of Not-Knowing*. Cham: Springer.

Bail, C. A. (2021). *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton, NJ: Princeton University Press.

Barberá, P. (2020). Social media, echo chambers, and political polarization. In N. Persily & J. A. Tucker (Eds.), Social Media and Democracy: The State of the Field, Prospects for Reform (pp. 34–55). Cambridge: Cambridge University Press.

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: Lessons from interdisciplinary research. *Annals of the International Communication Association*, 48(2): 139–156.

Capozzi Lupi, A. T. E., Cignarella, A. T., Frenda, S., Lai, M., & Stranisci, M. A. (2023). Debunker Assistant: A support for detecting online misinformation. In *Proceedings of CLiC-it 2023*, Turin: CEUR, pp. 413–420.

Carter, J. A., Clark, A., Kallestrup, J., Palermos, S. O., Pritchard, D. (eds). (2018). *Extended Epistemology*. Oxford: Oxford Academic.

Clark, A. (2017). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Clark, A., Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1): 7–19.

Ferrara, E. (2023). The rise of generative AI and the threat of synthetic misinformation. *Communications of the ACM*, 66(12): 20–23.

Frenda, S., Basile, V., Plank, B., & Stranisci, M. A. (2025). Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*, 59(1), 1–38.

Gorwa, R., & Garton Ash, T. (2020). Democratic transparency in the platform society. In N. Persily & J. A. Tucker (Eds.), Social Media and Democracy: The State of the Field, Prospects for Reform (pp. 286–312). Cambridge: Cambridge University Press.

Konstas, I., Ashraf, S., Ren, A., & Naseem, U. (2024). Large language models for misinformation detection: Challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12: 845–861.

Levin, M. (2019). The computational boundary of a “self”: Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 10, 2688.

Levin, M. (2022). Technological approach to mind everywhere: An experimentally grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, 16, 768201.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, 17(2) : 141–161.

Putnam, H. (1978). Reference and Understanding. In Id., *Meaning and the Moral Sciences*. London: Routledge and Kegan Paul, 97–119.

Resch, M. M., et al. (Eds.). (2026). *Trust and Disinformation*. Cham: Springer.

Sunstein, C. R. (2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ: Princeton University Press.

Theil, S. (2022). Private censorship and structural dominance: Why social media platforms should have obligations to their users under freedom of expression. Cambridge Law Journal, 81(3), 645–672.

Urbinati, A., Lai, M., Frenda, S., & Stranisci, M. A. (2025). Are you sure? Measuring model bias in content moderation through uncertainty. In *Findings of ACL: EMNLP 2025*. Bangkok: ACL, pp. 980–995.

Uscinski, J., Littrell, S., Klofstad C. (2024). The importance of epistemology for the study of misinformation, *Current Opinion in Psychology*, 57: https://doi.org/10.1016/j.copsyc.2024.101789.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation. *Social Media + Society*, 6(1), 1–13.

Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy*. Strasbourg: Council of Europe Publishing.

Population Replacement Conspiracy Theories Detection on Telegram and News Headlines: Benchmarking LLMs and BERT models in Portuguese and Italian

Erik Bran Marino, Renata Vieira
Universidade de Évora, CIDEHUS
Évora, Portugal
erik.marino@uevora.pt, renatav@uevora.pt

Abstract

Disinformation has become a serious threat to the democratic stability of Western societies, with various conspiracy theories spreading from fringe spaces to mainstream media and politics. While some of these theories may seem merely absurd and harmless, others pose significant risks. Among the most dangerous are Population Replacement Conspiracy Theories (PRCTs), which promote the false narrative of a deliberate demographic substitution through immigration. Despite their disinformative nature, increasing widespread and documented connections to extremist violence and political polarization, current computational detection models primarily target COVID-19 or general conspiracy theories, lacking specialized annotated corpora and approaches for identifying PRCTs in multilingual contexts. In this work, we present the first systematic benchmark for PRCT detection in Portuguese and Italian.

1. Introduction

Population Replacement Conspiracy Theories (PRCTs) constitute a family of false narratives claiming deliberate orchestration of demographic substitution through immigration and differential birth rates (Bracke and Hernández Aguilar, 2023). From a theoretical perspective, PRCTs represent an intersection within the *Information Disorder* framework (Wardle and Derakhshan, 2017). While often circulating as *mis-information* when shared by unaware users, their strategic use in political rhetoric frequently aligns with *dis-information*, intentionally disseminated to cause public harm.

These theories, including the Great Replacement, White Genocide, Kalergi Plan, and Eurabia, have demonstrably dangerous real-world consequences, motivating acts of extremist violence such as the Utøya attack (2011) and the Christchurch massacre (2019) (Wojtasik, 2020; Ekman, 2022). Recent research confirms that PRCT endorsement correlates with violent intentions, Anti-Muslim prejudice, and support for exclusionary policies (Obaidi et al., 2022), while their mainstreaming into political discourse raises political polarization and threatens democratic institutions (Marino et al., 2024).

In this work, we present the first systematic benchmark for PRCT detection in Romance languages, evaluating open-weights Large Language Models and fine-tuned encoders across Italian and Portuguese. Our findings reveal that: (1) Deepseek-V3 generally performs best due to its larger configuration; (2) the reasoning-optimized LLM DeepSeek-R1-14B, fine-tuned on Telegram data, achieves the highest performances through cross-domain transfer; and (3) efficient language-specific encoders remain highly competitive for high-throughput applications.

Limited data availability in such an underexplored domain leads to confounded variables; thus, cross-domain results should be interpreted as preliminary hypotheses rather than causal evidence.

2. Related Works

Despite documented harms discussed above, and their frequent proliferation on digital environments, computational detection of PRCTs remains underdeveloped. Existing conspiracy theory detection research focuses predominantly on COVID-19 misinformation, climate discourse, general conspiracy narratives, or platform-specific phenomena (Pogorelov et al., 2021; Miani et al., 2022; Cheatham et al., 2022; Weinzierl and Harabagiu, 2022; Langguth et al., 2023; Gambini et al., 2024; Maggini et al., 2025). While these contributions advance conspiracy detection methodologies, they do not address the specific linguistic and semantic patterns characterizing PRCT discourse.

Moreover, multilingual PRCT detection remains limited, with few isolated efforts on Youtube English datasets (Marino et al., 2025), leaving Portuguese and Italian languages critically underrepresented in computational approaches to PRCT narrative detection. Understanding how computational models perform across these linguistically related but politically differentiated contexts is essential for developing robust automated detection systems capable of monitoring extremist discourse in multilingual transnational spaces.

Cross-domain hate speech detection has demonstrated that models trained on one dis-

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 104–113
12 May 2026.

course domain can transfer to others with varying degrees of success. Toraman et al. (2022) examined transfer learning across five hate domains (religion, gender, race, politics, sports) in English and Turkish, finding that 96% of target domain performance could be recovered by training on other domains in English (92% in Turkish), with gender and religion domains generalizing better than sports-specific content. Markov et al. (2021) demonstrated that stylometric and emotion-based features provide robust cross-domain hate speech detection across English, Slovene, and Dutch, outperforming word and character n-gram features under cross-domain conditions and significantly boosting deep learning models when combined in an ensemble. Pamungkas and Patti (2019) addressed both cross-domain and cross-lingual abusive language detection across ten datasets in English, Italian, Spanish, and German, showing that systems trained on general abusive language datasets produce cross-domain robust systems capable of detecting more specific types of abusive content, and that domain-independent multilingual lexicons (HurtLex) facilitate knowledge transfer between domains and languages.

While these studies establish that cross-domain transfer is feasible for general hate speech categories, whether similar patterns hold for the task of PRCT detection across different media types and Romance languages remains an open empirical question. We address this gap by presenting the first systematic benchmark for PRCT detection in Portuguese and Italian. Our contribution is twofold: (1) we compare model architectures spanning fine-tuned transformer encoders and open-weights large language models, (2) to the best of our knowledge, we provide the first empirical evidence for cross-domain transfer in PRCT detection, comparing in-domain performance against domain shift scenarios.

Our experimental design evaluates performance across two distinct communication paradigms: informal, community-oriented Telegram messages and formal, public-oriented news headlines. Good performance in both domains would suggest a generalizability that could extend to other media, such as webpages, parliamentary speeches, or political manifestos.

3. Methodology

3.1. Datasets

We evaluate models on two annotated corpora regarding migration in European Portuguese and Italian. This language pairing was selected not only because of data availability and their shared Romance language roots, but also because both countries are currently experiencing different but comparable surges in far-right political narratives and PRCT mainstreaming. The corpora represent distinct communicative contexts: Portuguese Telegram messages (informal, conversational) and Italian news headlines (formal, public-oriented). We summarize the key characteristics of both corpora in the *Dataset Overview* panel.

The Portuguese Telegram corpus is a subset of the *Mute Cods* dataset (Laken et al., 2026), extracted from public channels documented as disseminating extremist and conspiracist content. It contains 919 training and 231 test messages with 15.7% PRCT prevalence. Messages average 383 characters. The annotation was conducted by domain experts in linguistics and social sciences following detailed annotation guidelines. To ensure consistency, data were at least double-annotated, with disagreements resolved through discussion to refine the guidelines, achieving overall a moderate inter-annotator agreement (Krippendorff's $\alpha = 0.58$).

The Italian news corpus is derived from the *PartisanLens* dataset (Maggini et al., 2026), which collected immigration-related headlines programmatically via Media Cloud between 2020 and 2024. The subset used comprises the Italian-only data with 434 training and 131 test headlines (averaging 81 characters), of which in total 161 are PRCT. This dataset was annotated for PRCT by two independent experts, with a third senior annotator resolving any disagreements, resulting in a substantial agreement (Krippendorff's $\alpha = 0.774$).

The observed inter-annotator agreement levels align with the inherent complexity of annotating multidimensional social constructs like PRCTs. As Matamoros-Fernández and Farkas (2021) highlight, moderate agreement is standard in social media research due to the ambivalence of online communication and context collapse. Our Telegram agreement ($\alpha = 0.58$) specifically reflects this reality, where conspiratorial narratives frequently blend with humor, irony, and informal vernacular. In such subjective tasks, chance-corrected metrics capture genuine human interpretive variation of implicit framing rather than poor annotation quality (Plank, 2022; Hemm et al., 2024).

Dataset Overview

Corpus	Lang	Train	Test	PRCT%	Len	IAA
Telegram	PT	919	231	15.7	383	α=0.58
News	IT	434	131	28.5	81	α=0.77

Details: *Len* indicates the average character length per sample; *IAA* represents the Inter-Annotator Agreement measured via Krippendorff's α.

3.2. Models

We evaluate seven model architectures across different learning paradigms: four large language models (Mistral-7B-Instruct-v0.3, Phi-4 14B, DeepSeek-R1-14B, DeepSeek-V3) and three transformer encoders (XLM-RoBERTa-base 270M multilingual, Albertina-280M Portuguese-specific, UmBERTo-110M Italian-specific). Depending on the experimental phase, models are evaluated either in a zero-shot setting or after supervised fine-tuning (using LoRA for LLMs and standard fine-tuning for encoders). Our experimental process is systematically divided into four distinct phases.

3.3. Experimental Design

The task is formally modeled as a binary text classification problem, where a given input text must be classified as either containing PRCT narratives (positive class) or not (negative class).

