IMT School for Advanced Studies, Lucca
Lucca, Italy

A multidisciplinary approach to combat disinformation in online discourse

PhD Program in Systems Science

Track in Computer Science and System Engineering

XXXVI Cycle

By

Manuel Pratelli

2026

The dissertation of Manuel Pratelli is approved.

PhD Program Coordinator: coordinator, IMT School for Advanced Studies Lucca

Advisor: Prof. Rocco De Nicola, IMT School for Advanced Studies Lucca

Co-Advisor: Dr. Marinella Petrocchi, CNR IIT, Pisa, Italy

Co-Advisor: Prof. Fabio Pinelli, IMT School for Advanced Studies Lucca

The dissertation of Manuel Pratelli has been reviewed by:

Prof. Roberto Di Pietro, KAUST, Saudi Arabia

Prof. Carolina Scarton, University of Sheffield, UK

IMT School for Advanced Studies Lucca
2026

Contents

List of Figures

List of Tables

Acknowledgements

Questo lavoro rappresenta un punto di arrivo di un percorso lungo e intenso, che non sarebbe stato possibile senza il supporto, diretto e indiretto, di molte persone.

Il primo e più sentito ringraziamento va alla mia famiglia, che mi ha sostenuto in ogni fase di questi anni. A mia moglie Ilaria, che ha creduto nella mia scelta di intraprendere un dottorato anche quando è arrivata in un momento della vita non proprio "convenzionale", e che mi è stata accanto con comprensione e incoraggiamento costante. Ai miei genitori, mia mamma e mio babbo, che in tutti gli anni di studio non hanno mai smesso di credere in me e di darmi fiducia. A mio fratello Jonathan, mio zio Patrizio e ai miei nonni, Ivo, Tina, Graziella e Rosa, che con la loro presenza, i loro gesti semplici e la loro cura hanno regalato serenità e senso di unione a me e alla mia famiglia. E soprattutto ai miei figli, Miriam, Elia e Mattia, che sono la luce del mio cuore e la motivazione più grande in tutto ciò che faccio.

I risultati raggiunti in questo dottorato non sarebbero stati possibili senza tutte le persone incontrate lungo il percorso, che hanno contribuito, ciascuna a modo proprio, alla mia crescita personale e scientifica. Un ringraziamento va a tutti i colleghi e amici del gruppo SysMa e della Scuola IMT, con cui ho condiviso questi anni: tra questi i miei advisors Marinella, Rocco e Fabio, ai quali sono profondamente grato per la guida ed il supporto. Un ringraziamento speciale va a Marinella: sono convinto che senza di lei il mio percorso non sarebbe mai stato lo stesso. La sua fiducia costante (per niente banale da trovare) mi ha permesso di crescere e di affrontare le sfide che via via si presentavano.

Desidero inoltre ringraziare Fabio Saracco, a cui devo molto della prima fase del mio percorso di ricerca, ed a tutti i colleghi IMT che hanno reso questo percorso più ricco e stimolante: Lillo, Mirco, Emilio, Gabriele, John, Silvia, Margherita, Roberto, Yuri, Stefano, Simone, Daniele, e tutti gli altri che, anche se non nominati singolarmente, hanno lasciato un segno.

Un sincero ringraziamento anche ai colleghi conosciuti più recentemente al CNR: Stefano, Amaury, Matteo, Clara, Tiziano, Paolo, Marco, Ilaria, Michela e Giuseppe, con cui ho avuto il piacere di collaborare e confrontarmi.

Guardando indietro, sono molto felice di aver intrapreso questo percorso. È stata una scelta guidata dalla passione per la ricerca e dalla fiducia nel suo valore, con la speranza di poter contribuire, anche solo in parte, a qualcosa di utile per la società e per le persone. È questo lo spirito che spero di riuscire a trasmettere anche alle persone che amo, a partire dalla mia famiglia nelle scelte che faranno.

Vita

February 3, 1988	Born, Lucca, Italy
July 5, 2012	B.Sc. in Computer Engineering Final Mark: 104/110 University of Pisa, Italy
November 27, 2015	M.Sc. in Computer Engineering Final Mark: 110/110 University of Pisa, Italy
November 2015 – October 2019	Software Engineer Backend Java developer MetodoIn s.r.l, Pisa
November 2019 – October 2020	Research Collaborator Project: Tools for Fighting FakEs (TOFFEe) IMT School for Advanced Studies Lucca
November 2020 – Present	Ph.D. in CSSE IMT School for Advanced Studies Lucca
November 2024 – Present	Researcher (Level III) IIT-CNR, Pisa, Italy

Publications

1. G. Caldarelli, R. De Nicola, M. Petrocchi, M. Pratelli, and F. Saracco, "Flow of online misinformation during the peak of the COVID-19 pandemic in Italy," *EPJ Data Science*, vol. 10, no. 1, 34, 2021.
2. R. De Nicola, M. Petrocchi, and M. Pratelli, "On the efficacy of old features for the detection of new bots," *Information Processing & Management*, vol. 58, no. 6, 102685, 2021.
3. M. Pratelli and M. Petrocchi, "A Structured Analysis of Journalistic Evaluations for News Source Reliability," in *Proc. of the 16th International AAAI Conference on Web and Social Media (ICWSM) Workshops*, 2022.
4. M. Mattei, M. Pratelli, G. Caldarelli, M. Petrocchi, and F. Saracco, "Bow-tie structures of Twitter discursive communities," *Scientific Reports*, vol. 12, no. 1, 12944, 2022.
5. M. Pratelli, M. Petrocchi, F. Saracco, and R. De Nicola, "Swinging in the States: Does disinformation on Twitter mirror the US presidential election system?," in *Companion Proc. of the ACM Web Conference*, pp. 1395–1403, 2023.
6. M. Pratelli, M. Petrocchi, F. Saracco, and R. De Nicola, "Online disinformation in the 2020 U.S. election: swing vs. safe states," *EPJ Data Science*, vol. 13, no. 1, 25, 2024.
7. M. Pratelli, F. Saracco, and M. Petrocchi, "Entropy-based detection of Twitter echo chambers," *PNAS Nexus*, vol. 3, no. 5, pgae177, 2024.
8. J. Bianchi, M. Pratelli, M. Petrocchi, and F. Pinelli, "Evaluating Trustworthiness of Online News Publishers via Article Classification," in *Proc. of the 39th ACM/SIGAPP Symposium on Applied Computing*, pp. 671–678, 2024.
9. M. Pratelli, F. Saracco, and M. Petrocchi, "Unveiling News Publishers Trustworthiness Through Social Interactions," in *Proc. of the 16th ACM Web Science Conference*, pp. 139–148, 2024.
10. M. Pratelli, F. Saracco, and M. Petrocchi, "TROPIC - Trustworthiness Rating of Online Publishers through online Interactions Calculation," in *Proc. of the European Conference on Information Retrieval*, pp. 407–412, 2025.
11. M. Pratelli, J. Bianchi, F. Pinelli, and M. Petrocchi, "Evaluation of Reliability Criteria for News Publishers with Large Language Models," in *Proc. of the 17th ACM Web Science Conference*, pp. 179–188, 2025.

12. M. Pratelli and M. Petrocchi, " Evaluating the Simulation of Human Personality-Driven Susceptibility to Misinformation with LLMs," in *Proc. of the 28th European Conference on Artificial Intelligence*, 2025.

Presentations

1. M. Pratelli, "A Structured Analysis of Journalistic Evaluations for News Source Reliability," at *News Media and Computational Journalism (MEDIATE) Workshop, International AAAI Conference on Web and Social Media (ICWSM)*, Virtual, 06/06/2022.
2. M. Pratelli, "Swinging in the States: Does disinformation on Twitter mirror the US presidential election system?," at *Cyber Social Threats (CySoc) 2023 Workshop, ACM Web Conference Companion*, Austin, Texas, USA, 30/04–04/05/2023.
3. M. Pratelli, "Unveiling News Publishers Trustworthiness Through Social Interactions," at *16th ACM Web Science Conference (WebSci '24)*, Stuttgart, Germany, 23/05/2024.
4. M. Pratelli, "TROPIC – Trustworthiness Rating of Online Publishers through online Interactions Calculation," at *28th European Conference on Information Retrieval (ECIR '25)*, Lucca, Italy, 07/04/2025.
5. M. Pratelli, "Evaluation of Reliability Criteria for News Publishers with Large Language Models," at *17th ACM Web Science Conference (WebSci '25)*, New Brunswick, New Jersey, USA, 22/05/2025.
6. M. Pratelli, "Evaluating the Simulation of Human Personality-Driven Susceptibility to Misinformation with LLMs," at *28th European Conference on Artificial Intelligence (ECAI '25)*, Bologna, Italy, 27/10/2025.

Abstract

To address the challenge of combating online disinformation, this thesis proposes an interdisciplinary approach combining network science and computer science methodologies.

The contributions of this thesis are the results of parallel studies. One research focus was understanding the dynamics behind the online spread of low-quality or unreliable content during major socio-political events. By analyzing social data, we found that factors such as polarized political communities, echo chambers, and offline elements such as specific voting systems are key indicators of significant flows of unreliable information.

Building on these insights, we developed strategies to counteract the spread of unreliable content. In this context, we developed (i) a novel methodology to detect echo chambers based on social interactions by modelling the concept of echo chambers with network science models and (ii) an automated system to estimate the trustworthiness of news publishers considering in turn (a) automating experts-designed reliability criteria using large language models (LLMs), (b) only users interactions within social media discussion and last, (c) an analysis of articles' textual content to ascertain publisher trustworthiness.

This thesis presents a comprehensive framework that utilizes advanced techniques for analysing online social media discourse to combat online disinformation. A prototype software tool, currently under development, aims to implement this framework by providing support to various facets of social media analysis, including early assessments of news publishers' trustworthiness, evaluation of user behaviors related

to the propensity to disseminate untrustworthy content, identification of participants in shared discourse or echo chambers, and augmentation of human expert efforts to classify the trustworthiness of previously unclassified news publishers.

Chapter 1

Introduction

The creation and spread of disinformation and misinformation represent one of the most urgent and complex challenges facing contemporary society. Disinformation refers to the deliberate dissemination of false or manipulated content, often through coordinated campaigns involving bots and troll accounts. Misinformation, by contrast, concerns the sharing of false or misleading information without malicious intent ([117]).

The World Economic Forum has recently identified mis- and disinformation as among the most critical global risks in the short term[1]. The rapid and large-scale circulation of false but persuasive content can shape public opinion, polarize electorates, fuel social tensions, and exacerbate ongoing crises—from international conflicts to public health emergencies—ultimately undermining collective security and eroding trust in democratic institutions [117, 203].

Research in the field of fake news science shows that misinformation and disinformation often spread within tightly-knit communities, amplified by polarization dynamics and echo chambers [15, 60], and are further propagated by automated accounts such as social bots [75, 201]. The emergence of large language models (LLMs) has further intensified this problem by enabling the massive generation of personalized, contextually plausible content that is increasingly difficult to distinguish from

[1] https://www.weforum.org/publications/global-risks-report-2025/

authentic information—especially during critical events such as presidential elections or international crises [73].

Yet, the very technologies that give rise to new risks also offer the potential for innovative solutions. As a researcher, I believe it is essential for the scientific community to proactively address this challenge by developing tools and strategies aimed at mitigating the impact of disinformation and strengthening citizens' informational resilience over the long term. This thesis proposes a multidisciplinary approach to combating online disinformation.