Phase 1 establishes the zero-shot baseline performance. We evaluate all four LLMs (Mistral-7B, Phi-4, DeepSeek-R1-14B, DeepSeek-V3) in their vanilla configurations separately on each test set. DeepSeek-V3 was specifically included to provide a performance upper bound against a state-of-the-art (SOTA) model. This identifies the inherent reasoning capabilities of each model prior to any task-specific adaptation. To verify robustness beyond limited test sets (n=231 PT, n=131 IT), we additionally evaluate Mistral-7B and DeepSeek-R1-14B zero-shot on complete datasets (train+test combined, n=1150 PT, n=565 IT) to check whether overall the performances keep being consistent.

Phase 2 investigates domain-specific fine-tuning. We implement LoRA fine-tuning exclusively on the models that demonstrated strong potential in Phase 1 and whose parameter size allowed feasible training on our hardware infrastructure. Specifically, Mistral-7B and DeepSeek-R1-14B were selected for fine-tuning. Phi-4 was excluded due to its uncompetitive zero-shot performance, while DeepSeek-V3 was excluded because its massive parameter size rendered fine-tuning computationally prohibitive in our setup. We fine-tuned these LLMs, alongside language-specific encoders (UmBERTo for Italian, Albertina for Portuguese), on single-domain configurations (Telegram-only and News-only) to evaluate in-domain versus cross-domain transferability.

Phase 3 evaluates merged-domain fine-tuning. We test whether combining both datasets during training yields domain-invariant representations that improve overall classification. We evaluate Mistral-7B, DeepSeek-R1-14B, and the multilingual encoder XLM-RoBERTa trained on the merged corpus, comparing the results against the domain-specific models from Phase 2.

Phase 4 constructs a Pareto frontier analysis across all evaluated configurations to identify non-dominated models optimizing distinct regions of the accuracy-efficiency trade-space. This allows us to find optimal model choices depending on the priority of the task (maximum accuracy, balanced performance, or maximum throughput).

We report macro-averaged F1 (equal class weighting) and binary F1 (positive class focus) as primary metrics, alongside accuracy, precision, and recall. LLMs use temperature 0 for deterministic outputs with structured prompts defining task requirements and output format. Fine-tuned encoders use class-weighted cross-entropy loss with learning rate 2e-5, batch size 16, and early stopping on validation F1-macro. All experiments use seed 42 for reproducibility. LoRA employed rank 8, alpha 16 on 16 layers, learning rate 1e-5, batch size 2, 600 iterations.

3.4. Computational Infrastructure

With the exception of DeepSeek-V3, which was accessed via official API calls due to its massive parameter size (671 B), all experiments were conducted locally on a MacBook Pro with an Apple M3 Max chip (16-core CPU, 40-core GPU) and 64 GB of unified memory. Local LLM inference utilized the MLX framework optimized for Apple Silicon, while encoder fine-tuning employed PyTorch.

4. Results

4.1. Phase 1: Zero-Shot Baseline Performance

Phase 1 established the inherent capabilities of LLMs prior to fine-tuning (Table 1). DeepSeek-V3 demonstrated the strongest zero-shot performance on Italian News (F1-macro 0.879) and good performances on Portuguese Telegram (F1-macro 0.758). We do not take into account Deepseek-V3 time-analysis, as it was run on a different setup compared to the other models. DeepSeek-R1-14B followed closely, showing highly competitive reasoning capabilities (F1-macro 0.870 on IT; 0.781 on PT: the highest result), albeit with significantly slower inference due to its chain-of-thought architecture. Mistral-7B offered a balanced trade-off, maintaining good F1-macro scores (0.702 IT, 0.744 PT) with the lowest inference latency among LLMs (0.64s–0.80s). Conversely, Phi-4 exhibited overly conservative classification behavior, resulting in high false negative rates and the lowest F1-macro scores across both sets (0.611 IT, 0.697 PT). Full dataset evaluation confirmed test set representativeness: Mistral-7B maintained F1-macro stability (test: 0.702/0.744 vs full: 0.704/0.704), as

		News ITA						Telegram PT					
Model	**Type**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**
Phi-4	Vanilla	0.763	0.873	0.613	0.611	0.367	8.39s	0.879	**0.819**	0.656	0.697	0.462	7.67s
Mistral-7B	Vanilla	0.771	0.737	0.688	0.702	0.559	**0.64s**	**0.887**	0.816	0.707	0.744	0.552	**0.80s**
DeepSeek-R1-14B	Vanilla	0.893	0.883	0.860	0.870	0.816	12.86s	0.866	0.752	**0.830**	**0.781**	**0.643**	15.10s
DeepSeek-V3	Vanilla	**0.901**	**0.896**	**0.866**	**0.879**	**0.827**	1.69s	**0.887**	0.802	0.729	0.758	0.581	1.78s

Table 1: Phase 1: Zero-shot baseline performance of Large Language Models. Bold indicates best performance per test set.

did DeepSeek-R1-14B (test: 0.870/0.781 vs full: 0.820/0.752).

4.2. Phase 2: Domain-Specific Fine-Tuning

Building upon Phase 1, we proceeded to fine-tune Mistral-7B and DeepSeek-R1-14B. DeepSeek-V3 and Phi-4 were excluded from this and subsequent phases due to, respectively, hardware limitations and poor baseline performance. Table 2 presents the results of models trained strictly on single domains, including language-specific encoders.

Mistral-7B benefited consistently from domain-matched training: Telegram-only fine-tuning elevated its F1-macro to 0.819 on Portuguese evaluation, while News-only fine-tuning reached 0.781 on Italian headlines. However, this accuracy gain incurred a substantial inference cost (4.5-4.7s per sample vs 0.64-0.80s zero-shot). DeepSeek-R1-14B demonstrated exceptional cross-domain performance: when fine-tuned strictly on Portuguese Telegram, it achieved F1-macro 0.892 on Italian news: the highest score observed across all experiments. Paradoxically, this same model degraded on its own training domain (F1-macro 0.655 on PT Telegram, below its zero-shot baseline of 0.781). This suggests architectural sensitivity where reasoning-optimized models may suffer forgetting on informal, fragmented content while successfully generalizing conspiratorial structures to formal discourse.

Language-specific encoders showed mixed results: Albertina achieved strong in-domain performance on Portuguese Telegram (F1-macro 0.774) with exceptional speed (0.018s), whereas UmBERTo did not performe well on Italian news (F1-macro 0.589).

4.3. Phase 3: Merged-Domain Fine-Tuning

To test whether combining training data from both languages and domains would yield a more robust, generalized classifier, we evaluated Mistral-7B, DeepSeek-R1-14B, and the multilingual encoder XLM-RoBERTa on a merged training corpus (Table 3). The findings indicate that merging datasets did not yield significant improvements and, in several cases, degraded performance.

For Mistral-7B, merged training resulted in F1-macro 0.734 on Italian News and 0.770 on Portuguese Telegram. These scores fall short of the peak performances achieved via single-domain fine-tuning (0.781 on News FT and 0.819 on TG FT, respectively). DeepSeek-R1-14B exhibited a similar trend, dropping to F1-macro 0.833 on News and 0.633 on Telegram under the merged configuration. The multilingual encoder XLM-RoBERTa achieved modest scores (F1-macro 0.661 IT, 0.670 PT) but stood out for its remarkable efficiency (0.005s per sample). Overall, the data suggests that in PRCT detection, exposing models to highly heterogeneous stylistic registers (formal vs. informal) and languages simultaneously may introduce interference rather than constructive transfer.

4.4. Phase 4: Pareto-Optimal Model Selection

Phase 4 constructs the Pareto frontier separately for each language to identify optimal deployment strategies for large-scale annotation. DeepSeek-V3 was explicitly excluded from this analysis: while it served as a SOTA benchmark in Phase 1, its reliance on external API calls makes its inference latency incomparable to locally executed models, rendering any efficiency comparison unfair. A model configuration is Pareto-optimal if no alternative achieves both higher F1-binary (positive class) and faster inference within the target language evaluation. Table 4 presents Pareto-optimal language configurations across all tested models.

For Italian news headlines, four configurations occupy the Pareto frontier. XLM-RoBERTa provides maximum throughput (200 samples/second) at moderate accuracy (F1-binary 0.568). Mistral-7B with Telegram LoRA achieves balanced performance (F1-binary 0.688, 4.50s per sample), representing a 12 percentage point improvement over XLM-RoBERTa with 900-fold slower inference. DeepSeek-R1-14B zero-shot occupies the quality-focused tier (F1-binary 0.816, 12.86s), while DeepSeek with Telegram LoRA achieves

		News ITA						Telegram PT					
Model	**Type**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**
Mistral-7B	TG FT	0.771	0.748	0.786	0.753	0.688	4.50s	**0.896**	0.797	**0.848**	**0.819**	**0.700**	4.62s
Mistral-7B	News FT	0.840	0.869	0.752	0.781	0.667	4.72s	**0.896**	**0.824**	0.746	0.777	0.613	4.73s
DeepSeek-R1-14B	TG FT	**0.908**	**0.892**	**0.892**	**0.892**	**0.850**	16.67s	0.853	0.716	0.630	0.655	0.393	15.94s
DeepSeek-R1-14B	News FT	0.901	0.881	0.887	0.884	0.840	15.90s	0.857	0.726	0.689	0.705	0.492	16.34s
UmBERTo	News FT	0.710	0.644	0.588	0.589	0.367	**0.005s**	—	—	—	—	—	—
Albertina	TG FT	—	—	—	—	—	—	0.874	0.761	0.790	0.774	0.623	**0.018s**

Table 2: Phase 2: Performance of models fine-tuned on single domains (Telegram or News). Bold indicates best performance per test set.

		News ITA						Telegram PT					
Model	**Type**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**
Mistral-7B	Merged FT	0.817	**0.871**	0.707	0.734	0.586	4.52s	**0.900**	**0.860**	0.726	**0.770**	**0.597**	4.58s
DeepSeek-R1-14B	Merged FT	**0.855**	0.827	**0.840**	**0.833**	**0.771**	16.20s	0.823	0.648	0.623	0.633	0.369	15.60s
XLM-RoBERTa	Merged FT	0.687	0.660	0.684	0.661	0.568	**0.005s**	0.749	0.663	**0.783**	0.670	0.508	**0.005s**

Table 3: Phase 3: Performance of models fine-tuned on the merged dataset (News + Telegram). Bold indicates best performance per test set.

maximum accuracy (F1-binary 0.850, 16.67s), the highest score observed across all Italian evaluations.

For Portuguese Telegram messages, three distinct configurations emerge. XLM-RoBERTa provides maximum throughput but with lower accuracy (F1-binary 0.508, 0.005s). Albertina fine-tuned on Telegram offers Portuguese-specific optimization (F1-binary 0.623, 0.018s), processing 55.6 samples per second while achieving 11 percentage point improvement over XLM-RoBERTa. Mistral-7B with Telegram LoRA FT reaches maximum accuracy (F1-binary 0.700, 4.62s), representing the optimal choice for quality-critical Portuguese Telegram annotation tasks with 26-fold slower inference than Albertina but 7.7 percentage point F1 improvement.

The choice among Pareto-optimal models depends on operational constraints and target language. For News Headlines Italian data annotation, DeepSeek-R1-14B with Telegram LoRA provides maximum accuracy (F1-binary 0.850) suitable for quality-critical applications, while Mistral-7B with Telegram LoRA offers balanced performance at faster inference. For Portuguese data annotation, Mistral-7B with Telegram LoRA achieves maximum accuracy (F1-binary 0.700), while Albertina provides adequate performance with great speed advantage for throughput-critical workflows. All other evaluated configurations are Pareto-dominated within their respective language evaluations.

5. Error Analysis

Zero-shot LLMs demonstrate conservative behavior with high Precision and lower Recall, minimizing false positives at the cost of missing true PRCT instances. Conversely, fine-tuned encoders show inverse patterns, with lower Precision but higher Recall, capturing more true positives while also accepting more false positives. Error rates demonstrate contrasting relationships with text length across the two domains. Italian news headlines show decreasing error rates as length increases, indicating that longer headlines provide sufficient context. Portuguese Telegram messages exhibit the inverse pattern, where PRCT claims embedded within extended narratives become diluted.

Cross-domain transfer reveals unexpected asymmetries. Mistral-7B fine-tuned exclusively on Portuguese Telegram achieves competitive performance on Italian news (F1-macro 0.753). Qualitative examination shows coverage of implicit formulations, such as weaponization metaphors, elite manipulation claims, and presuppositional framing. One hypothesis is that exposure to explicit conspiratorial language during training facilitates recognition of implicit PRCT formulations. DeepSeek-R1-14B (TG FT) demonstrates markedly different behavior, achieving exceptional performance on formal news (F1 0.892) but degrading on the training domain itself. This suggests architectural sensitivity where reasoning-optimized models benefit from cross-domain exposure on formal discourse but suffer forgetting on informal content.

To systematize our qualitative findings, Table

Model Configuration	Type	$F1_B$	Time/sample	Samples/sec	Recommended for
Italian News Headlines					
XLM-RoBERTa	Merged FT	0.568	0.005s	200.0	Maximum throughput
Mistral-7B	TG FT	0.688	4.50s	0.22	Balanced performance
DeepSeek-R1-14B	TG FT	**0.850**	16.67s	0.06	Maximum accuracy
Portuguese Telegram Messages					
XLM-RoBERTa	Merged FT	0.508	0.005s	200.0	Maximum throughput
Albertina	TG FT	0.623	0.018s	55.6	Balanced performance
Mistral-7B	TG FT	**0.700**	4.62s	0.22	Maximum accuracy

Table 4: Language-specific Pareto-optimal models identified through Phase 4 analysis. We highlight configurations suitable for distinct application scenarios: maximum accuracy favors offline social science analysis, while maximum throughput suits real-time content flagging. Each section shows dominant configurations for the target language (no alternative achieves both higher $F1_B$ and faster inference).

5 categorizes the most frequent error patterns and classification behaviors across models and domains. Model-specific error patterns reveal distinct behavioral profiles. For instance, Phi-4 exhibits extreme conservatism, missing 77.5% of PRCT instances, including headlines with literal replacement terminology. Mistral-7B demonstrates balanced behavior but struggles with distinguishing fear-inducing rhetoric from actual conspiratorial framing, occasionally missing military metaphors. Albertina exhibits domain-appropriate sensitivity for Portuguese informal content but struggles with historical PRCT references and demographic complaint rhetoric. DeepSeek-R1-14B fine-tuned on Telegram systematically misses serialized documentary content and specific dog-whistles (e.g., remigration) in its own fine-tuning domain.

A prominent source of false positives across all models occurs when they encounter affective language, fear-inducing rhetoric, or violent imagery that lacks the essential element of conspiratorial orchestration. Conversely, false negatives exhibit more diverse patterns: models systematically miss conspiracy theories expressed through indirect discussion, presupposition, or serialized content that distributes conspiratorial claims across multiple installments, requiring contextual understanding beyond a single-message scope.

6. Discussion

This work situates PRCTs within the broader framework of Information Disorder (Wardle and Derakhshan, 2017). Specifically, PRCTs typically operate as *dis-information*—false narratives deliberately created to inflict harm or sow societal division—though they are frequently propagated by genuine believers as *mis-information*. Understanding the contextual and cultural factors that differentiate formal news propagation in Italy from informal, community-driven Telegram messaging in Portugal is essential for developing comprehensive NLP systems capable of addressing the full spectrum of information disorders.

Models fine-tuned on Portuguese Telegram achieve higher F1 scores when evaluated on Italian news headlines than models trained on Italian news itself. Mistral TG FT reaches F1=0.753 on Italian news versus F1=0.688 for News FT and F1=0.729 for Merged FT. DeepSeek TG FT achieves F1=0.892 on Italian news, the highest score across all configurations.

This pattern is consistent with multiple non-exclusive hypotheses: (1) Telegram's explicit conspiratorial discourse provides richer linguistic signals than news headlines' implicit formulations; (2) longer texts (Telegram avg. 383 chars vs news 81 chars) enable more robust feature learning; (3) Italian linguistic features facilitate classification regardless of training corpus; (4) class distribution effects (28.5% PRCT in news vs 15.7% in Telegram) interact with model calibration.

DeepSeek-R1-14B exhibits paradoxical behavior: its Telegram-trained variant achieves F1=0.892 on Italian news (cross-domain evaluation) yet deteriorates to F1=0.655 on Portuguese Telegram (in-domain), performing worse than its own zero-shot baseline (F1=0.781). This pattern suggests forgetting during fine-tuning on informal discourse, where the model's chain-of-thought optimization for structured reasoning may conflict with Telegram's fragmented linguistic patterns. Conversely, Mistral-7B benefits consistently from domain-matched training across both datasets, with in-domain fine-tuning yielding monotonic improvements over zero-shot baselines.

Finally, the operational trade-off between predictive performance and computational cost dictates distinct deployment strategies. While LLMs offer superior reasoning for complex narratives, their

Error Pattern	Model (Behavior)	Domain	Representative Example
Implicit Formulations (Weaponization, Elites)	Mistral TG FT [True Positive]	News	*Germania li usa come armi per destabilizzarci* (Germany uses them as weapons to destabilize us)
Over-Conservatism (Literal replacement terms)	Phi-4 [False Negative]	News	*La sostituzione etnica di Meloni: 830mila immigrati* (Meloni's ethnic replacement: 830 thousand immigrants)
Fear-Inducing Rhetoric (Without orchestration)	Mistral-7B, Encoders [False Positive]	News	*Immigrati islamici pronti a colpire in Italia* (Islamic immigrants ready to strike in Italy)
Demographic Complaints (Historical context)	Albertina [False Negative]	TLgram	*Número de imigrantes nas escolas públicas aumenta 47% em dois anos* (Number of immigrants in public schools increases 47% in two years)
Serialized Content (& Specific dog-whistles)	DeepSeek TG FT [False Negative]	TLgram	*REMIGRΛZÓN ÚNICΛ SOLUZÓN* (REMIGRATION ONLY SOLUTION)
Reporting/Discussing (Presuppositional framing)	All Models [False Negative]	News	*Migranti e sostituzione etnica: critiche bipartisan a Lollobrigida* (Migrants and ethnic replacement: bipartisan criticism of Lollobrigida)

Table 5: Systematization of the most common error patterns and classification behaviors identified in the qualitative analysis.

inference latency renders them less suitable for real-time monitoring of high-velocity streams compared to efficient encoders like XLM-RoBERTa. As an example application, we could use a tiered architecture. Encoders would filter massive datasets first, passing only the ambiguous data to LLMs for detailed checking. While LLMs demonstrate superior performance in our experiments, recent work demonstrates that hybrid approaches combining shallow learning with theoretically-grounded features can outperform LLMs on related discourse tasks (Bassi et al., 2025). Given that, as Marino et al. (2025) found, PRCT discourse exhibits distinctive linguistic markers such as religious language, power dynamics, higher negative tone and conflict framing, future work should investigate whether operationalizing these patterns as explicit features could enhance detection while reducing computational costs.

7. Conclusions

This work provides a comparative analysis of large language models and fine-tuned encoders for Population Replacement Conspiracy Theory detection, offering the first empirical benchmark of their cross-corpus performance between informal Portuguese and formal Italian discourse contexts. We evaluated seven model architectures across informal Telegram messages and formal news headlines, establishing baseline performance metrics and identifying cross-corpus evaluation patterns.

A model fine-tuned on Portuguese Telegram data achieve F1-binary 0.850 (DeepSeek-R1-14B) when evaluated on Italian news headlines. This cross-corpus performance is consistent with domain transfer but could equally reflect linguistic, length-based, or distributional factors given the systematic confounding in our design. Controlled experiments with balanced language-domain combinations are required to isolate causal mechanisms.

Pareto frontier analysis identified language-specific optimal configurations balancing accuracy and computational efficiency: DeepSeek TG FT for Italian data (F1-binary 0.850, 16.67s per sample) and Mistral TG FT for Portuguese data (F1-binary 0.700, 4.62s per sample). These models provide practical deployment options for large-scale annotation (for instance, high-volume documents for offline social analysis), with encoder alternatives (XLM-RoBERTa, Albertina) offering throughput-critical solutions where speed outweighs accuracy requirements, which may be a better choice for online or run-time PRCT spotting. Future work should expand coverage to additional languages, model architectures, and test generalization to other media formats.

8. Limitations

A primary limitation of this study, as discussed in Section 6, is the systematic confounding of language, domain, text length, and class distribu-

System Prompt Configuration

You are a classifier specialized in detecting Population Replacement Conspiracy Theories (PRCT) in text. Your task is to determine if the given text contains or supports PRCT narratives.

DEFINITION OF PRCT:
Population Replacement Conspiracy Theories claim that native populations in Western countries are being systematically replaced by non-European (non-Christian, non-white) peoples through immigration and demographic changes as part of a deliberate orchestrated plan.

INDICATORS OF PRCT:

- Explicit references to "Great Replacement", "White Genocide", "Kalergi Plan", "Eurabia"
- Dog whistles: "Fourteen Words", "1488", "ZOG", "remigration", "demographic warfare", "invasion"
- Framing immigration as "invasion", "ethnic substitution", or "cultural erasure"
- Claims of orchestrated plans by elites/globalists to replace native populations
- Suggesting demographic changes are intentional and part of a hidden agenda

CLASSIFICATION:

- **YES:** The text mentions, supports, or alludes to PRCT narratives
- **NO:** The text does not contain PRCT content (may discuss immigration without conspiracy framing)

IMPORTANT:

- You will only see the text of the message. Base your decision only on that.
- General anti-immigration sentiment WITHOUT conspiracy elements = NO
- Concerns about demographic changes WITHOUT deliberate orchestration claims = NO
- If dog whistles are present, then, it's likely = YES

OUTPUT FORMAT: Respond with a valid JSON object:

{ "prct": "YES" or "NO" }

TEXT TO CLASSIFY: [[TEXT]]

Figure 1: The system prompt used for zero-shot classification with LLMs.

tion, which restricts our ability to isolate the specific drivers of cross-corpus performance. Consequently, our findings regarding cross-domain transfer should be interpreted strictly as preliminary hypotheses rather than definitive evidence.

Furthermore, the computational demands of fine-tuning LLMs imposed strict constraints on our experimental setup. Consequently, LoRA adaptations were executed using a single initialization seed. While reporting single-run performance is a pragmatic and standard approach for establishing initial baselines in resource-constrained LLM research, it inherently precludes the calculation of cross-run variance and formal statistical significance testing. We encourage future research with expanded computational budgets and data size to conduct multi-seed evaluations to establish robust confidence intervals. Plus, while LoRA substantially reduces computational costs and memory requirements, enabling experimentation across multiple model architectures, prior work demonstrates that LoRA can underperform full fine-tuning on certain tasks, particularly those requiring extensive parameter updates for domain adaptation. The magnitude of this performance gap in our PRCT detection context remains unquantified. Temporal coverage spans 2021–2024 without diachronic analysis, though PRCT narratives demonstrably evolve across decades. This temporal constraint limits conclusions about model robustness to evolving conspiracy formulations. Finally, the zero-shot evaluation employed a single English prompt template without few-shot example selection. This methodological choice was motivated by our objective to establish a uniform, rigorous baseline performance generalizable to deployment contexts lacking annotated data, given that LLMs often exhibit more stable alignment with English instructions even on multilingual tasks. A systematic prompt comparative study—including testing the same models under plain instruction, language-specific prompts (Italian and/or Portuguese), and few-shot or chain-of-verification (CoVe) prompts—is left as a next step for future research.