In this dissertation, the terms *trustworthiness* and *reliability* are used interchangeably to denote the extent to which an informational content or a news outlet deserves to be believed. As detailed in Section 2.3.3, this measure results from a structured evaluation procedure based on multiple quality indicators, referred to as *criteria*. Although the interchangeable use of these terms may appear natural, we note that in several domains—for example, the social sciences—they are often distinguished: *trustworthiness* typically refers to a subjective assessment, whereas *reliability* is understood as an objective property. However, we found no consensus on this distinction in the literature.

1.1 Contributions of this thesis

In this thesis, we propose a novel approach that integrates network science with computer science methodologies to address the challenge of disinformation.

In the initial phase of our research, we embark on an exploratory journey, leveraging data harvested from social media amidst complex socioeconomic events. Our investigation posits the influence of offline user attributes (such as ideological beliefs or political affiliations) and systemic factors (including specific electoral frameworks) on social media behavior, manifesting in unique networking patterns. Employing the principles of network science alongside rigorous quantitative methodologies, we delve into the identification and characterization of these distinctive network configurations. This analysis aims to unravel the

underlying conditions conducive to the spread of disinformation in the digital sphere. By shedding light on these mechanisms, our work aims to contribute to the understanding of how disinformation takes root and proliferates online, offering insights that could inform strategies to fight their diffusion.

The first contribution of this thesis delves into the dynamics of online disinformation across diverse settings, shedding light on the complex interplay between the spread of content produced by low-credibility sources and several key factors and specifically: (i) users' political affiliations, as explored in [33, 127, 159, 160]); (ii) the characterization of accounts, particularly focusing on automated entities, discussed in [159, 160]; and (iii) the influence of the electoral system, examined in [159, 160].

Our analysis of the online debate surrounding the topic of the current pandemic, as reported in [33], reveals that, in Italy, discussions on this predominantly scientific topic are deeply divided along political lines. A notable finding is the disproportionate activity of a center-right group in spreading low-quality content, with 96% of all un-trustworthy news shared by political groups originating from this cohort.

Further, in [159, 160], we present evidence from the United States during a pre-election period, demonstrating how the spread of disinformation on Twitter is influenced by the U.S. electoral system, particularly through the dichotomy of swing versus safe states. Our observations indicate a significant focus of disinformation efforts on swing states, where traffic and the presence of untrustworthy news are markedly higher. Specifically, users affiliated with the Republican party are more likely to disseminate low-credibility news, with this group accounting for 91% of all links to untrustworthy news sources at the (political) community level. Moreover, our research finds that among classified users, automated accounts (bots) play a more prominent role than human-operated accounts in the spread of untrustworthy news regarding swing states.

The second contribution of this thesis stems from insights from the above mentioned contributions and a new line of research initiated in the course of my PhD. This contribution consists of the definition and

implementation of two novel methodologies designed to support disinformation mitigation, **which constitute the core innovations of this work**.

The first methodology, introduced in [155], is for the identification of echo chambers on online social media platforms. As case study, we consider the COVID-19 vaccination debate on Twitter/X in Italy. The methodology is grounded in the foundational definition of echo chambers, which are characterized by the simultaneous presence of two key phenomena: (i) interactions among users who share similar opinions and (ii) homogeneous exposure to the same news content. By leveraging techniques from network science, we detect the co-occurrence of these phenomena and their overlap within online discussions.

Our study in [155] focuses on the Italian COVID-19 vaccine discourse and reveals the prominent role of echo chambers in shaping discussions, particularly within right-leaning political communities. These echo chambers not only foster cohesive narratives among their members but also act as fertile environments for the spread of disinformation.

As previously mentioned, the methodology, based on the combined analysis of content exposure and user interactions, was validated using real-world data from Twitter/X, leading to the identification of nine distinct echo chambers. Specifically, these were composed of users who: (i) are densely connected and frequently retweet each other; (ii) are politically polarized, all belonging to communities aligned with the same political ideology; (iii) are highly active — despite representing only 0.35% of the total user base, they account for nearly 30% of the total activity within their political cluster; and (iv) consistently disseminate low-credibility content, often adopting radical ideological positions that intensify over time.

Further, a manual inspection of the most influential echo chambers revealed that their members continue to endorse extreme views on a variety of polarizing issues — including the war in Ukraine, immigration, and LGBT rights — and that conspiracy theories related to vaccines remain a central theme in their discourse, even years after the initial data collection.

This methodology, on one hand, it represents a novel and scalable approach for detecting echo chambers in social media environments; on the other, it serves as a powerful analytical tool to uncover behavioral patterns and ideological polarization among online users. The insights gained from this initial methodology, which leverages the analysis of social interactions, serve as a basis for certain developments related to the second methodology introduced.

The second methodology implements the automatic evaluation of the reliability of news outlets and articles, based on established principles of quality journalism.

The primary objective is to overcome the limitations of manual expert annotation—namely, high costs, limited scalability, and inconsistent evaluations due to subjective interpretations of guidelines—by leveraging reproducible computational tools. These limitations became particularly evident during my work as an expert annotator for the Global Disinformation Index (GDI), an experience that led to the publication of the official report on the Italian media landscape [148], and served as the starting point for this research direction.

The resulting body of work adopts a multidisciplinary approach, integrating artificial intelligence and network science to automate and scale the reliability assessment of news articles and publishers. Two main approaches have been explored so far:

1. In [157], we introduce, for the first time in the literature, a method for automatically estimating the reliability of online news outlets based solely on user–publisher interactions on social media, without relying on textual analysis or fact-checking. The method identifies a subset of users, termed *voters*, whose propensity to share high- or low-quality information is used as a proxy to infer the credibility of unlabelled news sources. This approach is rooted in the findings of [155], which show that user engagement patterns reflect the trustworthiness of the underlying narratives. Specifically, the dynamics of content dissemination provide critical signals for assessing the reliability of news outlets, with high- and low-quality content exhibiting distinct sharing behaviors. Our method enables

the evaluation of previously unassessed news sources, making it particularly suitable for low-resource settings, e.g., In terms of the difficulty of finding annotators, the possibly limited resources available to compensate them, and the tight deadlines for completing the task. This work, also lays the foundation for the development of the TROPIC software tool [156], released for the research community.

2. The second approach, presented in [158], investigates the use of Large Language Models (LLMs) to automate the reliability assessment of news outlets and articles by performing a structured analysis of the news, guided by the journalism criteria identified in [152]. In that prior work, we conduct a comparative analysis of the evaluation frameworks adopted by GDI[2] and NewsGuard[3], both of which rely on expert-based criteria assessed by diverse annotator panels. Our findings reveal strong conceptual alignment across criteria, substantial coverage, and overall agreement in classification outcomes.

   Building on these insights, [158] evaluates whether LLMs, in zero-shot settings, can replicate human expert judgments and help resolve inter-annotator disagreements. Agreement is measured using standard inter-rater metrics, and the model's ability to resolve disagreements is quantitatively assessed. Despite the absence of fine-tuning, the results show that LLMs can effectively reproduce human evaluations and reduce the costs of the annotation process, highlighting their potential as assistive tools for expert annotators. The study also reveals latent biases in the models, such as increased sensitivity to political bias and sensationalism compared to human annotators, opening the door to future investigations into their real-world applicability.

   On a related front, [17] explores the use of article-level textual features to assess the publisher reliability through a multi-label classi-

[2] [192]
[3] [136]

fication task, using a BERT-based model to assign reliability scores to individual articles.

1.2 Structure of the thesis

This thesis is structured into six main chapters, each addressing a distinct aspect of the overarching research question concerning the detection and mitigation of online disinformation. The organization follows a logical progression from foundational concepts and empirical observations to methodological innovations and applied solutions.

- **Chapter 2 – Related Work and Methods** provides the conceptual and methodological foundations of the thesis. It offers a review of the literature at the intersection of network science, disinformation studies, and computational social science, with a particular focus on prior works related to disinformation spread, echo chamber detection, and the evaluation of news credibility. This chapter draws upon and integrates material from several of the author's previous publications [17, 33, 57, 127, 152, 155, 157–160].
  Note. The state of the art reported here is aligned with the knowledge available at the time of publication of the referenced works, and may therefore not fully reflect the most recent advancements at the time of submission of this thesis.

- **Chapter 3 – The spread of Disinformation Dynamics on Twitter: Case Studies and Analysis** presents the first core contribution of this thesis. Through empirical case studies conducted in both Italy and the United States, this chapter investigates the influence of political affiliations, user types (including bots), and electoral systems on the dissemination of low-credibility content on Twitter/X. The findings are drawn from works published in [33, 127, 159, 160].

- **Chapter 4 – Disinformation and echo chamber detection during the COVID-19 vaccination debate in Italy** introduces the first methodological innovation of the thesis: a novel, scalable approach for the

detection of echo chambers in online social media, based on the joint analysis of user interactions and content exposure. This chapter is based on the methodology and empirical analysis presented in [155].

- **Chapter 5 – Automatically Evaluating News Publisher and Articles Trustworthiness** introduce the second methodological contribution of this thesis. The chapter begins by introducing the work in [152], which provides a solid conceptual basis and compiles a set of widely used and expert-designed criteria for assessing the reliability of news content. These criteria serve as the backbone for a structured approach aimed at automating the evaluation process. Building upon this framework, the chapter then explores the potential of Large Language Models (LLMs) to replicate expert annotations by following these structured guidelines, as investigated in [158]. This line of research assesses the capabilities of LLMs to perform reliability evaluations in a scalable and reproducible manner, while addressing key challenges related to human annotation and inter-annotator agreement.

- **Chapter 6 – Social interactions analysis for Trustworthiness rating and the development of a first SW prototype** concludes the presentation of the second core contribution of the thesis. It introduces the methodology proposed in [157], which estimates the trustworthiness of news sources based on user–publisher interaction patterns on social media. The chapter also describes the design and implementation of the `TROPIC` software tool [156], which operationalizes this approach and is made available to the research community. In addition, the chapter presents a complementary line of research focused on article-level trustworthiness assessment, using text-based features and deep learning techniques, as discussed in [17].

- **Chapter 7 – Conclusions** summarizes the key findings of the thesis, discusses their implications, and outlines several directions for future research. Particular emphasis is placed on the integration of

AI-driven approaches into disinformation detection pipelines, the ethical considerations surrounding automated judgments, and the potential for future applications in education, media literacy, and platform governance.

Each chapter is self-contained and contributes to a coherent narrative that spans from empirical investigations to methodological advances, with the ultimate goal of enhancing our understanding of online disinformation and improving our capacity to counter it through computational means.

1.3 Self-plagiarism disclosure

In accordance with academic integrity and ethical publication standards, this thesis includes content that has been previously published in peer-reviewed journals and conference proceedings. While the structure of the dissertation and the narrative have been adapted to the requirements of a doctoral thesis, portions of the text, figures, and results are derived from the author's previously published works. These contributions are appropriately cited in the relevant sections of the thesis, and detailed below, in relation to each chapter.

This disclosure aims to ensure transparency and to explicitly acknowledge the reuse of content authored by the candidate in prior scientific publications. In all cases, the intellectual property remains with the author(s), and the reuse of material complies with the copyright and licensing agreements of the respective publishers.

- **Chapter 2** (Related Work and Methods): draws upon and integrates literature reviews and theoretical insights presented in [17, 33, 57, 127, 152, 155, 157–160].

- **Chapter 3** (The spread of Disinformation Dynamics on Twitter: Case Studies and Analysis): is based on content and results previously discussed in [33, 57, 127, 159, 160], including methodological aspects and case studies.

- **Chapter 4** (Disinformation and echo chamber detection during the COVID-19 vaccination debate in Italy): incorporates material, analysis, and figures from the work published in [155].