9. Ethical Considerations

This research involves the analysis of extremist narratives and conspiracy theories, which inherently include discriminatory and harmful language. To ensure privacy and ethical compliance, all Telegram messages were sourced exclusively from publicly accessible channels through official APIs, and no personally identifiable information was retained or analyzed. The Italian news headlines consist entirely of public journalistic records. Furthermore, given the sensitive and potentially distressing nature of the textual data, the annotation process was conducted exclusively by domain experts in linguistics and social sciences, with periodic meetings, rather than vulnerable and isolated crowd-workers, thereby attempting to mitigate the psychological risks associated with exposure to extremist content. Finally, we explicitly acknowledge the risk of false positives, such as misclassifying legitimate, albeit polarized, political discourse on immigration as conspiratorial. Consequently, these systems are intended to serve as assistive diagnostic tools that require human-in-the-loop oversight, rather than fully autonomous moderation mechanisms. All research activities were carried out in accordance with the ethical guidelines and with the formal approval of the research center's Ethical Review Board.

Acknowledgments

This work was supported by the HYBRIDS project, which has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee (Grant Number: EP/X036758/1). The work is partially supported by the Portuguese Science Foundation as part of the projects CEECIND/ 01997/2017 and UIDP/00057/2025. The content of this work reflects only the authors' view and the funding agencies are not responsible for any use that may be made of the information it contains.

Bibliography

Davide Bassi, Erik Bran Marino, Renata Vieira, and Martin Pereira. 2025. Old but gold: LLM-based features and shallow learning methods for fine-grained controversy analysis in YouTube comments. In *Proceedings of the 12th Argument mining Workshop*, pages 46–57, Vienna, Austria. Association for Computational Linguistics.

Sarah Bracke and Luis Manuel Hernández Aguilar. 2023. *The Politics of Replacement: From "Race Suicide" to the "Great Replacement"*. Routledge, London.

Susan Cheatham, Per E Kummervold, Lorenza Parisi, Barbara Lanfranchi, Ileana Croci, Francesca Comunello, Maria Cristina Rota, Antonietta Filia, Alberto Eugenio Tozzi, Caterina Rizzo, et al. 2022. Understanding the vaccine stance of italian tweets and addressing language changes through the covid-19 pandemic: Development and validation of a machine learning model. *Frontiers in Public Health*, 10:948880.

Mattias Ekman. 2022. The great replacement: Strategic mainstreaming of far-right conspiracy claims. *Convergence*, 28(4):1127–1143.

Margherita Gambini, Serena Tardelli, and Maurizio Tesconi. 2024. The anatomy of conspiracy theorists: unveiling traits using a comprehensive twitter dataset. *Computer Communications*, 217:25–40.

Ashley Hemm, Sandra Kübler, Michelle Seelig, John Funchion, Manohar Murthi, Kamal Premaratne, Daniel Verdear, and Stefan Wuchty. 2024. Are you serious? handling disagreement when annotating conspiracy theory texts. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 124–132, St. Julians, Malta. Association for Computational Linguistics.

Katarina Laken, Erik Bran Marino, Paloma Piot, Davide Bassi, Søren Fomsgaard, Michele Maggini, Renata Vieira, Marcos Garcia, and Sara Tonelli. 2026. Mute Cods: A Multilingual Telegram Dataset with Benchmark Models for Conspiracy Theory Detection. In *Proceedings of the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2026)*. Forthcoming.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, 6(2):443–484.

Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo. 2025. Detecting hyperpartisanship and rhetorical bias in climate journalism: A sentence-level Italian dataset. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*,

pages 168–187, Vienna, Austria. Association for Computational Linguistics.

Michele Joshua Maggini, Paloma Piot, Anxo Pérez, Erik Bran Marino, Lúa Santamaría Montesinos, Ana Lisboa Cotovio, Marta Vázquez Abuín, Javier Parapar, and Pablo Gamallo. 2026. Partisanlens: A multilingual dataset of hyperpartisan and conspiratorial immigration narratives in european media. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1186.

Erik Marino, Jesus M. Benitez-Baleato, and Ana Sofia Ribeiro. 2024. The polarization loop: How emotions drive propagation of disinformation in online media—the case of conspiracy theories and extreme right movements in southern europe. *Social Sciences*, 13(603).

Erik Bran Marino, Davide Bassi, and Renata Vieira. 2025. Linguistic markers of population replacement conspiracy theories in youtube immigration discourse. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 670–679.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & new media*, 22(2):205–224.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2022. Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, 54(4):1794–1817.

Milan Obaidi, Jonas Kunst, Simon Ozer, and Sasha Y Kimel. 2022. The "great replacement" conspiracy: How the perceived ousting of whites can evoke violent extremism and islamophobia. *Group Processes & Intergroup Relations*, 25(7):1675–1695.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. Wico text: a labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In *Proceedings of the 2021 workshop on open challenges in online social networks*, pages 21–25.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Maxwell Weinzierl and Sanda Harabagiu. 2022. Vaccinelies: A natural language resource for learning to recognize misinformation about the covid-19 and hpv vaccines. *arXiv preprint arXiv:2202.09449*.

Karolina Wojtasik. 2020. Utøya–christchurch–halle. right-wing extremists' terrorism. *Security Dimensions. International and National Studies*, (33):84–97.

Mapping Discourse Reframing: A Multi-Layer Network Approach to Italian HPV Vaccine Discourse on X (2010-2024)

Lorella Viola
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
l.viola@vu.nl

Abstract

Understanding how online narratives travel through coalitions is critical for identifying information disorder, yet computational analyses often rely on conservative network constructions that erase initially sparse but salient signals. This paper proposes a novel multi-layer framework that captures low-frequency signals of emerging information disorder allowing for locating where online discourse is reframed and amplified over time. The use case is 14 years of Italian discourse on X regarding the Human Papillomavirus (HPV) vaccine across three pivotal epochs (2010–2024). Utilizing hashtag co-occurrence networks, we introduce a dual-layer approach. We first identify robust core discourse coalitions through conservative community detection, revealing a stable prevention-oriented backbone contrasted with increasingly separable skepticism coalitions. We then introduce a 'coverage' layer and project fringe hashtags into core coalitions based on weighted connectivity. Using a manually labelled set of skeptical and conspiratorial seed tweets, we demonstrate that this core–coverage projection significantly improves the recovery of long-tail, problematic hashtags while preserving an interpretable coalition structure. Our findings characterize the structural maturation of polarized narratives and provide a methodology for mapping how discourse is reframed and amplified by information disorder over time.

Keywords: HPV vaccine, information disorder, X, online community mapping, hashtags, health discourse, conspiracy

1. Introduction

Discourses rise and fall in popularity not arbitrarily, but because they align with the prevailing social, political, and economic contexts in which they are created and perpetuated (Robertson, 1998). For example, public conceptualisations and discussions of health reflect more than just health: they also offer privileged viewpoints to understand how ideas, beliefs and explanations for what health is and what determines it change over time, for example, in response to increased scrutiny or societal shifts. In turn, understanding the structural and temporal evolution of health discourse is essential to identify mechanisms of information disorder and inform democratic countermeasures. This is particularly important where harms include public health risks such as the erosion of trust in medical recommendations.

Especially in digital ecosystems, misleading, conspiratorial, or polarizing claims rarely propagate in isolation (Viola, 2025; Carpiano et al., 2023; Durmaz and Hengirmen, 2022; Quintana et al., 2022; Gunaratne et al., 2019); instead, they travel through narrative coalitions that infiltrate core public health narratives and reframe topics, institutions, and responsibilities. Despite the exponential growth in scholarship mapping the impact and content of disinformation (Wiggins, 2023; Bonnevie et al., 2021; Chen et al., 2021; Stabile et al., 2019; Van Prooijen and Van Vugt, 2018), a significant computational challenge remains: information disorder signals often reside in the 'long tail' of discourse, that is a large number of niche or short-lived elements which in a communication ecosystem individually appear only a few times, but collectively make up a substantial share of discourse diversity (Kordumova et al., 2016). Because emerging disinformation or conspiratorial narratives may initially present as sparse, low-frequency signals, traditional network analyses that rely on conservative filtering (e.g., high edge-count thresholds) risk erasing the very salient signals that characterize the early stages of narrative reframing. This erasure limits our ability to observe how fringe skepticism matures into robust, structurally separable coalitions that can eventually contest the prevention-oriented backbone of public health.

This paper addresses this gap by proposing a novel multi-layer framework designed to capture low-frequency signals without sacrificing the interpretability of the core discourse. The use case is 14 years of Italian discourse on X (formerly Twitter) surrounding the human papillomavirus (HPV) vaccine. This vaccine has historically been a lightning rod for anti-vaccination campaigns and gender-based disinformation (Viola, 2025; Kornides et al., 2023; Calo et al., 2021) offering a unique longitudinal lens into how discourse reframing undermines public trust. Specifically, this study maps the structural maturation of narrative coalitions as manifested through hashtags in Italian posts about

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 114–125
12 May 2026.

HPV across three pivotal epochs (2010–2014, 2015–2019, 2020–2024). The assumption is that hashtags can be interpreted as cultural products (La Rocca and Boccia Artieri, 2022; Zappavigna, 2016; Weller et al., 2013) and that these cultural products act as markers of conceptual associations and communal identity. Using hashtag co-occurrence networks, we introduce a dual-layer approach: a 'core' layer defined by conservative community detection (k-core ≥ 2) to identify the stable thematic backbone, and a 'coverage' layer (edge count ≥ 1) that captures the long tail. By projecting these fringe hashtags into the core coalitions based on weighted connectivity, we provide a methodology for locating where and how health narratives are reframed and amplified over time. Using a manually labelled set of skeptical and conspiratorial seed posts, we demonstrate that this projection significantly improves the recovery of problematic hashtags that hijack health narratives and that would otherwise be lost in traditional network constructions.

2. Previous studies

Research on health discourse has long emphasized its temporal dimension, showing how shifts in public sentiment and dominant narratives can shape, sometimes rapidly, policy agendas and institutional responses (McPhail and Bombak, 2015; Berridge, 2007; Lawrence, 2004). This work highlights the dynamic interplay between public priorities, political decision-making, and the framing of responsibility and risk. A well-known example is the case of the HPV vaccine in Japan, when following a surge in media reports of unconfirmed adverse events, often framed through conspiracy-adjacent narratives of government and pharmaceutical collusion, in 2013 the Japanese Ministry of Health, Labour and Welfare (MHLW) suspended its proactive recommendation for the vaccine (Yagi et al., 2024). This suspension lasted nearly nine years, ending only in April 2022 and it is frequently cited in academic literature as a cautionary tale of how information disorder and safety concerns can lead to the total collapse of a public health program (Larson and Broniatowski, 2021).

Studies foreground how health meanings evolve over time and why such evolution matters for understanding both governance and public contestation around health. With specific reference to vaccine discourse, recent work has examined the growth and transformation of anti-vaccine activism and the changing structure of online vaccine discussion (Britt, 2023; Carpiano et al., 2023; Crupi et al., 2022; Durmaz and Hengirmen, 2022; Mønsted and Lehmann, 2022; Quintana et al., 2022). Across studies, several recurring patterns have emerged: increasing attention to vaccination topics online; highly polarised interaction dynamics in which users preferentially engage with like-minded others; and a growing alignment between vaccine skepticism and political identity, including far-right affiliation. Importantly, research has shown that anti-vaccination groups, though smaller than pro-vaccination communities, can account for a disproportionate share of content and remain highly active (Gunaratne et al., 2019). These dynamics are often mediated by trust: information tends to be accepted not primarily on accuracy grounds, but on whether it circulates within trusted in-groups (Quintana et al., 2022). This important literature, however, has often addressed vaccine discourse at a broad level (without focusing on a specific vaccine) or concentrates primarily on COVID-19 vaccination. Moreover, English-language studies substantially outnumber those in other languages, limiting our understanding of how vaccination narratives evolve in different sociolinguistic and political contexts. This motivates further work on non-English case studies and on vaccine-specific discourses, such as the Italian online debate around the HPV vaccine.

In terms of online cultural artifacts, scholarship on hashtags has expanded significantly over the past two decades (La Rocca and Boccia Artieri, 2022). While hashtags originated as indexing tools for grouping content, research shows that they have evolved into multi-functional instruments that operate simultaneously as thematic markers, navigation devices, and forms of self-expression (e.g., #metoo). Because hashtags combine descriptive signals with affective and communicative meanings, they can be studied through linguistic, cultural, and media perspectives. A growing body of work argues that hashtags operate as speech acts and cultural objects: they acquire meaning through collective use and are continually redefined via interaction. From this perspective, hashtags are not merely employed by users but are also transformed by them, contributing to the construction of social reality (Budnik et al., 2019). This approach has been particularly influential in studies of hashtag activism, where hashtags function as rallying points for collective action and solidarity, enabling mobilisation and identity formation around political and social causes (Dobrin, 2020; Ross, 2020; Sebeelo, 2021). In these contexts, hashtags can become condensed manifestos of movements and potent markers of affiliation (La Rocca and Boccia Artieri, 2022).

This scholarship supports the view of hashtags as cultural symbols carrying semantic content that is continuously renegotiated. This conceptualisation motivates the present study's focus on longitudinal changes in HPV-related hashtag use in Italian posts as a way to understand evolving health nar-

ratives and their entanglement with broader social dynamics.

3. The HPV vaccine

HPV, the most prevalent sexually transmitted infection globally, affects both men and women, with an estimated 80%-90% of individuals contracting it during their lifetime (Kombe Kombe et al., 2021). Of the 100 types of HPV, 13 may induce cancer, with cervical cancer being the most common HPV-related cancer and the fourth most common cancer among women worldwide (WHO, 2022). In terms of public reception, since its approval by the Food and Drugs Administration (FDA) and the European Medical Agency (EMA) in 2006, Gardasil, the HPV vaccine licensed by Merck in the same year, has faced multiple anti-vaccination campaigns around the world. Concerns have included fears of promoting sexual promiscuity, providing false security, mandatory vaccination issues, access disparities, unreported adverse effects, and corruption (Chen et al., 2021; Briones et al., 2012; Vamos et al., 2008). These concerns have been amplified through media like Andi Reiss's 2018 documentary 'Sacrificial Virgins', which sparked controversy and government pressure in Australia (Gwynne, 2020), works like 'The HPV Vaccine On Trial' (Holland et al., 2018) and retracted articles on fertility issues (DeLong, 2018).

In addition to the already cited Japanese case, disinformation and conspiracy theories, typically revolving around recurring tropes, have also been identified in the discourse of HPV anti-vaccination campaigns (Smith and Gorski, 2024; Khalil et al., 2023; Wakefield, 2022). In China, some conspiracists claim that the HPV vaccine is both a profit-driven scheme by the government and a biological weapon from western nations aimed at eradicating the Chinese ethnic group (Chen et al., 2021). Broader global narratives include accusations that governments and pharmaceutical companies fabricate data on vaccine efficacy and safety (Khalil et al., 2023), allegations that the vaccine causes infertility and primary ovarian insufficiency (POI), depopulation conspiracy theories (Viola, 2025; Smith and Gorski, 2024), and claims that vaccinated individuals shed spike proteins, causing illness in others (Wakefield, 2022).

To complicate the information space even more, some theories have been substantiated, such as the 2017 Italian exposé on corruption during the Gardasil approval process (Borella, 2017; Gabanelli and Valesini, 2016), thus making Italy a particularly valuable use case for the present study. Italy was also the first European country to adopt a population-based HPV vaccination strategy for 12-year-old girls (Mennini et al., 2022; Gabutti et al., 2021). Additional age groups were proposed as secondary targets, including 25-year-old women already involved in HPV screening services and a potential third cohort of women between 12 and 25 years old. The 2017–2019 National Immunisation Plan (NIP) expanded to include both sexes as primary targets for HPV vaccination during adolescence, preferably before sexual debut (Mennini et al., 2022). Currently, the vaccine is offered for free; it is not mandatory but recommended for boys and girls between 11 and 15 years of age, with some regions extending the gratuity up until 26 years of age.

4. Data and methodology

Data retrieval was conducted through targeted queries that extracted posts from X containing specific hashtags, including #HPV, #Gardasil, #papillomavirus, #papilloma.[1] Even though posts were collected with the platform language parameter set to Italian (lang=it), an additional post-hoc language filter was necessary because the platform's language metadata is not fully reliable, mostly due to language tags produced by automated classifiers. Language identification was performed using fastText's pre-trained language identification model (lid.176), and only tweets predicted as Italian (it) were kept for downstream analysis. The resulting corpus spans from 2 January 2010 to 30 December 2024 ($\approx$5,476 days, ≈ 15 years) and contains 4,895 unique post records from 2,252 unique usernames (1,910 unique user IDs). The 2,394 hashtags are present in 904 posts (18.47%), comprising 754 unique hashtags. The dataset also includes several attributes such as the post texts, likes, replies, reposts, shares, and quotes count. A detailed description is given in Table 4 in the Appendix. The dataset was finally pseudonymised and can be provided upon request to the author.

To map long-term changes in Italian HPV vaccine discourse, the study models hashtags as markers of discourse by analysing their co-occurrence structure over time. As already discussed, the methodology was designed to overcome the loss of early information disorder signals inherent in longitudinal social media analysis whereas traditional network approaches are often applied with aggressive filtering to ensure structural interpretability. To recover such long-tail signals and identify where information disorder typically originates, we implemented a dual-layer hashtag co-occurrence framework that maintains the rigour of conservative community

[1] Data collection was performed in 2025 on the Apify platform by executing a Twitter/X scraping Actor (Actor ID: nfp1fpt5gUlBwPcor) via Apify Actor runs (Apify API v2 / Apify Console), and exporting the resulting dataset from the Actor run output.

Epoch	Posts	Hashtags	Bigrams
2010–2014	1,277	640	812
2015–2019	1,514	1,095	1,932
2020–2024	2,104	659	1,127

Table 1: Post, hashtag and hashtag-bigram counts per epoch.

detection while capturing the emerging narrative reframing found in lower-frequency discourse.

Based on their timestamp, posts were first segmented into three pivotal epochs: 2010–2014 (initial policy implementation), 2015–2019 (policy Expansion), and 2020–2024 (COVID and Post-COVID). Following the research design (i.e., hashtags as cultural products that signal conceptual association and communal alignment) all hashtags were extracted. For each post and year, we then constructed an undirected weighted co-occurrence table of hashtag pairs (bigrams), where an edge between hashtags h_i and h_j is created whenever they appeared in the same post and the edge weight equalled the number of posts in that year in which the pair co-occurs. Yearly edges were aggregated within each epoch to obtain an epoch-specific weighted hashtag co-occurrence graph. Table 1 illustrates the distribution of posts, hashtags and hashtags bigrams across epochs.

From the distribution of hashtags and hashtags bigrams across epochs, interesting observations can already be drawn. There is a clear surge in both unique hashtags and bigram connections during the 2015–2019 epoch. This aligns with the 'Policy Expansion' phase, suggesting that the discourse became more dense and diverse. Moreover, in all epochs, the number of bigrams significantly exceeds the number of hashtags, indicating a highly interconnected network where hashtags are rarely used in isolation.

To identify the stable thematic backbone of the discourse, we first construct a core layer. In this layer, we apply conservative network constraints to filter out noise and ensure the robustness of the identified coalitions, specifically, only edges with a weight ≥ 2 (co-occurrence in at least two unique posts) are retained. To ensure that every hashtag is part of a conversation circle rather than just a lonely tag attached to a single main topic, we also apply k-core constraint ≥ 2, pruning all nodes that are not part of a subgraph where every node is connected to at least two others.

Community structure within each epoch graph was identified using a hierarchical Girvan-Newman community detection procedure (Girvan and Newman, 2002); this algorithm reflects the number of shortest paths that pass through an edge. By progressively removing edges with the highest betweenness, it effectively disassembles the network along its most significant connections, revealing peripheral community structures within the graph. This method is especially suited to smaller or moderately-sized networks where the clarity of the hierarchical structure is paramount. Using this algorithm, we detected the core coalitions, that is the most stable discourse. To enable diachronic comparison, communities were matched across epochs using Jaccard overlap of their core hashtag sets, producing stable community identifiers.

We then computed the coverage layer. This layer is meant to capture the information disorder signals, which are often sparse and initially disconnected from the mainstream, as already discussed. This layer retains every hashtag co-occurrence (edge weight ≥ 1) and removes the k-core constraint. This inclusive approach ensures that low-frequency hashtags such as emerging skepticism markers are captured rather than discarded. Finally, we implement a core–coverage projection. Rather than treating the two layers as separate entities, we project the coverage hashtags into the existing core coalitions based on their weighted relative connectivity. Specifically, a hashtag from the coverage layer was assigned to a core coalition if it shared an edge with at least one node in that core coalition and its strongest weighted connection (or the sum of its connections) gravitates towards that specific cluster. Hashtags that shared no edges with any core nodes remained unassigned. This method allowed us to observe how fringe, long-tail narratives attempt to infiltrate or reframe established core narratives, for example showing how a conspiracy hashtag might attach itself to a core prevention node.

Finally, to validate the effectiveness of this projection, we manually labelled a set of seed posts containing verified skeptical or conspiratorial content. Seed posts were selected via Critical Discourse Analysis (CDA)-guided manual identification of canonical vaccine-controversy frames as found in the literature (e.g., safety harms, corruption/pharma capture, freedom/control, depopulation rhetoric, institutional distrust). The author then manually flagged posts in the corpus that clearly instantiated these frames and used the hashtags appearing in these posts to build a small seed lexicon for weak supervision. By measuring the recovery rate of these seeds within the projected layer versus a standard single-layer k-core graph, we demonstrate that the Core–Coverage framework captures significantly more salient information disorder signals while preserving an interpretable structural map of the discourse evolution.

5. Analysis and Results

First, we report the size of the *largest connected component* (LCC), i.e., the number of nodes in the

Epoch	Layer	N	E	C	LCC	W-in
2010–2014	Core	46	99	3	40	0.90
2010–2014	Coverage	150	403	5	137	0.76
2015–2019	Core	86	224	7	69	0.95
2015–2019	Coverage	349	1,169	16	312	0.79
2020–2024	Core	40	84	7	24	1.00
2020–2024	Coverage	268	858	17	220	0.84

Table 2: Core vs. coverage diagnostics across epochs. N = nodes; E = edges; C = connected components; LCC = size of the largest connected component; W-in = within-community edge share.