- **Chapter 5** (Automatically Evaluating News Publisher and Articles Trustworthiness): reuses results, experimental setups, and methodological insights from [152, 158].

- **Chapter 6** (Social interactions analysis for Trustworthiness rating and the development of a first SW prototype): is derived from the findings and methodological developments presented in [17, 156, 157].

For completeness, it is worth noting that the work presented in [159] represents an extended and refined version of the preliminary results previously discussed in [160], with additional data analysis and theoretical contributions.

The reuse of these materials has been done with the intent to provide a coherent and unified view of the research trajectory, and to clearly highlight the original contributions developed during the doctoral program.

1.4 Declaration

I hereby declare that this doctoral thesis has been written entirely by me. The work presented in this thesis is original and represents the result of my own research activities.

The thesis is based on a collection of scientific articles authored or co-authored by me, which are explicitly reported on a chapter-by-chapter basis in Section 1.3. Each chapter builds upon these published works and integrates them into a coherent and unified framework. Where necessary, minor adaptations have been made for consistency, clarity, and continuity, without altering the scientific content or the original results.

I also acknowledge the use of AI-based tools as a support for improving the clarity, readability, and stylistic consistency of the text. These tools were employed exclusively for language refinement and editorial

assistance (e.g., rephrasing, grammar, and style). They did not contribute in any way to the scientific content, research design, data collection, data analysis, interpretation of results, or conclusions of this thesis. All ideas, analyses, and scientific contributions presented in this work are entirely my own.

1.5 Reproducibility of the studies

1.5.1 Twitter Data Collection

To promote transparency, all Twitter datasets used in this Thesis have been made publicly available. Below, we summarize each dataset, including collection period, methodology, keyword lists, and release location.

COVID-19 pandemic in Italy **Reference:** [33] — presented in Chapter 3

Description: Using Twitter's Streaming API (see Section 1.5.1), we collected approximately 4.5 million tweets in Italian between February 21 and April 20, 2020. The dataset was obtained using keyword- and hashtag-based tracking related to the COVID-19 pandemic. The full corpus initially included tweets in multiple languages; Italian tweets were filtered post hoc. Temporary interruptions occurred due to connection issues (February 27 and March 10), and occasional rate-limiting by the Twitter API was observed. However, the entropy-based network validation procedure applied later (see Section 2.3.2) mitigates concerns about completeness.

Keywords used: `coronavirus, ncov, covid, SARS-CoV2, #coronavirus, #coronaviruses, #WuhanCoronavirus, #CoronavirusOutbreak, #coronaviruschina, #coronaviruswuhan, #ChinaCoronaVirus, #nCoV, #ChinaWuHan, #nCoV2020, #nCov2019, #covid2019, #covid-19, #SARS_CoV_2, #SARSCoV2, #COVID19`

Access: `https://toffee.imtlucca.it/datasets`

2020 U.S. Presidential election Twitter debate **References:** [159, 160] — presented in Chapter 3

Description: This dataset includes approximately 5.3 million tweets collected via Twitter's Streaming API from October 27 to November 3, 2020, i.e., the week before the U.S. presidential election. Data were collected using combinations of keywords related to four swing states and four safe states, paired with candidate names (Trump and Biden).

Keywords used: `arizona biden, arizona trump, florida biden, florida trump, michigan biden, michigan trump, pennsylvania biden, pennsylvania trump, new jersey biden, new jersey trump, indiana biden, indiana trump, washington biden, washington trump, louisiana biden, louisiana trump`

Access: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ANBPTC`

Italian COVID-19 vaccination debate **References:** [155, 158] — dataset used in Chapter 4 and Chapter 6

Description: Using Twitter's Streaming API, we collected approximately 1.87 million tweets in Italian between September 1 and September 24, 2021. The dataset comprises about 136,000 unique users, and roughly 220,000 tweets contain URLs. The data collection was driven by keywords linked to the vaccination campaign and public discourse around it.

Keywords used: `vax, vaccino, vaccini, vaccinarsi, novax, Astrazeneca, Pfizer-BioNTech, Moderna, Sputnik, greenpass` (English meanings: *vaccination terms, anti-vax, COVID-19 vaccines, vaccination certificate*)

Access: `https://figshare.com/articles/dataset/Anonymized_data_for_the_analysis_of_Entropy-based_detection_of_Twitter_echo_chambers_/25460962/1`

Change in Twitter property and the advent of Twitter/X In late October 2022, Twitter, Inc., the American social media company, underwent

a significant transformation when it became the property of Elon Musk[4]. This transition ushered in a series of radical changes and reforms that embraced both managerial and technical aspects.

One pivotal development of note for the scientific community was the discontinuation of Twitter's free API tier by February 2023, to be replaced with a 'basic paid tier'[5]. For researchers and developers, this change meant that Twitter content was no longer available for research purposes without subscribing to a significantly different paid plan. It also presented a challenge in terms of rehydrating the datasets currently in use. Despite the fact that our dataset was collected during a period of free access (from October 27 to November 3, 2020), the policy appears to be unchanged up to now [6]. Thus, we acknowledge the potential obstacles to the reproducibility of the experiments conducted using Twitter/X data.

However, we maintain that the proposed methodology intriduced in this Thesis remains highly adaptable to other online social networks.

1.5.2 Code and software availability

To ensure full reproducibility of the experiments and tools developed in this Thesis, the following code and software resources are publicly available:

- **Reliability Criteria Evaluation with LLMs** — accompanying the study presented in Section 5.2 [158]:
  `https://github.com/manuelP88/ReliabilityCriteriaEvaluationWithLLM`
- **TROPIC Demo Platform** — accompanying the study presented in Chapter 6.1.6 [156]:
  `https://tropic.iit.cnr.it/`

[4] https://www.nytimes.com/2022/10/27/technology/elon-musk-twitter-deal-complete.html

[5] https://twitter.com/XDevelopers/status/1621026986784337922

[6] https://developer.twitter.com/en/developer-terms/policy#4-e

Chapter 2

Related Work and Methods

2.1 The spread of Disinformation on social media

The proliferation of online disinformation has become a pressing global concern, often likened to a modern-day plague. During major events—such as elections, public health crises, or geopolitical conflicts—disinformation frequently takes center stage, sowing confusion and distrust among the public[1], with potentially serious offline consequences.

The rise of the internet and social media platforms has democratized access to information and significantly increased the diversity of news sources, including contributions from individual users. While this development offers many benefits, it has also enabled the unchecked dissemination of low-quality and misleading content. Such unmediated communication channels have contributed to the pollution of public discourse across several domains, including politics, healthcare, education, and environmental issues [23].

Recognizing the severity of the problem, the European Union took a formal stance as early as 2020. In the Joint Communication entitled

[1] `https://www.un.org/en/countering-disinformation`. All URLs were last accessed on January 26, 2024.

"Tackling COVID-19 Disinformation – Getting the Facts Right" [2], the High Representative of the Union for Foreign Affairs and Security Policy explicitly stated the need for coordinated action:

> "Combating the flow of disinformation, misinformation [...] calls for action through the EU's existing tools, as well as with Member States' competent authorities [...] enhancing citizens' resilience."

Further underlining the global relevance of the issue, the World Economic Forum has recently listed mis- and disinformation among the most critical short-term global risks[3]. The rapid, large-scale circulation of false yet persuasive narratives can influence public opinion, polarize electorates, heighten social tensions, and intensify ongoing crises—from international conflicts to public health emergencies. These dynamics ultimately threaten collective security and undermine trust in democratic institutions [117].

Scientific research on fake news and disinformation highlights how such content often spreads within tightly-knit communities. These dynamics are exacerbated by polarization and the formation of echo chambers [15, 60], and are further amplified by the activity of automated accounts such as social bots [75, 201].

In the remainder of this section, we review the related work on the spread of disinformation on social media, with a particular focus on two major domains: the COVID-19 pandemic (including vaccine-related discussions) and political elections. We also discuss the key underlying mechanisms that enable and sustain disinformation campaigns. These include the emergence of polarized and ideologically aligned user clusters—often referred to as *discursive communities*—the role of echo chambers in reinforcing biased narratives, and the strategic use of bots and other forms of automation to artificially boost the visibility and credibility of misleading content.

These topics are examined in detail in the following subsections: disinformation related to pandemic (Section 2.1.1), election-related disinfor-

[2](June 10, 2020, `https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=JOIN:2020:8:FIN`)

[3]`https://www.weforum.org/publications/global-risks-report-2025/`

mation (Section 2.1.2), echo chambers (Section 2.1.3), and the role of bots (Section 2.1.4).

2.1.1 Disinformation during the COVID-19 pandemic

Covid-19 Related Work As in any disaster, natural or otherwise, people are exposed to online misinformation. This is the case of COVID-19 too: the physical pandemic was quickly complemented by the so-called COVID-19 infodemic, i.e., the diffusion of a great amount of low-quality information about the virus. Academia has stepped up its efforts to combat this infodemic. Here, we briefly review some of the most relevant articles in the area.

Nevertheless, the analysis of the existence and diffusion of polarised / biased / false stories about COVID-19 has immediately attracted several scholars, which are focusing on different facets of these phenomena, such as: the most searched terms on Google related to COVID-19 [170], the existence of Facebook groups experiencing an extreme exposure to disinformation [39], the change in the type of information on Twitter during the evolution of the pandemic [82] and the *disinformation epidemiology* on various online social platforms [44].

Rovetta et al., in [170], explore the internet search activity related to COVID-19 from January to March 2020, to analyse article titles from the most read newspapers and government websites, 'to investigate the attitudes of infodemic monikers circulating across various regions and cities in Italy'. The study reveals a growing regional and population-level interest in COVID-19 in Italy, highlighting how the majority of searches concern -often unfounded- remedies against the disease.

Work in [82], by Gallotti et al., develops an Infodemic Risk Index to depict the risk of exposure to false information in various countries around the world. Regarding healthcare news, the authors find that even before the rise of the pandemic, entire countries were exposed to false stories that can severely threaten public health.

Hossaini et al. [103] release COVIDLies, a dataset of 6,761 expert-annotated tweets to evaluate the performances of existing NLP systems

in detecting false stories about COVID-19. Still regarding datasets, work by Zhou et al. [218] presents ReCOVery, a repository of more than 2k news articles on Coronavirus, together with more than 140k tweets testifying the spreading of such articles on Twitter. Chen et al., in [42], present to the scientific community a multilingual COVID-19 Twitter dataset that they have been continuously collecting since January 2020. Celestini et al., in [39], collect and analyse over 1.5 M COVID-19-related posts in Italian. Findings are that, although controversial topics associated to the origin of the virus circulate on social networks, discussions on such topics is negligible compared to those on mainstream news websites.

Pierri et al., in [149], provide public access to online conversations of Italian users around vaccines on Twitter. The authors report a consistent amount of low-credibility information already circulating on Twitter alongside vaccine-related conversations. Still regarding COVID-19 vaccination campaigns, De Verna et al. collect a Twitter dataset of English posts, giving statistics about hashtags, URLs, and number of tweets over time, through a dashboard.

Sharma et al, in [181], consider the role of Twitter bots in the pandemic online debate. By moving away from the research trend of detecting bot squads, on the basis of features concerning coordination and synchronous behavior among a group of accounts, they propose an approach to automatically uncover coordinated group behaviours from account activities and interactions between accounts, based on temporal point processes.