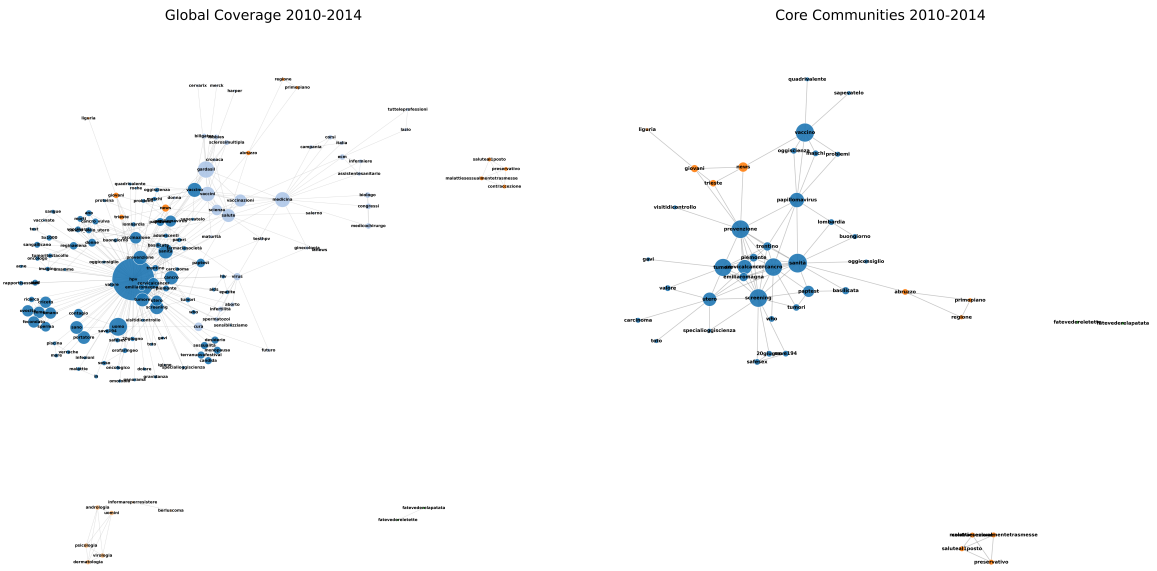
largest subgraph in which all nodes are mutually reachable via paths, together with the total number of disconnected components in the network (implicitly). Coverage networks are larger and typically include more peripheral components, while the core networks concentrate mass into fewer, denser structures. Second, we report the *within-community edge share*, defined as the fraction of edges whose endpoints fall within the same detected community. As expected, this share is substantially higher in the core networks (2010–2014: 0.90 vs. 0.76; 2015–2019: 0.95 vs. 0.79), indicating that sparsification produces a clearer modular backbone, whereas the coverage layer preserves more cross-community bridges and niche connections that increase between-community linking. The results are displayed in Table 2.

Figures 1 and 3 display the two layers as network graphs. The graphs show how the discourse expanded dramatically particularly between the first two epochs. The coverage layer, which represents the full conversational landscape, more than doubled in size in the epoch 2015-2019, growing from 150 nodes to 349 nodes, whereas the core layer grew from 46 to 86 nodes, indicating that more hashtags achieved the structural robustness required to be considered part of the core discourse. At the same time, while the network expanded, its density decreased, reflecting a more fragmented and specialized discourse where different communities become increasingly distinct. In coverage, the discourse is organized around the HPV anchor, as evidenced by the very high connectivity of the #hpv node compared to other tags. Around it, stable health frames can be observed such as screening/prevention/cancer and medical/vaccine vocabulary (e.g., #screening, #prevenzione, #cancro, #tumore, #paptest, #vaccini, #medicina, and #gardasil). This suggests that early discourse is largely public-health informational and HPV is framed as prevention/cancer risk with a supporting vaccine thread. In contrast, the core is much smaller and notably does not contain #hpv (nor edges involving #hpv). This confirms that the core network foregrounds a tight prevention/screening/cancer backbone showing that the clinical/preventive narrative family is the most robust structure.

The second epoch (2015–2019) reveals the rise of narrative coalitions. While the prevention backbone persists (#prevenzione, #salute), a highly robust skepticism coalition has moved into the core. This community includes terms like #vaccinegate, #dna, #genoma, and #rna. The presence of these terms in the core indicates they are no longer fringe noise but have formed a structurally dense and persistent narrative coalition. The coverage layer captures the external political and social pressure. We see #lorenzin (referring to the Italian mandatory vaccination law), #novax, and pharmaceutical actors like #merck. These terms are highly active but remain in the coverage layer, suggesting they are widely distributed across conversations rather than being anchored in a single tight thematic core.

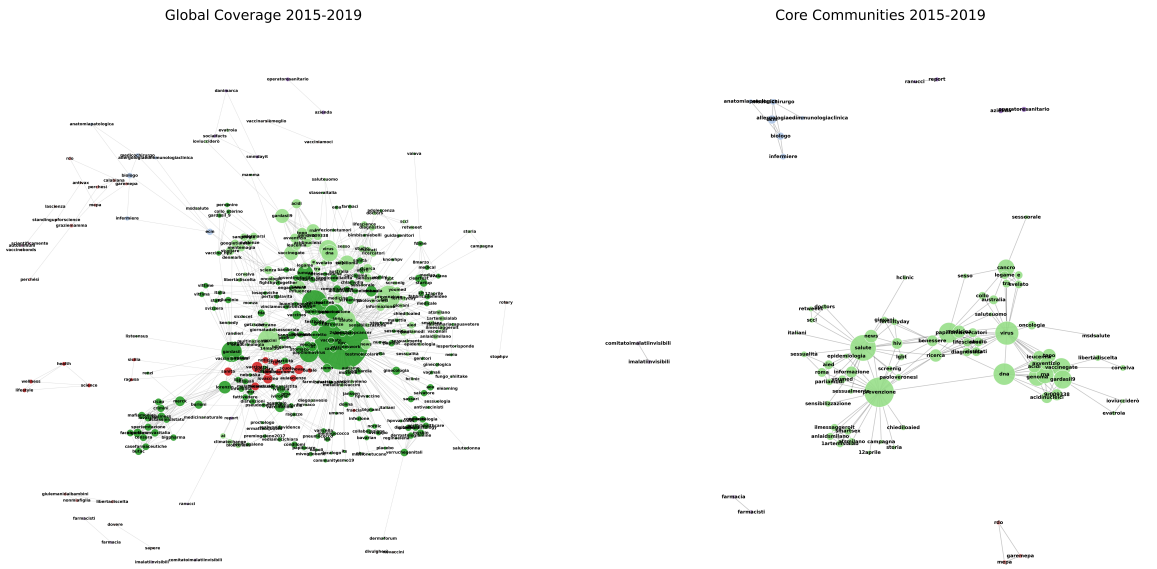
The analysis of the final epoch (2020–2024) reveals a profound structural and thematic shift that characterizes the maturation stage of information disorder in the Italian HPV discourse. In this period, the network dynamics reflect a discourse deeply influenced by the post-COVID-19 environment and a consolidation of skeptical narratives. While the 2015–2019 epoch marked a peak in discursive expansion, 2020–2024 shows a consolidation. The coverage layer contains 268 nodes, slightly lower than the previous peak (349) but maintaining a high level of connectivity. The most striking result of this epoch is the content of the core layer. Unlike earlier periods where the core was dominated by institutional prevention, the 2020–2024 core has become the site of a highly robust, internationalized skepticism coalition. This is especially visible in the core community that features high-frequency hashtags such as #bigpharma, #eugenetica (eugenics), #ogm, and #who/#oms (World Health Organization). The presence of #covid_19 within this same core cluster demonstrates that HPV discourse is no longer an isolated medical topic: it has been absorbed into a broader framework including other health topics. Moreover, the appearance of #thedefender (associated with the Children's Health Defense) and #hpvvaccine alongside global pharmaceutical names like #merck and #glaxo indicates that the Italian discourse is now tightly synchronized with international anti-vaccination narratives.

Another very significant result is the reversal of the 2010–2014 epoch, whereas many institutional and prevention-oriented terms have moved to the coverage layer in this later epoch. While hashtags like #hpv, #prevenzione, #salute, and #screening remain the most frequent in terms of total mentions (weighted degree), they are now part of the long tail in relation to the highly connected core of skepticism. The coverage layer also captures a broader range of general health services (#mam-



(a) Coverage network (2010–2014). (b) Core network (2010–2014).

Figure 1: Hashtag co-occurrence networks for 2010–2014 (coverage vs. core).



(a) Coverage network (2015–2019). (b) Core network (2015–2019).

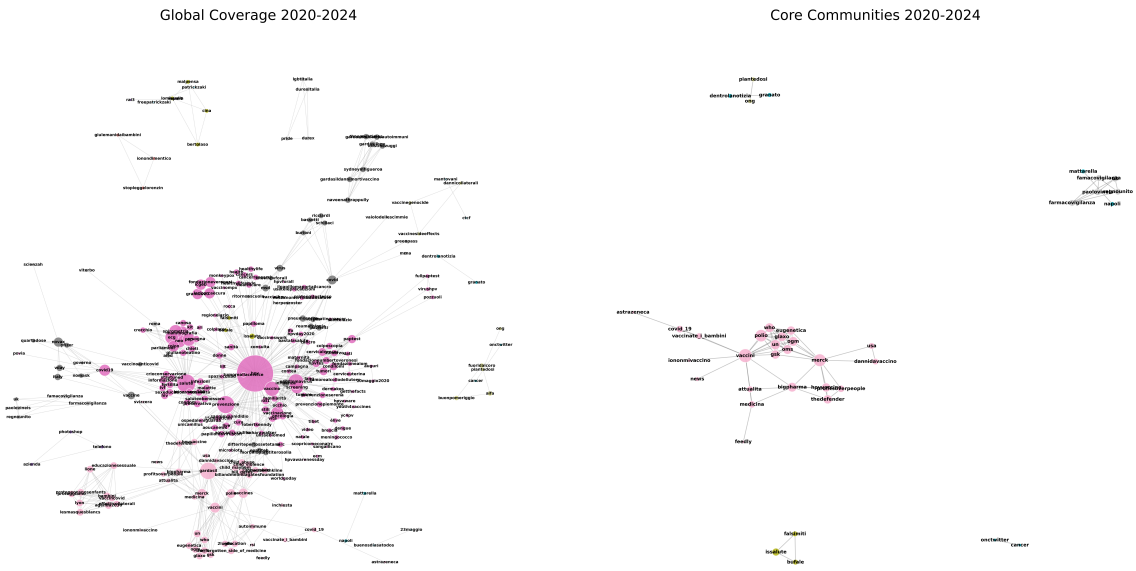
Figure 2: Hashtag co-occurrence networks for 2015–2019 (coverage vs. core).

mografia, #spirometria, #ecg, #paptest) suggesting that while the general public still discusses HPV in the context of screening, these conversations are structurally more fragmented compared to the tight, self-reinforcing core of reframed discourse. In other words, HPV-related discussion persists, but it co-exists with (and is partially recontextualised by) broader vaccine controversy, including anti-establishment and COVID-adjacent frames.

From an information-disorder perspective, this doesn't merely show that new narratives appear, but that they alter the structural coupling between themes, whether by creating bridges that connect previously separate frames, or by forming parallel clusters that remain partly segregated yet reshape the overall discourse environment.

6. Validation

To validate the methodology, we connected hashtag communities to higher-level information-disorder narratives by constructing a small seed



(a) Coverage network (2020–2024).

(b) Core network (2020–2024).

Figure 3: Hashtag co-occurrence networks for 2020–2024 (coverage vs. core).

lexicon of narrative-indicative hashtags for two frames: *skeptical/anti-vaccine* and *conspiracy*. The skeptical/anti-vaccine frame included posts expressing doubt or opposition to vaccination, such as safety/efficacy concerns and institutional distrust framed as skepticism while the conspiracy frame was used to identify posts alleging coordinated deception or malicious intent, e.g., Big Pharma plots, depopulation/sterilization claims, censorship narratives. The author manually annotated a small subset of randomly chosen posts (≈ 10 per year); posts were flagged conservatively only when the narrative signal was unambiguous. From this subset, we identified 86 seed posts (69 conspiratorial and 17 skeptical). These posts yielded a seed-hashtag lexicon comprising 52 conspiracy-associated hashtags, 5 skeptical hashtags, and 3 overlapping hashtags. This lexicon was then used as weak supervision to calculate community-level enrichment scores, indicating whether seed-linked content was overrepresented in a given community relative to its size. The list of seed hashtags is in Table 5 in the Appendix whereas Table 3 shows two examples of posts for both categories.

For each epoch and community, the seed enrichment score was computed as the ratio between (i) the community's share of all seed mentions and (ii) the community's baseline share of total hashtag mentions. Communities with enrichment > 1 are over-represented in the corresponding narrative seeds relative to their size, proving that the community contains a higher concentration of skeptical or conspiratorial content than the rest of the network. Figure 4 shows how in each epoch, only a minority

Label	Italian post	English translation
Skeptical	VIDEO HPV: I GIOVANI CHIEDONO PIU' INFORMAZIONE SUL VACCINO	HPV VIDEO: Young people ask for more information about the vaccine
Skeptical	Tumori: vaccino hpv non convince mamme	Cancer: the HPV vaccine does not convince mothers
Conspiracy	Yes, è proprio un giornalista ad aver iniziato, con una serie di articoli scandalosi... per quanto riguarda la comunità scientifica, no comment	Yes, it was a journalist who started it, with a series of scandalous articles... as for the scientific community, no comment
Conspiracy	non sono i giornali ad aver affossato certe teorie ma la stessa comunità scientifica internazionale.. il caso wakefield è addirittura peggio	It wasn't the newspapers that sank certain theories, but the international scientific community itself... the Wakefield case is even worse

Table 3: Examples of posts flagged by the skeptical and conspiracy seed indicators, with English translations.

of hashtags belong to the core backbone used for community detection, while a larger set of hashtags can be projected onto the core community structure based on their connectivity to core nodes. A third group remains unassigned, indicating peripheral or disconnected usage that cannot be reliably linked

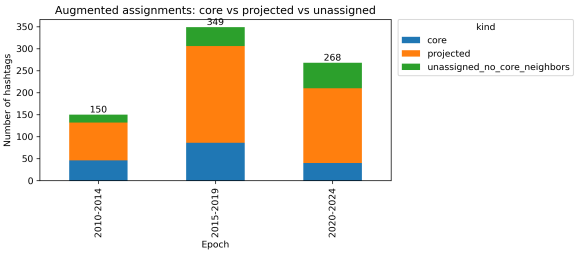


Figure 4: Augmented assignments by epoch.

to a community. Figure 6 in the Appendix reports the distribution of the projection support score (the summed weight of edges linking a projected hashtag to its assigned core community).

As an example of the core vs coverage enrichment overlay, Figure 5 displays the enriched graph for the third epoch (2020–2024). Nodes are coloured by community-level seed enrichment (yellow = baseline, red = high enrichment; log-scaled). The core network (left) shows the sparsified backbone used for community detection; the coverage network (right; largest connected component) retains peripheral hashtags and bridging edges. High enrichment appears in a small number of network communities rather than across the network as a whole. This indicates that information-disorder markers cluster into coherent sub-discourses with consistent co-occurring hashtags, instead of being randomly or uniformly scattered throughout the HPV/vaccine conversation. Importantly, because enrichment is defined relative to community size, concentrated high values reflect disproportionate over-representation of seed markers within particular sub-narratives.

7. Discussion

The 14-year longitudinal analysis (2010–2024) provides empirical evidence for the lifecycle of HPV vaccine discourse in Italian X through narrative coalitions. In the three epochs, clear core clusters shifts can be identified: institutional discourse forms the core in 2010-2014 and skepticism is sparse and located in the long tail. In the second epoch (2015–2019), narrative coalitions around controversy and skepticism mature into a core cluster that competes with the prevention backbone. Finally, in the last epoch (2020–2024), the *information disorder* coalition solidifies into a dominant core. The HPV discourse is no longer about the vaccine but about reframing health as a political and conspiratorial struggle against global institutions.

Methodologically, the dual-layer approach proved effective for detection of early information disorder signals. The coverage networks ensured that peripheral and bridging material was not excluded, while the core networks continue to identify the most robust associations that were least sensitive to low-frequency noise. The fact that in 2020–2024, the core backbone contains almost no links between different communities suggests that late-period discourse is more segmented into distinct frames.

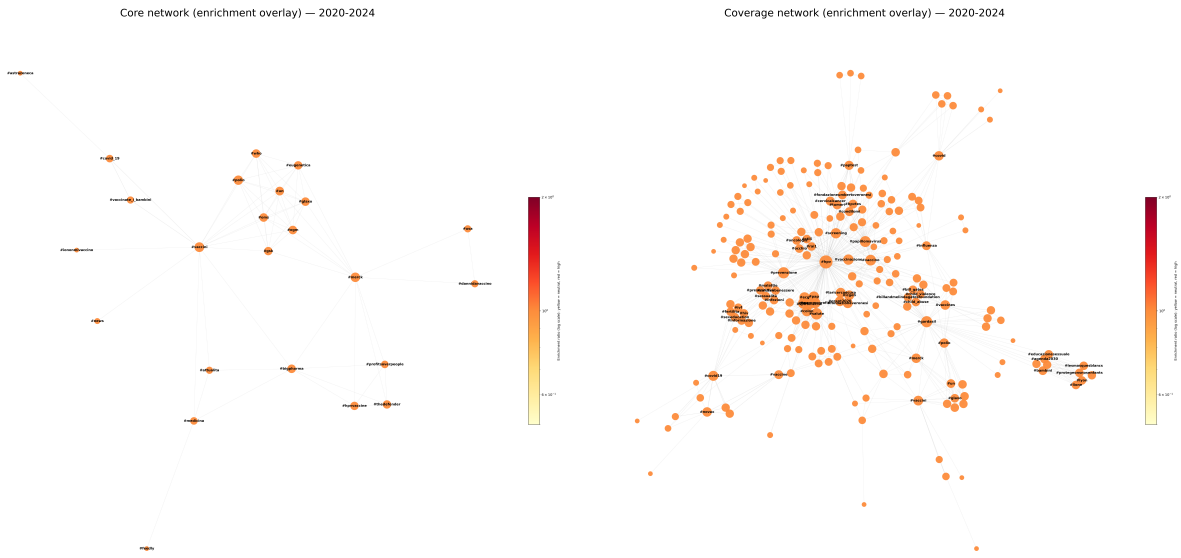
Finally, by using manually labelled seed posts (skeptical/conspiratorial) to validate the projection, the method effectively treats hashtags as cultural products, providing a sociolinguistic lens onto how hashtags can hijack narratives. This provides a clear empirical basis for interpreting the results as a form of conceptual change in which stable health concepts persist but become embedded in a reconfigured and more polarised information environment. The findings demonstrate that information disorder is also matter of structural shifts in connectivity. In turn, this highlights the importance of the methodology introduced here: it captured that the projected influence of institutional discourse still exists but has lost its central structural dominance.

8. Conclusion

This paper examined the temporal evolution of Italian discourse about the HPV vaccine on X through a graph-based analysis of hashtag co-occurrence. By modelling hashtags as cultural objects that mark discourse and tracking their co-occurrence structure across three epochs (2010-2014, 2015-2019, 2020-2024), cross-epoch community matching, and augmented projection, we show that the HPV conversation undergoes a clear process of conceptual reconfiguration.

Beyond the HPV case, the proposed methodology provides a transparent and reusable framework for studying conceptual change in online discourse and for linking network communities to information-disorder narratives via seed enrichment. The innovation of the method lies in its attempt to bridge the gap between Discourse Studies and Network Science, specifically by addressing the loss of signal problem in longitudinal online discourse studies.

The Core–Coverage projection effectively solves the long tail issue which typically results in graphs structurally biased towards mainstream/dominant discourse. By capturing the coverage layer and then projecting those fringe nodes into the core, this framework creates a computational zoom lens that allows to see when sparse, low-frequency hashtags are reframing attempts that attach themselves to established narrative backbones. In this way, the analysis is moved beyond what is being said to how the structure of the argument matures.



(a) Core network (backbone).

(b) Coverage network (LCC).

Figure 5: Core vs coverage enrichment overlay (2020–2024).

9. Limitations

While this study provide several insights, limitations must be acknowledged. First, the dataset is subject to the X's API access constraints and policy requirements and observed temporal differences may partially reflect changes in platform usage and data availability rather than discourse alone. Second, hashtag-based analysis captures a salient but incomplete layer of meaning-making: many tweets contain no hashtags, hashtags are used strategically (for visibility, irony, or audience targeting), and co-occurrence networks do not directly encode stance, intent, or factual accuracy. Therefore, the reliance on hashtags as primary signals may overlook important discourse without hashtags. Third, community structure depends on modelling choices (epoch boundaries, edge thresholds, top-k pruning, k-core filtering, and community detection settings). While we mitigate this by reporting core vs. coverage results and by using augmented projection with explicit support scores, different parameterisations may yield alternative granularities of communities.

The seed-based enrichment approach provides weak supervision for narrative-family mapping but is sensitive to seed selection and ambiguity (e.g., topic anchors that co-occur with narrative markers), therefore the seed-based validation may introduce bias depending on selection criteria. Enrichment serves as a guide for qualitative validation rather than a definitive classifier of conspiracy or skepticism. Additionally, while the method is generalizable in principle, further evaluation on other domains or datasets would strengthen the claims. Future work should extend the analysis with larger corpora and validate narrative-family mapping through systematic annotation and inter-coder agreement, enabling stronger claims about misinformation, disinformation, and malinformation dynamics.

10. References

Virginia Berridge. 2007. *Marketing health: smoking and the discourse of public health in Britain, 1945-2000*. Oxford University Press, Oxford ; New York. OCLC: ocn123797267.

Erika Bonnevie, Allison Gallegos-Jeffrey, Jaclyn Goldbarg, Brian Byrd, and Joseph Smyser. 2021. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *Journal of communication in healthcare*, 14(1):12–19. ISBN: 1753-8068 Publisher: Taylor & Francis.

Alessandra Borella. 2017. Report - Reazioni avverse - 17/04/2017 - Video.

Rowena Briones, Xiaoli Nan, Kelly Madden, and Leah Waks. 2012. When Vaccines Go Viral: An Analysis of HPV Vaccine Coverage on YouTube. *Health Communication*, 27(5):478–485. Publisher: Routledge _eprint: https://doi.org/10.1080/10410236.2011.610258.

Brian C. Britt. 2023. The Evolution of Discourse in Online Communities Devoted to a Pandemic. *Health Communication*, 38(5):1041–1053. Publisher: Routledge _eprint: https://doi.org/10.1080/10410236.2021.1991618.

Ekaterina Budnik, Violetta Gaputina, and Vera Boguslavskaya. 2019. Dynamic of hashtag functions development in new media: Hashtag as an identificational mark of digital communication in social networks. In *Proceedings of the XI International Scientific Conference Communicative Strategies of the Information Society*, pages 1–5, St. Petersburg Russian Federation. ACM.