A lot of work examines Twitter, because of the availability of public APIs for data gathering. Instead, Yang et al. [212] analyse and compare the presence of links pointing to low-credibility content both on Twitter and Facebook. Misinformation 'superspreaders' and evidences of coordinated sharing of false stories about COVID-19 are present on both the platforms. Still at a narrower granularity, Cinelli et al., in [44], carry on a massive analysis on Twitter, Instagram, YouTube, Reddit and Gab. The authors characterize COVID-19 information spreading from questionable sources, finding different volumes of misinformation in each platform.

This brief literature overview on the COVID-19 infodemic, although not exhaustive, highlights that the spread of misinformation on pandemic-related issues on the internet and social media is a major issue. Scientists propose various methods to detect false information about the virus. Aligned with this line of research, in Section 3.1 of this manuscript we quantify the *effective* level of misinformation about the pandemic exchanged on Twitter during late winter and early spring in 2020 in Italy, with a special focus on the role of the Italian political communities.

2.1.2 Disinformation in political elections

To give a glaring example, on 3 September 2021, Jacob Anthony Angeli Chansley was sentenced to 41 months in prison for obstruction of justice. Chansley, also known by various nicknames such as "QAnon Shaman", participated with other far-right activists in the attack on the United States Capitol on January 6, with the intention of disrupting the certification of election results. He was convinced by online disinformation campaigns about fraud against former President Donald Trump in the election[4] and a known conspiracy theorist [195]. This egregious news episode is just the tip of the iceberg of a series of consequences that the proliferation of disinformation online has on society around election time. A nationwide survey in the U.S. after the 2018 midterm elections found that trust in the electoral system dropped significantly after Republican supporters spread unsubstantiated rumors of fraud online, driving voters away from politics, despite fact-checking efforts to disprove such rumors [13].

Numerous scholars have delved into the U.S. 2016 and 2020 presidential elections, seeking to gauge the extent to which digital disinformation influenced Trump's victory/defeat, yet a definitive answer remains elusive. For example, the study by Georgacopoulos et al. [86] reveals that in the three months leading up to the 2016 election, fake news supporting Trump was shared on Facebook nearly four times more than the eight million fake news items supporting Clinton.

[4] `https://en.wikipedia.org/wiki/Jacob_Chansley`.

Examining more than 170 million tweets exchanged on Twitter in the five months leading up to the same election, [21] found that trustworthy news stories overwhelmingly came from journalistic sources and verified Twitter accounts. In contrast, conspiracy theories, fake news, and highly partisan news largely originated from unofficial Twitter clients, posted by unknown users who often disappeared from the platform, or through automated accounts commonly referred to as social bots. Shao et al. [179] also highlighted the role of Twitter bots, showing how these bots were primarily responsible for the early spread of disinformation by engaging influential accounts through mentions and replies.

The evolutionary adaptation of bots, characterized by their increasing ability to evade detection techniques, is well documented [75] and [54]. Luceri et al. [124], for example, found that from the 2016 presidential election to the 2018 midterm elections, political discussion bots evolved to the point where they became increasingly indistinguishable from humans.

Electoral politics and social media The study presented in Section 3.2 investigates whether the peculiar structure of the U.S. presidential election system—where physical campaigning is particularly intense in swing states—also manifests in online campaigning. From this perspective, we found it interesting to examine several works that have studied the relationship between electoral politics and the use of social media.

In particular, one paper used opinion mining techniques to examine in real-time the correspondence between exit poll results and the opinions of Twitter users in the week leading up to the 2020 U.S. election [10]. In fact, it was possible to predict the president-elect in 10 of the 11 states considered to be swing states in 2020, even exceeding the percentages of the most recent physical exit polls. In this case, we could say that there was more than a campaign mimicry, there was a match precisely in the prediction of the winner.

Similarly, a series of articles focused on the influence that exposure to Facebook and Instagram feeds had on the voting decisions of U.S. citi-

zens during the 2020 presidential election campaign[5].

In a first paper [94], nearly 45k users from the two platforms were recruited. One control group retained feed visualization settings dictated by the platforms' algorithms, while the others had their settings altered, allowing users to see the most recent feeds. Despite these changes, being fed content based on history rather than interests reportedly had no effect on the users' political attitudes and off-platform behaviors.

A second paper was interested in the effects of viewing news shares on Facebook, again during the U.S. 2020 election period [95]. The removal of re-shares significantly reduced the amount of political news, including content from untrusted sources. Despite this, there were no changes in political polarization or individual-level political attitudes in the subgroup that did not see shared content. From these studies, it appears, as above, that online users were little influenced by the content proposed by the platform and generated by other users, i.e. their voting intention was not surprisingly changed.

Many works on the relationship between electoral politics and social media focus on Europe, and in particular the European Parliament elections. The number of studies is probably due to the different facets that this particular election has in Europe, from the number of elected officials from each country, the presence of numerous local parties, and the doubt that campaigns are driven more by individual member state interests than by all as a community. In fact, according to experts, the European Parliament elections are experienced by European citizens as 27 different election campaigns, one for each member state[6].

One of the study's findings is similar to ours: the more citizens the candidate is expected to represent, the more activity the candidate has on the social network. Thus, when the physical campaign becomes heavy and complex to manage, the campaign activity is changed online. How-

[5]Social Media and the 2020 election: `https://www.princeton.edu/news/2023/07/28/social-media-polarization-and-2020-election-insights-spias-andrew-guess-and`

[6]Make or break for the EU? Europeans vote in June with far right on the rise: `https://www.theguardian.com/world/2024/jan/03/make-or-break-for-the-eu-europeans-vote-in-june-with-far-right-on-the-rise`

ever, the social activity of the Member of the European Parliament -or MEP- candidates is limited to the election period, indicating that social networks are not used to cement a relationship with the electorate, but rather for the sole purpose of garnering votes. Some studies justify the use of social networks by candidates as a means of broadcasting only on the grounds that interaction with users leads to insults and harassment. This was the result of an analysis in [194], which found that the content of tweets directed at MEP candidates was often rude and harassing.

A continuation of the study in [141] is the one in [56], where the authors looked at outgoing members of the European Parliament who, after the 2014 elections, ran again as candidates in the same elections in 2019. The purpose of the work was to see whether the different candidates focused their online campaigns more on themselves or on the party they represented. One finding that emerged, compared to running a physical campaign, was that the relatively low-cost nature of social media allowed some politicians to simultaneously campaign as individuals and as *party animals* in a way that analog campaigning could not.

Social campaigns for the 2019 European Parliament have been widely studied in the literature. The article collection in [97] analyzes how political parties in 12 member states used Facebook in the lead-up to elections. Again, the overall message is that social media was used to persuade the public to vote for the candidate, rather than using the platform to interact and mobilize voters.

This is also the conclusion of other work that has examined the relationship between social campaigning and electoral politics in the run-up to national elections in Europe. One example is the study in [25], in which two electoral events in Britain, in 2015 and 2017, are considered. What comes out is that, once again, the tendency of candidates to use the social medium is to broadcast the program and make propaganda, and not to interact with the electorate, even with the idea of maintaining solid contact after the election.

Disinformation flows in U.S. Presidential elections In the introduction, we already cited some analyses on detecting online disinformation

flows in the periods leading up to the 2016 and 2020 U.S. presidential elections [86], [21], [179]We can also cite [70], where the authors analyzed both mainstream and social media coverage of the 2016 U.S. presidential election. Their analysis revealed the asymmetric nature of the media landscape, with Twitter displaying a more partisan tendency. Donald Trump's campaign primarily emphasized immigration, whereas Hillary Clinton's coverage tended to emphasize various scandals. Right-wing media tended to favor pro-Trump outlets, while left-wing media focused on traditional objective journalism.

As mentioned at the beginning of the article, Howard et al. in [104] conducted a study centered on analyzing tweets related to swing and safe states during the pre-election period of the 2016 U.S. presidential election. Their findings revealed a significant concentration of polarized news in tweets associated with swing states with a significant number of presidential electors. This work serves as a valuable precursor to our analysis (see Section 3.2). It is important to note, however, that the election we examine differs from the one in their study. In addition, our research focuses on the behavior of social bots, and most importantly, we have refined our dataset by employing a process based on complex network analysis to filter out noise.

2.1.3 Echo-chambers on online social media

In the virtual world, individuals often seek information that confirms their existing beliefs and engage with users who share similar viewpoints. This behavior leads to the formation of echo chambers, defined as "bounded, enclosed media spaces that have the potential to both amplify the messages delivered within them and insulate them from rebuttal" Del Vicario et al. [60], Garrett [85], Jamieson and Cappella [109], and Zollo et al. [219].

While the creation of echo chambers is worrisome on its own for the radicalization of online –and, therefore, also offline– debates, their effects are also more dangerous when they are affected by misinformation. With the advent of the internet and Online Social Networks, the pro-

duction and fruition of information feature less mediated procedures, where content and quality do not always go through a rigorous editorial process [40, 83, 198]. Moreover, the fast processing of data reduces the capability to objectively judge the information provided [145], possibly accelerating the diffusion of low-quality posts. Intuitively enough, the presence of unreliable pieces of information can increase the formation of echo chambers in which prejudicial, biased, misleading, if not outright fake news is exalted.

Echo-chamber detection on OSMs

The detection of echo chambers has been generally approached by the literature starting from online content whose nature is known a priori. Through the analysis of the social accounts that interact with specific content, e.g. via likes, shares, retweets, and comments, it has been shown how information relating to specific narratives attracts distinct communities. Work in [60], by Del Vicario *et al.*, focuses on public Facebook pages divided into two groups: conspiracy theories and news about science (conspiracy theories are 'the pages that disseminate alternative, controversial information, often lacking supporting evidence' [60]). The findings are that users are divided into homogeneous clusters: by analysing the accounts that share news about science and conspiracies, they are bound by ties of friendship in the network. Quoting the authors: 'different contents generate different echo chambers, characterized by the high level of homogeneity inside them'.

Homogeneity is not only about friendship, but also about emotional approach and reaction to debunking attempts. Zollo *et al.*, in [220], establish how users polarised on conspiracies express more negative feelings in their comments than users polarised on science news. Work in [219] confirms how the echo chamber paradigm goes hand in hand with the confirmation bias phenomenon –the users' tendency to look for, prefer and interpret information in line with their thoughts [114, 140], while ignoring or downplaying evidence that contradicts their beliefs: interactions with debunking posts (i.e., posts that provide fact-checked infor-

mation to specific topics) are overwhelmingly from users biased towards science or non-biased users.
The above examples show how echo chambers emerge by analysing thematic pages and noting that users divide into distinct communities according to the page topic. Going deeper, it also emerges that consecutively sharing users are linked by friendship links on the network.

Interestingly for the purpose of our contribution (see Chapter 4), other studies have instead analysed the dynamics of information exposure by considering the news URLs present in the posts. This is the case, e.g., of work by Weaver *et al.* [204], in which the network of densely-connected news articles is constructed. It starts from the number of news URLs shared by each user, to arrive at the weighted network of news URLs in which the weights between two URLs identify how many users have reshared the URL pair. Leveraging a state-of-the-art community detection algorithm, communities of co-shared news items are found, distinct in terms of political leaning (i.e., left-leaning and right-leaning).
Guarino *et al.*, in [92], consider public Facebook pages, without however knowing *a priori* the kind/quality/reputability of their content. Focusing on the activity of users sharing links to pieces of online news, the authors construct the bipartite network of users/shared URLs and apply the Bipartite Configuration Model (BiCM) introduced in [172] to project the bipartite network on the two levels, the user level and the URL level. Applying the BiCM assures that two accounts (resp., two URLs) are connected if the number of URLs shared by both the accounts (resp., if the number of accounts sharing both the URLs) is so large that it cannot be explained by the degree distribution of the two layers only.

2.1.4 Bots on online social media

Bots are computer algorithms often employed for malicious purposes: they distribute spam, promote public figures, and ultimately distort public opinion. To counter their proliferation, several detection approaches have been developed, mostly relying on supervised and unsupervised

classifiers that exploit a wide range of account features, from simple metadata to computationally expensive behavioral traces obtainable through the Twitter public APIs.

One of the first documented cases of social media manipulation occurred during the 2010 Massachusetts election, when a handful of automated Twitter accounts spread misinformation about the Democratic candidate [133]. Although Twitter quickly banned the accounts, the damaging content had already propagated across platforms, even reaching Google search results. This early example highlighted the societal risks posed by bots. Since then, automated accounts have repeatedly influenced major events, including the 2016 and 2020 U.S. elections [181], the Brexit referendum, and debates on immigration and the COVID-19 pandemic [30, 32].

Over the past decade, bots have grown increasingly sophisticated [54, 75]. Fake accounts can now build friendship networks, sustain conversations with real users without immediate detection, and act in highly coordinated teams. These behaviors make them difficult to identify with traditional classifiers, which often analyze accounts in isolation. As Menzer observes, "There is no shortage of research challenges, even 10 years later, to try to identify this kind of manipulation" [185].

Detection of social bots

Research on bot detection spans roughly 15 years, with the first works appearing around 2010 [133, 213]. Early approaches (2010–2014) relied mainly on supervised machine learning and the analysis of individual accounts: classifiers were applied separately to each account, which was then labeled as bot or human [51].

From 2014 onwards, researchers began shifting attention to groups of accounts, aiming to detect coordinated behaviors. Techniques explored during this phase included detecting anomalies in synchronicity and activity patterns [87, 112], loosely synchronized actions [36], similarities in behavioral sequences [53], and distributions of reputation scores [202]. This collective perspective proved crucial, as many bots act in teams that are more easily revealed through group-level analysis.

In recent years, these two strands have converged into the development of general-purpose classifiers. These models aim to detect both individual bots and accounts that belong to coordinated groups, thus broadening their applicability. For example, Sayyadiharikandeh et al.[174] proposed an ensemble of specialized supervised models tailored to different bot classes, whose outputs are combined through a voting scheme. Similarly, Yang et al.[210] designed a scalable framework to process Twitter's full public stream in real time by exploiting low-cost features, mainly from account profiles. Comparable low-complexity features had already been successfully used to detect fake followers [52], confirming their continued relevance in more recent detection systems.

Other contributions investigate the effectiveness and complementarity of existing detectors. Schuchard and Crooks [176], for instance, applied three state-of-the-art systems—Botometer [200], DeBot [41], and BotHunter [14]—to accounts active during the 2018 U.S. midterm elections. Their study revealed poor agreement among detectors, with only a handful of accounts unanimously labeled as bots, underscoring the need for multi-method approaches. Ferrara [72] also explored lightweight features, such as geolocation and activity statistics (e.g., tweet frequency, retweet ratios), to build a simple yet effective detection system.

Hybrid approaches have also been explored. El-Mawass et al. [128] proposed a cascade of classifiers that feeds outputs into a probabilistic graphical model, enabling the propagation of beliefs to similar accounts. This reduces false positives and improves recall while maintaining precision.

Alongside these account-level strategies, several frameworks focus explicitly on coordinated activity. Tools such as BotSlayer [105] highlight emerging campaigns by detecting anomalies in hashtags, links, and trending media. Similarly, Sharma et al. [181] proposed a temporal point process framework to uncover coordinated behaviors from account interactions.

Recent work has also examined bots in specific contexts, especially U.S. elections. For example, Bellutta et al.[11] found abnormal surges in account creation during the 2020 U.S. elections, associated with suspi-

ciously homogeneous behavior and the sharing of low-credibility sources. Ferrara et al.[74] revealed systematic efforts linking bots, hyper-partisan outlets, and conspiracy groups, while Linhares et al.[122] identified communities spreading allegations of election fraud, comprising both suspended and self-deleted accounts. Yang et al.[207] highlighted hyperactive accounts that flooded Twitter with political content during the 2018 midterms, often displaying suspicious patterns of activity.

Overall, the field of bot detection has evolved from early single-account classifiers, through group-level analyses of coordinated activity, to modern attempts at unifying both perspectives into general-purpose systems. Yet, challenges remain, as detection methods often disagree, bots continuously adapt, and coordinated manipulation campaigns persist.

2.2 Profiling news media reliability

The rise of online misinformation has made it increasingly difficult for readers to assess the quality and reliability of the content they consume. While fact-checking individual claims remains an essential tool, it is often impractical for novel or rapidly evolving information. To address this, researchers and organizations have turned to source-level evaluation, assessing the trustworthiness of entire news outlets as a proxy for content quality. This approach allows for more scalable interventions and offers a promising complement to content-level analysis.

2.2.1 Expert-based evaluations of news trustworthiness

Several non-profit and journalistic organizations—such as [136], the [192] (GDI), [130], and [1]—have developed expert-based systems to evaluate the trustworthiness of news publishers. These evaluations are grounded in well-defined, transparent criteria aligned with professional journalistic standards, including credibility, transparency, factual accuracy, and editorial responsibility [137, 192].

While each organization employs slightly different evaluation protocols, their approaches share a reliance on trained expert annotators and

structured rating frameworks. For instance, GDI uses country-specific experts who conduct multi-stage reviews that include ownership investigation and systematic content analysis. Similarly, NewsGuard's criteria assess whether outlets repeatedly publish false information, clearly label advertising, distinguish between news and opinion, and disclose ownership structures.

Despite methodological variations, multiple studies have demonstrated a substantial degree of agreement among these institutions, indicating both inter-rater reliability and cross-framework consistency [120, 152]. This convergence supports the validity of expert-based evaluations as benchmarks for media trustworthiness.

However, a major limitation of these approaches lies in their scalability. Because the evaluations are manually conducted, they often require several weeks or months of sustained expert work, making them resource-intensive and challenging to update in real time. These constraints have motivated a growing interest in developing automated methods capable of replicating or approximating expert assessments at scale.

Automated approaches The term *infodemic*, popularized during the COVID-19 pandemic, highlights the urgent need for scalable and robust mechanisms to counter the growing volume of online misinformation. Initiatives such as the [68], along with international organizations like the United Nations and the World Health Organization (WHO), have emphasized the importance of fostering trusted information ecosystems. These efforts have called for increased collaboration among fact-checkers, researchers, and communication professionals.

While such initiatives contribute significantly to public awareness and the debunking of falsehoods, their impact is inherently limited in scope. This limitation has spurred growing interest in automated approaches that can scale the evaluation of online information credibility. In recent years, the academic, journalistic, and technological communities have increasingly focused on automating the assessment of trustworthiness—particularly of news publishers.

In this thesis, we explore three distinct automated approaches to eval-

uating the reliability of online news content and its sources. Previous work in the literature can be broadly categorized into the following two paradigms (and possible hybrid versions).

- **Content-based Classification of News Articles and Publishers**
  In this approach, the credibility of news articles is inferred from their textual content. Typically, supervised machine learning models are trained on labeled datasets, where labels are often derived from the known trustworthiness of the articles' publishers. A detailed review of these content-based classification methods is provided in Section 2.2.2.

- **Social Interaction-based Classification of News Articles and Publishers**
  As discussed in Section 2.2.3, this line of research leverages user interactions on social media—such as retweets, replies, and URL sharing—as proxies for estimating the trustworthiness of news articles and publishers. Rather than analyzing content directly, these methods rely on patterns of engagement e.g, surrounding news URLs. Although these techniques offer scalability and the potential for near real-time monitoring, they may sacrifice the interpretability offered by content-based models.

The recent emergence of Large Language Models (LLMs) introduces a new paradigm, in which models are prompted to assess the reliability of publishers or their content based on expert-defined criteria. This approach aims to simulate expert annotation at scale. A review of existing work in this area is provided in Section 2.2.4. An open research question in this domain concerns the extent to which LLM-generated assessments align with those produced by human experts.

2.2.2 Content-Based classification of news articles

Several studies have proposed to infer the credibility of news publishers by analyzing the content of individual articles. This typically involves

training classifiers using labeled datasets, where article-level labels are derived from the known trustworthiness of their sources.

Earlier efforts in this domain utilized traditional machine learning techniques such as Support Vector Machines, often trained on bag-of-words (BoW) representations or n-grams [88, 175]. These approaches, while foundational, were limited in their ability to capture the nuances of language and article structure.

More recent methods have shifted toward using deep learning, particularly transformer-based architectures like BERT, for fine-grained article classification [20, 162]. These studies frame the problem as either binary (trustworthy vs. untrustworthy) or multiclass classification, based on source-derived article labels. For instance, Przybyla [162] achieves high accuracy when the source of an article is known during training, while Bianchi et al. [17] explore multiclass prediction based on NewsGuard-derived scores. Additionally, some works analyze structural features of news articles. For example, Dai et al. [55] examine whether articles follow typical journalistic structures such as the Inverted Pyramid, while Heravi et al. [100] build on these ideas to map writing style to credibility cues. The field has also seen contributions from shared tasks such as SemEval and CLEF. The 2023 SemEval Task 3 [150] focused on detecting the category, framing, and persuasion techniques in online news across multiple languages and genres. Other challenges, such as CheckThat!@CLEF [135], addressed infodemic-era misinformation, encouraging systems that flag articles or tweets worth fact-checking. Framing and topic detection—core components in understanding persuasive and potentially biased narratives—have been the subject of targeted research as well [37, 67, 123]. These perspectives enrich article-level classification efforts by adding contextual layers that go beyond surface-level features. Despite substantial progress, challenges remain regarding multilingual generalizability, dataset availability, and the robustness of label sources. Many models still rely heavily on the assumption that article credibility mirrors that of its publisher, an approximation that may not always hold in practice.

2.2.3 Social interaction-Based models

Recent literature has explored the use of user interactions on social platforms as a proxy for evaluating the credibility of publishers [90, 157]. Rather than focusing on the textual content of articles, these studies examine behavioral traces—how users engage with URLs shared by different outlets, including posting, reposting, and commenting behaviors.

Such approaches are grounded in earlier findings showing that false or misleading information tends to spread more rapidly and broadly than truthful content [179, 203]. For example, Gravino et al. [90] analyze the alignment between user interests, as reflected in Google Trends, and the news supply from online outlets. They show that content from untrustworthy publishers more closely matches public interest than that from mainstream outlets, suggesting a mechanism through which such publishers capture attention by tailoring content to audience preferences. This work supports the idea that interaction signals—particularly those generated around low-credibility sources—may be valuable indicators of trustworthiness.

A key advantage of interaction-based models is their scalability and language-independence. Since they rely on network dynamics and behavioral signals, rather than linguistic features, they can be deployed across languages and regions with minimal adaptation. Furthermore, they enable real-time monitoring, making them suitable for early detection and rapid response to emerging information threats.

Building on this line of work, more recent graph-based approaches have modeled the news ecosystem as a heterogeneous network comprising users, articles, and publishers [65, 182, 183, 191]. These models often leverage Graph Neural Networks (GNNs) to jointly learn the credibility of all entities in the graph. For instance, Shu et al. [183] propose TriFN, a tri-relationship embedding framework that captures user-news, publisher-news, and user-user interactions for fake news classification. Dou et al. [65] introduce a preference-aware method that fuses user engagement signals with content-based features via GNNs to assess the credibility of news articles. Shrestha et al. [182] propose a joint learn-

ing framework that simultaneously predicts the trustworthiness of news articles, users, and publishers using Relation Graph Convolutional Networks.