William A. Calo, Melissa B. Gilkey, Parth D. Shah, Anne-Marie Dyer, Marjorie A. Margolis, Susan Alton Dailey, and Noel T. Brewer. 2021. Misinformation and other elements in HPV vaccine tweets: an experimental comparison. *Journal of Behavioral Medicine*, 44(3):310–319.

Richard M. Carpiano, Timothy Callaghan, Renee DiResta, Noel T. Brewer, Chelsea Clinton, Alison P. Galvani, Rekha Lakshmanan, Wendy E. Parmet, Saad B. Omer, Alison M. Buttenheim, Regina M. Benjamin, Arthur Caplan, Jad A. Elharake, Lisa C. Flowers, Yvonne A. Maldonado, Michelle M. Mello, Douglas J. Opel, Daniel A. Salmon, Jason L. Schwartz, Joshua M. Sharfstein, and Peter J. Hotez. 2023. Confronting the evolution and expansion of anti-vaccine activism in the USA in the COVID-19 era. *The Lancet*, 401(10380):967–970. Publisher: Elsevier.

Li Chen, Yafei Zhang, Rachel Young, Xianwei Wu, and Ge Zhu. 2021. Effects of Vaccine-Related Conspiracy Theories on Chinese Young Adults' Perceptions of the HPV Vaccine: An Experimental Study. *Health Communication*, 36(11):1343–1353.

Giuseppe Crupi, Yelena Mejova, Michele Tizzani, Daniela Paolotti, and André Panisson. 2022. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:102–113.

Gayle DeLong. 2018. RETRACTED ARTICLE:[A lowered probability of pregnancy in females in the USA aged 25–29 who received a human papillomavirus vaccine injection]. *Journal of Toxicology and Environmental Health, Part A*, 81(14):661–674. Publisher: Taylor & Francis.

Diana Dobrin. 2020. The Hashtag in Digital Activism: A Cultural Revolution. *Journal of Cultural Analysis and Social Change*, 5(1):03.

Nihal Durmaz and Engin Hengirmen. 2022. The dramatic increase in anti-vaccine discourses during the COVID-19 pandemic: a social network analysis of Twitter. *Human Vaccines & Immunotherapeutics*, 18(1):2025008. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21645515.2021.2025008.

Milena Gabanelli and Valesini. 2016. Se il dirigente prescritto in un caso di corruzione rappresenta l'Italia in Ue. *Corriere della Sera*.

Giovanni Gabutti, Erica d'Anchera, Francesco De Motoli, Marta Savio, and Armando Stefanati. 2021. Human Papilloma Virus Vaccination: Focus on the Italian Situation. *Vaccines*, 9(12):1374.

Michelle Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826. ArXiv:cond-mat/0112110.

Keith Gunaratne, Eric A. Coomes, and Hourmazd Haghbayan. 2019. Temporal trends in anti-vaccine discourse on Twitter. *Vaccine*, 37(35):4867–4871.

Jacqueline Gwynne. 2020. Sacrificial virgins: Is Gardasil even necessary? *News Weekly*, (3028):6. Publisher: National Library of Australia.

Mary Holland, Kim Mack Rosenberg, and Eileen Iorio. 2018. *The HPV Vaccine On Trial: Seeking Justice for a Generation Betrayed*. Skyhorse Publishing. Google-Books-ID: OnV5DQAAQBAJ.

Joe Khalil, Sarah Boutros, Abdo Hassoun, Souheil Hallit, and Habib Barakat. 2023. Human papillomavirus vaccine knowledge and conspiracy beliefs among secondary school students in Lebanon. *BMC Pediatrics*, 23(1):363.

Arnaud John Kombe Kombe, Bofeng Li, Ayesha Zahid, Hylemariam Mihiretie Mengist, Guy-Armel Bounda, Ying Zhou, and Tengchuan Jin. 2021. Epidemiology and Burden of Human Papillomavirus and Related Diseases, Molecular Pathogenesis, and Vaccine Evaluation. *Frontiers in Public Health*, 8:552028.

Svetlana Kordumova, Jan Gemert, and Cees G. Snoek. 2016. Exploring the Long Tail of Social Media Tags. In *Proceedings, Part I, of the 22nd International Conference on MultiMedia Modeling - Volume 9516*, MMM 2016, pages 51–62, Berlin, Heidelberg. Springer-Verlag.

Melanie L. Kornides, Sarah Badlis, Katharine J. Head, Mary Putt, Joseph Cappella, and Graciela Gonzalez-Hernadez. 2023. Exploring content of misinformation about HPV vaccine on twitter. *Journal of Behavioral Medicine*, 46(1-2):239–252.

Gevisa La Rocca and Giovanni Boccia Artieri. 2022. Research using hashtags: A meta-synthesis. *Frontiers in Sociology*, 7:1081603.

Heidi J. Larson and David A. Broniatowski. 2021. Volatility of vaccine confidence. *Science*, 371(6536):1289.

Regina G. Lawrence. 2004. Framing Obesity: The Evolution of News Discourse on a Public Health Issue. *Harvard International Journal of Press/Politics*, 9(3):56–75.

Deborah McPhail and Andrea E. Bombak. 2015. Fat, queer and sick? A critical analysis of 'lesbian obesity' in public health discourse. *Critical Public Health*, 25(5):539–553. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09581596.2014.992391.

Francesco Saverio Mennini, Andrea Silenzi, Andrea Marcellusi, Michele Conversano, Andrea Siddu, and Giovanni Rezza. 2022. HPV Vaccination during the COVID-19 Pandemic in Italy: Opportunity Loss or Incremental Cost. *Vaccines*, 10(7):1133.

Bjarke Mønsted and Sune Lehmann. 2022. Characterizing polarization in online vaccine discourse—A large-scale study. *PLOS ONE*, 17(2):e0263746. Publisher: Public Library of Science.

Ignacio Ojea Quintana, Ritsaart Reimann, Marc Cheong, Mark Alfano, and Colin Klein. 2022. Polarization and trust in the evolution of vaccine discourse on Twitter during COVID-19. *PLOS ONE*, 17(12):e0277292. Publisher: Public Library of Science.

Ann Robertson. 1998. Shifting Discourses on Health in Canada: From Health Promotion to Population Health. *Health Promotion International*, 13(2):155–166.

Andrew S. Ross. 2020. Discursive delegitimisation in metaphorical *#secondcivilwarletters* : an analysis of a collective Twitter hashtag response. *Critical Discourse Studies*, 17(5):510–526.

Tebogo B. Sebeelo. 2021. Hashtag Activism, Politics and Resistance in Africa: Examining #ThisFlag and #RhodesMustFall online movements. *Insight on Africa*, 13(1):95–109.

Tara C. Smith and David H. Gorski. 2024. Infertility: A common target of antivaccine misinformation campaigns. *Vaccine*, 42(4):924–929.

Bonnie Stabile, Aubrey Grant, Hemant Purohit, and Kelsey Harris. 2019. Sex, Lies, and Stereotypes: Gendered Implications of Fake News for Women in Politics. *Public Integrity*, 21(5):491–502. Publisher: Routledge _eprint: https://doi.org/10.1080/10999922.2019.1626695.

Cheryl A. Vamos, Robert J. McDermott, and Ellen M. Daley. 2008. The HPV vaccine: framing the arguments FOR and AGAINST mandatory vaccination of all middle school girls. *Journal of School Health*, 78(6):302–309. ISBN: 0022-4391 Publisher: Wiley Online Library.

Jan-Willem Van Prooijen and Mark Van Vugt. 2018. Conspiracy Theories: Evolved Functions and Psychological Mechanisms. *Perspectives on Psychological Science*, 13(6):770–788.

Lorella Viola. 2025. 'Barren lesbians plotting sterilization': gender stereotypes and prejudices in health disinformation narratives, a cross-cultural analysis of social media of the HPV vaccine. In Catherine Tebaldi, Alistair Plum, and Christoph Purschke, editors, *Conspiracy as Genre: Narrative, Power and Circulation*. Bloomsbury Academic, London.

Andrew Wakefield. 2022. Infertility: A Diabolical Agenda Is Anti-Vaxx Sleight-of-Hand Propaganda.

Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, editors. 2013. *Twitter and society*. Number vol. 89 in Digital formations. Peter Lang, New York.

WHO. 2022. Cancer Today. Technical report, WHO.

Bradley Wiggins. 2023. 'Nothing Can Stop What's Coming': An analysis of the conspiracy theory discourse on 4chan's /Pol board. *Discourse & Society*, 34(3):381–398. Publisher: SAGE Publications Ltd.

Asami Yagi, Yutaka Ueda, and Tadashi Kimura. 2024. HPV Vaccine Issues in Japan: A review of our attempts to promote the HPV vaccine and to provide effective evaluation of the problem through social-medical and behavioral-economic perspectives. *Vaccine*, 42(22):125859.

Michele Zappavigna. 2016. Twitter. In Christian R. Hoffmann and Wolfram Bublitz, editors, *Pragmatics of social media*, volume 11, pages 201–224. Walter de Gruyter. Publisher: Walter de Gruyter GmbH & Co KG.

A. Appendix

Statistic	Value
Time span (UTC)	2010-01-02 to 2024-12-30
Duration	5,476 days (≈15.0 years)
Tweets (rows)	4,895
Unique users (user IDs)	1,910
Unique usernames	2,252
Tweets with ≥1 hashtag	904 (18.47%)
Unique hashtags (observed)	754
Likes (sum / median / max)	191,523 / 0 / 28,662
Replies (sum / median / max)	25,864 / 0 / 1,314
Retweets (sum / median / max)	36,562 / 0 / 2,261
Quotes (sum / median / max)	5,215 / 0 / 425

Table 4: Descriptive statistics for the Italian-language Twitter/X corpus. Hashtag presence is computed across all hashtag fields (including quoted/reply tweet objects when available).

Seed set	Unique seed hashtags
Skeptical	#feedly, #gardasil, #hpv, #medicina, #papilloma
Conspiracy	#algoritmo, #antivax, #assassinidistato, #aulascienze, #bigpharma, #burioni, #business, #butac, #casefarmaceutiche, #censura, #checoincidenza, #cicap, #crimini, #facebook, #gardasil, #giulemanidaibambini, #giulemanidaigiovani, #hpv, #ionondimentico, #iovaccini, #iovaccino, #laverità, #libertadiscelta, #lorenzin, #mafiamedica, #malattiesessuali, #medicina, #merck, #nonmiafiglia, #notizie, #novax, #papillomavirus, #portatore, #pseudomedicina, #salute, #sanita, #sano, #scienza, #scuola, #senzacategoria, #speriemtazione, #standingupforscience, #stopleggelorenzin, #strinic, #teamvaxitalia, #ultimora, #uominiedonne, #uomo, #vaccinarsi, #vaccinazione, #vaccinazioni, #vaccini
Overlap	#hpv, #gardasil, #medicina

Table 5: Seed hashtag lexicon used to flag skeptical and conspiracy narratives (hashtags normalised to lowercase).

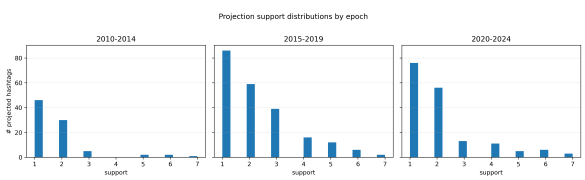


Figure 6: Distribution of projection support by epoch (sum of edge weights from a projected hashtag to its assigned core community).

Author Index