While most of these models operate at the article level, our recent work [157] introduces a novel approach to model publisher trustworthiness directly from social interactions, without relying on content analysis. Specifically, we analyze patterns of engagement with news URLs—referred to as *News Engagement Clusters* (NECs)—that emerge from ongoing social media discussions. These clusters group together links that have attracted attention from statistically significant groups of users, often revealing coherent behavioral signals associated with individual publishers.

This approach offers a complementary alternative to expert-based evaluations, such as those provided by organizations like NewsGuard or Media Bias/Fact Check, which are rigorous but time-consuming [120, 152]. It also differs from recent efforts that apply classification models (including large language models) to individual news articles [20, 162, 208], which require large labeled datasets and may struggle with generalization across domains.

To the best of our knowledge, our method is the first to apply complex network analysis of URL engagement patterns to infer publisher reputations from user behavior alone. By focusing on emergent user engagement structures, we offer a lightweight and scalable mechanism to assess the credibility of news sources—one that can complement both manual annotation pipelines and content-based classifiers.

2.2.4 Language models for trustworthiness estimation

A more recent and rapidly evolving direction explores the use of Large Language Models (LLMs) to perform source evaluations directly. Several studies have proposed sets of criteria to distinguish reliable from unreliable news sources. These criteria have been developed by (i) independent organizations, such as the [192] (GDI) and [136]; (ii) platforms like Facebook, through initiatives such as "[69]"; (iii) academic literature [101,

217]; and (iv) collaborative projects like [193] and [68].

These criteria vary in complexity and intended users. For instance, Facebook's guidelines are designed for lay users and offer simple tips to identify misinformation, whereas the frameworks proposed by GDI and NewsGuard are intended for expert annotators, such as professional journalists. Applying such expert-level criteria often requires domain-specific knowledge and nuanced judgment, which can be challenging for non-expert users [16]. Moreover, verifying subtle indicators—like political bias or context omission—may be particularly difficult for untrained raters.

At the same time, there is increasing evidence that both professional and crowd-sourced evaluations, at either the article or source level, can help reduce the spread of low-quality information [38, 113, 161]. Importantly, ratings based on expert criteria tend to have a stronger effect in reducing the intention to share unreliable content [38]. However, manual evaluations—especially those conducted at the granularity and consistency required by expert guidelines—pose major scalability challenges. This is particularly true for source-level ratings, which often rely on aggregated assessments of article-level dimensions [137, 192], and must be continuously updated to reflect new content.

Given these challenges, the automation of credibility assessments — particularly at the source level — has gained significant interest. Recent approaches include using social interaction signals to infer credibility without analyzing content [157], or classifying the reliability of individual articles to estimate source trustworthiness [17]. Another promising approach leverages LLMs to directly evaluate news publishers, either by analyzing article content or responding to structured prompts about outlets.

A representative example of this last direction is the work by Yang and Menczer [208], which prompts ChatGPT-3.5 to rate the credibility of over 7,000 news outlets using only their names as input. The outputs, on a 0–1 trustworthiness scale, are then compared with ratings from NewsGuard, MBFC, and human annotators [121]. Their findings show a moderate but statistically significant correlation between LLM-generated

and human expert ratings (Spearman's $\rho = 0.54$, $p < 0.001$), suggesting that LLMs can approximate human-level assessments using high-level input. However, such approaches often lack criterion-level granularity. For example, Yang and Menczer's prompt is intentionally minimalistic, eliciting a single trustworthiness score without probing specific dimensions of reliability. In contrast, our work [158] takes a more fine-grained approach by prompting the LLM to assess six distinct trustworthiness dimensions—derived from the expert criteria proposed by NewsGuard and GDI—across hundreds of news articles. Each LLM-generated response is then compared with human expert annotations. Our contributions (see Chapter 5) allows us to explore the alignment between LLMs and expert evaluations not just at the global trust score level, but also across specific, interpretable criteria. It also supports the identification of which dimensions are more amenable to automation, and which may still require human oversight. By doing so, our study contributes a novel, scalable framework for LLM-assisted evaluation of news content and creates a fine-grained, annotated dataset that can be used in future research. To the best of our knowledge, no prior work has proposed or tested a method for using LLMs to operationalize a full set of expert-designed credibility criteria in a structured and interpretable fashion. While some organizations like GDI (see [138]) use LLMs to flag content that matches disinformation narratives (through phrase-matching and filtering), and [139] employs AI tools in proprietary detection pipelines, these approaches either lack transparency or do not involve end-to-end automation of expert evaluation processes. Our contribution, therefore, addresses a key gap in the literature: providing a principled and replicable approach to automating detailed news reliability assessments using LLMs.

2.2.5 Challenges and open issues

Despite recent advances, the field of automated trustworthiness evaluation faces several persistent challenges and open gaps:

- **Lack of standardized criteria:** There is no universally accepted set

of evaluation criteria, which hinders replicability and cross-study comparisons.

- **Scarcity of benchmark datasets:** Ground truth datasets for trustworthiness evaluation are limited, often fragmented, and rarely aligned across studies, making systematic evaluation difficult.
- **Scalability of expert annotation:** While expert judgment remains the benchmark for evaluating content quality, it is not scalable to the vast and dynamic nature of online information.
- **Challenges in LLM-based approaches:** LLM-based methods show promise but are still experimental. Their effectiveness is sensitive to prompt design, and results may be influenced by model-specific training data. Moreover, transparency and interpretability remain open concerns.

As emphasized by Kozyreva et al.[116], automating source-level evaluations represents a crucial intervention to mitigate the spread of low-quality content online. Nevertheless, this remains a largely underexplored area, requiring further methodological advancements, the establishment of evaluation standards, and large-scale validation. In this context, the present thesis contributes to addressing these gaps through the studies presented in Chapters 5 and 6.

2.3 Methods

2.3.1 Entropy-based null models for complex network analysis

Recent advances in the analysis of online social networks have increasingly leveraged tools from network science to distinguish non-trivial patterns of social interaction from random noise. Among these, entropy-based null models have gained significant attention for their ability to provide unbiased benchmarks for network analysis. A comprehensive overview of these models is provided in the review by Cimini et al. [43].

Entropy-based methods have been applied in a wide range of contexts, including:

- the identification of discursive communities on Twitter based on retweet activity [9, 33, 34, 127, 159, 165, 166];
- the detection of non-trivial flows of disinformation [33, 45, 46, 157];
- the uncovering of coordinated behavior among automated Twitter accounts [27, 34];
- the analysis of information diets on Facebook [93];
- the identification of dominant subjects in public discourse [127, 165, 166].

The core idea behind entropy-based null models is to construct a maximally random network ensemble that still preserves certain observed topological features of the real system. In this way, they offer a model that is both general and tailored to the empirical network, enabling the detection of statistically significant deviations from randomness.

Formal definition of the null model. Let G^* be a real-world network and let $\vec{C}(G^*)$ denote a set of topological constraints observed in G^*. The goal is to define a null model: an ensemble of networks $\mathcal{G}$ that:

1. contains all possible graphs with the same number of nodes as G^*;
2. is maximally random except for preserving the constraints $\vec{C}$ on average.

To achieve this, one maximizes the Shannon entropy over the ensemble:

$$S = -\sum_{G \in \mathcal{G}} P(G) \ln P(G),$$

subject to the constraint that the expected value of $\vec{C}$ over the ensemble equals the observed value: $\langle \vec{C} \rangle_{\mathcal{G}} = \vec{C}(G^*)$.

This optimization leads to an exponential random graph model of the form:

$$P(G) \sim e^{-\vec{\theta}\cdot\vec{C}(G)},$$

where $\vec{\theta}$ is a vector of Lagrange multipliers determined by maximizing the likelihood of observing G^* [84, 110, 143, 187].

A fast and efficient implementation of these entropy-based models is provided by the Python library `NEMtropy`, available at: `https://pypi.org/project/NEMtropy/`.

In this thesis, we build on the methodology introduced in [43], and particularly make use of:

- the Bipartite Configuration Model (BiCM) [173], used for inferring similarity networks among verified users;
- its extension to *directed* bipartite networks [119];
- the general projection techniques for bipartite networks [171], applicable to both directed and undirected cases.

The entropy-based null models are central to the method for detecting discursive communities applied in this thesis (described in Section 2.3.2). In particular, the BiCM is applied to retweet networks to identify statistically significant clusters of verified users, which are then used to infer ideological groupings and support the analysis of online discourse.

Bipartite Configuration Model (BiCM)

Let us consider a bipartite network $\mathbf{G}^*_{\text{Bi}}$, in which the two layers are L and Γ. Define $\mathcal{G}_{\text{Bi}}$ the ensemble of all possible graphs with the same number of nodes per layer as in $\mathbf{G}^*_{\text{Bi}}$. It is possible to define the entropy related to the ensemble as [143]:

$$S = -\sum_{\mathbf{G}_{\text{Bi}}\in\mathcal{G}_{\text{Bi}}} P(\mathbf{G}_{\text{Bi}}) \ln P(\mathbf{G}_{\text{Bi}}), \tag{2.1}$$

where $P(\mathbf{G}_{\text{Bi}})$ is the probability associated to the instance $\mathbf{G}_{\text{Bi}}$. Now we want to obtain the maximum entropy configuration, constraining some

relevant topological information regarding the system. For the bipartite representation of verified and unverified user, a crucial ingredient is the degree sequence, since it is a proxy of the number of interactions (i.e. tweets and retweets) with the other class of accounts. Thus in the present manuscript we focus on the degree sequence. Let us then maximise the entropy (2.1), constraining the average over the ensemble of the degree sequence. It can be shown, [171], that the probability distribution over the ensemble is

$$P(\mathbf{G}_{\text{Bi}}) = \prod_{i,\alpha} (p_{i\alpha})^{m_{i\alpha}} \, (1 - p_{i\alpha})^{1 - m_{i\alpha}} \,, \tag{2.2}$$

where $m_{i\alpha}$ represent the entries of the biadjacency matrix describing the bipartite network under consideration and $p_{i\alpha}$ is the probability of observing a link between the nodes $i \in L$ and $\alpha \in \Gamma$. The probability $p_{i\alpha}$ can be expressed in terms of the Lagrangian multipliers x and y for nodes on L and Γ layers, respectively, as

$$p_{i\alpha} = \frac{x_i \, y_\alpha}{1 + x_i \, y_\alpha}. \tag{2.3}$$

In order to obtain the values of x and y that maximize the likelihood to observe the real network, we need to impose the following conditions [84, 187]

$$\begin{cases} \langle k_i \rangle = \sum_{\alpha \in \Gamma} p_{i\alpha} = k_i^* & \forall i \in L \\ \langle k_\alpha \rangle = \sum_{i \in L} p_{i\alpha} = k_\alpha^* & \forall \alpha \in \Gamma. \end{cases}, \tag{2.4}$$

where the $*$ indicates quantities measured on the real network.

In case of sparse network the formula (2.3) can be safely approximated with the Chung-Lu configuration model, i.e.

$$p_{i\alpha} \simeq x_i y_\alpha = \frac{k_i^* k_\alpha^*}{m},$$

where m is the total number of links in the bipartite network.

Bipartite Directed Configuration Model (BiDCM)

In the present subsection we will consider the case of the extension of the BiCM to *direct* bipartite networks and highlight the peculiarities of the network under analysis in this representation. The adjancency matrix describing a direct bipartite network of layers L and Γ has a peculiar block structure, once nodes are order by layer membership (here the nodes on L layer first):

$$\mathbf{A} = \left(\begin{array}{c|c} \mathbf{O} & \mathbf{M} \\ \hline \mathbf{N}^{\mathrm{T}} & \mathbf{O} \end{array} \right), \tag{2.5}$$

where the $\mathbf{O}$ blocks represent null matrices (indeed they describe links connecting nodes inside the same layer: by construction they are exactly zero) and $\mathbf{M}$ and $\mathbf{N}$ are non zero blocks, describing links connecting nodes on layer L with those on layer Γ and viceversa. In general $\mathbf{M} \neq \mathbf{N}$, otherwise the network is not distinguishable from an undirected one.

We can perform the same machinery of the section above, but for the extension of the degree sequence to a directed degree sequence, i.e. considering the in- and out-degrees for nodes on the layer L,

$$k_i^{\mathrm{out}} = \sum_{\alpha \in \Gamma} m_{i\alpha} \quad \text{and} \quad k_i^{\mathrm{in}} = \sum_{\alpha \in \Gamma} n_{i\alpha} \tag{2.6}$$

(here $m_{i\alpha}$ and $n_{i\alpha}$ represent respectively the entry of matrices $\mathbf{M}$ and $\mathbf{N}$) and for nodes on the layer Γ,

$$k_\alpha^{\mathrm{out}} = \sum_{i \in L} n_{i\alpha} \quad \text{and} \quad k_\alpha^{\mathrm{in}} = \sum_{i \in L} m_{i\alpha}. \tag{2.7}$$

The definition of the Bipartite *Directed* Configuration Model (BiDCM, [119]), i.e. the extension of the BiCM above, follows closely the same steps described in the previous subsection. Interestingly enough, the probabilities relative to the presence of links from L to Γ are independent on the probabilities relative to the presence of links from Γ to L. If $q_{i\alpha}$ is the probability of observing a link from node i to node α and $q'_{i\alpha}$ the probability of observing a link in the opposite direction, we have

$$q_{i\alpha} = \frac{x_i^{\mathrm{out}} y_\alpha^{\mathrm{in}}}{1 + x_i^{\mathrm{out}} y_\alpha^{\mathrm{in}}} \quad \text{and} \quad q'_{i\alpha} = \frac{x_i^{\mathrm{in}} y_\alpha^{\mathrm{out}}}{1 + x_i^{\mathrm{in}} y_\alpha^{\mathrm{out}}}, \tag{2.8}$$

where x_i^{out} and x_i^{in} are the Lagrangian multipliers relative to the node $i \in L$, respectively for the out- and the in-degrees, and y_α^{out} and y_α^{in} are the analogous for $\alpha \in \Gamma$.

In the present application we have some simplifications: the bipartite directed network representation describes users (on one layer) writing and retweeting posts (on the other layer). If users are on the layer L and posts on the opposite one and $m_{i\alpha}$ represents the user i writing the post α, then $k_\alpha^{\text{in}} = 1 \, \forall \alpha \in \Gamma$, since each message cannot have more than an author. Notice that, since our constraints are conserved on average, we are considering, in the ensemble of all possible realisations, even instances in which $k_\alpha^{\text{in}} > 1$ or $k_\alpha^{\text{in}} = 0$, or, otherwise stated, non physical; nevertheless the average is constrained to the right value, i.e. 1. The fact that k_α^{in} is the same for every α allows for a great simplification of the probability per link on $\mathbf{M}$:

$$q_{i\alpha} = \frac{(k_i^{\text{out}})^*}{N_\Gamma}, \tag{2.9}$$

where N_Γ is the total number of nodes on the Γ layer. The simplification in (2.9) is extremely helpful in the projected validation of the bipartite directed network [9].

Validation of the projected network

The information contained in a bipartite -directed or undirected- network, can be projected onto one of the two layers. The rationale is to obtain a monopartite network encoding the non trivial interactions among the two layers of the original bipartite network. The method is pretty general, once we have a null model in which probabilities per link are independent, as it is the case of both BiCM and BiDCM [172]. The method is graphically depicted in Fig. 1 in the case of BiCM; the case of BiDCM is analogous.

The first step is represented by the definition of a bipartite motif that may capture the non trivial similarity (in the case of an undirected bipartite network) or flux of information (in the case of a directed bipartite network). This quantity can be captured by the number of $V-$motifs

between users i and j [62, 173],

$$V_{ij} = \sum_{\alpha \in \Gamma} m_{i\alpha} m_{j\alpha}, \tag{2.10}$$

or by its direct extension

$$\mathcal{V}_{ij} = \sum_{\alpha \in \Gamma} m_{i\alpha} n_{\alpha j} \tag{2.11}$$

(note that $\mathcal{V}_{ij} \neq \mathcal{V}_{ji}$). We compare the abundance of these motifs with the null models defined above: all motifs that cannot be explained by the null model, i.e. whose p-value are statistically significance, are validated into the projection on one of the layers [171].

In order to assess the statistically significance of the observed motifs, we calculate the distribution associated to the various motifs. For instance, the expected value for the number of V-motifs connecting i and j in an undirected bipartite network is

$$\langle V_{ij} \rangle = \sum_{\alpha \in \Gamma} p_{i\alpha}\, p_{j\alpha}, \tag{2.12}$$

where $p_{i\alpha}$s are the probability of the BiCM. Analogously,

$$\langle \mathcal{V}_{ij} \rangle = \sum_{p \in P} q_{i\alpha}\, q'_{j\alpha} = \frac{(k_i^{\text{out}})^* (k_j^{\text{in}})^*}{N_\Gamma}, \tag{2.13}$$

where in the last step we use the simplification of (2.9) [9].

In both the direct and the undirect case, the distribution of the V-motifs or of the directed extensions is Poisson Binomial one, i.e. a binomial distribution in which each event shows a different probability. In the present case, due to the sparsity of the analysed networks, we can safely approximate the Poisson-Binomial distribution with a Poisson one [102].

In order to state the statistical significance of the observed value, we calculate the related p-values according to the relative null-models. Once we have a p-value for every detected V-motif, the related statistical significance can be established through the False Discovery Rate (*FDR*) procedure [12], which, respect to other multiple test hypothesis, controls the

number of False Positives. In our case, all rejected hypotheses identify the amount of V-motifs that cannot be explained only by the ingredients of the null model and thus carry non trivial information regarding the systems. In this sense, the validated projected network includes a link for every rejected hypothesis, connecting the nodes involved in the related motifs.

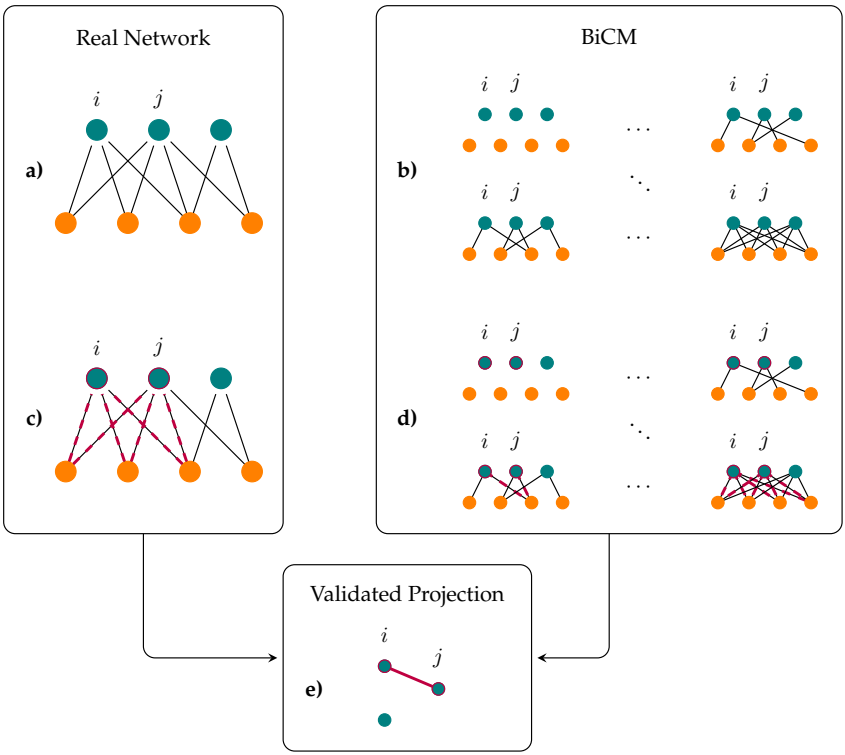


Figure 1: Schematic representation of the projection procedure for bipartite undirected networks. (a) An example of a real bipartite network. In the application, the two layers represent verified (turquoise) and unverified (orange) users. A link connects them if a retweet occurs. (b) The Bipartite Configuration Model (BiCM) ensemble includes all possible link realizations while preserving node degree constraints. (c) Two verified users (i and j) are shown with their common neighbors highlighted in magenta. (d) The overlap is tested against the BiCM null model for statistical significance. (e) If significant, a validated link is added between i and j in the projected network. Adapted from [34].

2.3.2 Entropy-based discursive community detection

In this subsection, we outline the methodology adopted in this thesis — particularly in Chapter 3 and Chapter 4 — for the detection of *discursive communities* (DiCos), i.e., groups of users contributing to the formation of a shared discourse. The procedure depicted in Figure 2 aims to assign community labels to social media accounts—particularly Twitter users—based on patterns of retweet interactions, with the overarching goal of inferring ideological alignment from collective engagement behavior.

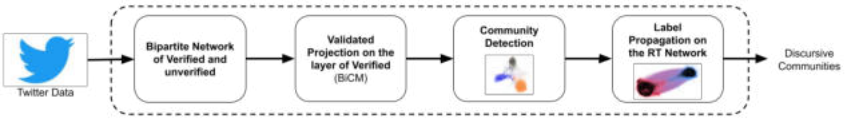


Figure 2: Entropy-based procedure for discursive community detection

The method leverages the structural properties of the retweet network, focusing on the interplay between *verified* and *unverified* users. Verified accounts, whose identity has been certified by the platform (e.g., journalists, politicians, public figures), are central to this analysis. These users are typically strong content producers and play a dominant role in shaping the discourse [9, 33, 127, 165].

Bipartite modeling and projection. The retweet network is represented as a bipartite graph, where one layer ($\top$) consists of verified users and the other ($\bot$) of unverified users. Edges are drawn from an unverified to a verified user if the former retweeted the latter. The underlying intuition is that verified users sharing a similar audience (i.e., retweeted by the same unverified users) are likely to convey similar narratives or ideological stances.

To quantify these similarities, we project the bipartite network onto the layer of verified users by counting the number of shared retweeters for each pair. However, raw co-retweet counts can be misleading, as they may simply reflect the popularity of certain accounts or the high activity

of certain users.

Entropy-based validation with BiCM. To address this, we employ an entropy-based null model—the *Bipartite Configuration Model* (BiCM) [43, 173]—which preserves the degree sequences of both layers (i.e., the number of retweets made by each unverified user and the number of retweets received by each verified user), while randomizing all other structural aspects. This model serves as a benchmark to assess whether the observed overlap in retweeters between two verified users is statistically significant or expected by chance.

If the number of shared retweeters for a given pair exceeds the BiCM-based expectation, a link is validated between the two verified users. The result is a monopartite, undirected, and unweighted network of verified users, where links represent statistically significant audience overlap.

Community detection and label propagation. We then apply the Louvain algorithm [19] to this validated network to extract communities of verified users—interpreted as ideologically coherent subgroups. Such an algorithm, despite being one of the most effective and popular, is also known to be order dependent [78]. To get rid of this bias, we apply it iteratively N times (N being the number of the nodes), after reshuffling the order of the nodes. Finally, we select the partition with the highest modularity.

Then, since verified users often have known identities and affiliations, these groups can be manually characterized post hoc (e.g., left-leaning vs. right-leaning, institutional vs. activist).

To expand the community labels to unverified users, we employ the label propagation algorithm of Raghavan, Albert, and Kumara [167] on the full retweet network. This step allows us to assign each unverified user to one of the previously identified verified-user communities, based on their retweet behavior.

Discursive communities. The final output of the procedure is a comprehensive partition of users into *discursive communities*. These are groups

of both verified and unverified users who engage in similar patterns of content sharing and are inferred to participate in the same segment of the public discourse. Crucially, this method does not rely on textual or semantic features but instead exploits structural signals emerging from retweet interactions.

Several works [9, 34, 45, 159, 165] have shown that this methodology provides a robust and scalable framework to study opinion dynamics and ideological polarization in online environments.

The importance of leveraging a validated projection for DiCo detection is also highlighted by our experiments (see Appendix A.2).

2.3.3 The NewsGuard Approach

The studies presented in this Thesis—particularly those in Chapter 3 and Chapter 4—rely on the credibility assessments provided by NewsGuard. Given the large-scale nature of our analyses we adopt a source-based approach: instead of evaluating individual news articles in isolation, we assess the reliability of the publishers responsible for producing them. Each news outlet is categorized according to its credibility and transparency, based on evaluations conducted by NewsGuard[7]. These assessments are performed by professional journalists who systematically review and annotate news domains using a standardized set of criteria.

This source-centric methodology is grounded in the premise that the intent and editorial practices of a publisher are central to determining the overall reliability of the information it disseminates [117]. Moreover, the impracticality of evaluating every article individually further justifies this approach, as it enables comprehensive and scalable analyses of information ecosystems [33, 180].

NewsGuard's framework evaluates each domain based on two main dimensions:

- **Credibility**, which includes criteria such as whether the outlet repeatedly publishes false information, fails to distinguish between news and opinion, or neglects to correct factual errors;

[7] `https://www.newsguardtech.com/`

- **Transparency**, which assesses whether ownership, editorial leadership, and article authorship are clearly disclosed, and whether advertisements are properly labeled[8].

Table 1 presents the labels used in this thesis to categorize domains based on NewsGuard's assessments[9]. In most cases, we limit our analysis to domains categorized as news sources. Exceptions include the study presented in [33], where a manual annotation process extended the categorization to include domains related to social networks (S), fundraising or petition platforms (F), marketplaces (M), political party outlets (P), institutional sites (IS), streaming platforms (ST), and search engines (SE).

In general, our focus is on quantifying the reputation of domains producing news content. Therefore, we exclude non-news entities such as platforms (P) and satirical sources (SA). The tags `T` and `N` represent trustworthy and untrustworthy news domains, respectively. These typically include outlets such as newspapers, magazines, television stations, and radio broadcasters.

Table 1: Tags for Domain Reputation Labeling According to NewsGuard

Label	Description
T	Trustworthy news domain
N	Untrustworthy news domain
P	Platform (e.g., `reddit.com`, `twitter.com`)
SA	Satirical source
UNC	Unclassified

NewsGuard: Source Reliability Assessment

In our research, NewsGuard's scoring system plays a central role in evaluating the reliability of news sources. Each outlet is assigned a *composite score ranging from 0 to 100*, based on performance across *nine journalistic*

[8] A detailed description of NewsGuard's evaluation process and criteria is available at: `https://www.newsguardtech.com/ratings/rating-process-criteria/`

[9] Throughout this work, a domain refers to the second-level domain (e.g., `nytimes.com`), excluding the top-level domain such as `.com` or `.org`.

criteria. A score of **60 or higher** is considered indicative of a trustworthy source.

In addition to the overall score, NewsGuard provides a detailed breakdown that shows which criteria were met or failed. This offers transparency into the evaluation process and allows researchers and readers to better understand the rationale behind each rating.

The NewsGuard Process

1. A trained NewsGuard analyst evaluates the website's content against nine criteria[10] (see Section C.2 Appendix C for the complete list of criteria, along with their respective weights).

2. Based on this assessment, the analyst drafts a written *"Nutrition Label"*, which includes:

    - A visual grid showing the site's performance on each criterion.
    - A written explanation describing the site's content, ownership, and the reasoning behind the rating.

3. If the site fails one or more criteria, NewsGuard contacts the publisher for comment. Any response is included in the Nutrition Label to represent the site's perspective.

4. The draft rating is reviewed and fact-checked by senior editors.

5. Ratings and Nutrition Labels are periodically updated to reflect changes in the site's practices or content.

All evaluations are conducted by trained journalists and editors, who apply consistent professional standards to assess a site's *credibility* and *transparency* practices.

[10] For a complete and up-to-date description of NewsGuard's rating process and criteria, visit `https://www.newsguardtech.com/ratings/rating-process-criteria/`. The description provided in this thesis reflects the process as outlined on the NewsGuard website at the time of our study. It is worth noting that the process or its description may have changed since then.

Chapter 3

The spread of Disinformation Dynamics on Twitter: Case Studies and Analysis

In this chapter, we present the first core contribution of the thesis: an empirical investigation of the dynamics underlying the spread of disinformation on social media, with a focus on Twitter/X. The analysis is grounded in case studies from Italy and the United States, covering key socio-political contexts such as the COVID-19 pandemic and the 2020 U.S. presidential elections.

We examine how the diffusion of low-credibility content is shaped by three main factors: (i) users' political affiliations, (ii) the presence of automated accounts (bots), and (iii) structural aspects of electoral systems. Adopting a multidisciplinary approach that combines network science and computational methods, we aim to uncover the mechanisms that drive the amplification of disinformation across different information ecosystems.

To this end, we employ network-based techniques to identify and characterize discursive communities, i.e., groups of users engaged in

shared political or ideological narratives, and analyze their role in exposure to and engagement with unreliable content. This is complemented by an analysis of user behavior, distinguishing between human and automated accounts, and by a comparative perspective on different electoral settings, such as swing versus safe states in the U.S.

The chapter is organized around two case studies. The first, focused on Italy during the COVID-19 pandemic (Section 3.1), highlights a strong political polarization of the debate, with low-credibility content predominantly shared within right-leaning communities. The second, centered on the U.S. electoral context (Section 3.2), shows that disinformation activity is concentrated in swing states and largely driven by Republican-affiliated users and bots.

Overall, these results provide a unified view of how individual-level attributes and systemic factors jointly shape the spread of disinformation, laying the groundwork for the methodological contributions developed in the following chapters.

3.1 Online disinformation during the peak of COVID-19 in Italy

3.1.1 Problem formulation and contributions

The COVID-19 pandemic has impacted on every human activity and, because of the urgency of finding the proper responses to such an unprecedented emergency, it generated a diffused societal debate. The online version of this discussion was not exempted by the presence of misinformation campaigns, but, differently from what already witnessed in other debates, the COVID-19 -intentional or not- flow of false information put at severe risk the public health, possibly reducing the efficacy of government countermeasures (see Section 2.1.1 for a detailed literature review).

The primary *research question* addressed in this section is focused on analyzing the *effective* impact of misinformation on the Italian societal debate on Twitter during the pandemic. This analysis specifically focuses on the role of various discursive communities. To this end, we

distinguish between the discursive communities of verified and unverified users, characterizing each group separately (refer to Section 2.3.2 for further details).

Main results We observe that the COVID-19 discussion on Twitter shows a clear division across groups that appear to be associated with different political orientations.We filter the network of retweets from random noise and check the presence of messages displaying URLs. By using NewsGuard's data (see Section 2.3.3), we assess the trustworthiness of the most recurrent news sites, among those tweeted by the political groups. The impact of untrustworthy posts reaches the 22.1% in the right and center-right wing community and its contribution is even stronger in absolute numbers, due to the activity of this group: 96% of all untrustworthy URLs shared by political groups come from this community.

Interestingly enough, we find that the main discursive communities are political, i.e., they involve politicians, political parties and journalists supporting a specific political ideal. While, at first sight, this may sound surprising - the pandemic debate was more on a scientific than on a political ground, at least in the very first phase of its abrupt diffusion -, it might be due to pre-existing *echo chambers* [76].

We then consider the news sources shared among the accounts of the various groups. Through a hybrid annotation approach, based on the judgments of independent journalists and annotation carried out by members of our team, we categorise such sources as trustworthy or untrustworthy (in terms of credibility of published news and transparency of editorial policies).

Finally, we extract the effective flow of content shared within the network: still following the approach of Ref. [9, 34], we extend the entropy-based methodology to a directed bipartite network of users and posts. In this sense, we are able to control not only the authorship activity and the retweeting attitude of the various accounts, but even the *virality* of the different messages, i.e., how many times a single message is shared.

The various political groups display different online behaviours. In

particular, the right wing community is more numerous and more active, even relatively to the number of accounts involved, than the other communities. Surprisingly enough, newly formed political parties, as the one of the former Italian Prime Minister Matteo Renzi, quickly imposed their presence on Twitter with a strong activity. Furthermore, the different political parties use different sources for getting information on the spreading on the pandemic. Notably, we experience that right and center-right wing accounts spread information from untrustworthy sources with a frequency almost 10 times higher than that of the other political groups. Due to their outstanding activity, their impact, in terms of number of d/misinforming posts in the debate, is much greater than that of any other group.

Contributions The present analysis contributes to understanding misinformation dynamics within the context of societal debates on Twitter during the COVID-19 pandemic, specifically within the Italian context.

- Analysis of Discursive Communities: The study delineates discursive communities on Twitter, focusing on both verified and unverified users, providing insights into how these communities interact and form based on shared information and retweets.

- Political Dynamics in COVID-19 Discussions: The analysis reveals a strong political division within the COVID-19 discourse, contrary to the expectation of purely scientific discussions. This suggests that the pandemic debate is significantly influenced by existing political affiliations and echo chambers. Specifically, the study highlights that the main discursive communities are not only scientific but also distinctly political, involving politicians, political parties, and journalists aligned with specific ideologies.

- Assessment of Information Trustworthiness and links with the political orientation: By utilizing the NewsGuard data (see Section 2.3.3), the research assesses the trustworthiness of frequently tweeted news sites by different political groups. It finds that the right and

center-right wing communities share a disproportionately high amount of information from low-quality (untrustworthy) sources, with these groups responsible for 96% of all untrustworthy URLs shared by political groups.

We previously discussed related work on the spread of online disinformation during the COVID-19 pandemic (see Section 2.1.1). In the following sections, we investigate, focusing on the Twitter dataset presented in Section 1.5.1, the dissemination of low-credibility information within major discursive communities. The analysis is divided into four parts. We begin in Section 3.1.2, which focuses on the identification and characterization of the main *Discursive Communities of Verified Users*. This step involves analyzing the structure of the retweet network to reveal coherent clusters of verified accounts, and understanding the type of content shared within each group. Next, in Section 3.1.3, we conduct a *Domain Analysis of Verified Users*, evaluating the credibility of the sources cited by users in these communities. In Section 3.1.4, we shift our attention to the *Validated Retweet Network*, examining the statistically significant (i.e., non-random) flows of information between users. This is followed by a more targeted investigation where we assess the credibility of domains shared through these effective flows. Finally, Section 3.1.5 focuses specifically on the *Propagation of Untrustworthy Domains* within the effective retweet network, allowing us to pinpoint how and where misinformation circulated most intensively.

3.1.2 Discoursive communities of verified users

Many studies in the field of social network analysis show that users are highly clustered per similar opinions [2, 48–50, 58, 59, 164, 219, 220]. Following the example of references [9, 34], we leverage this users' clustering in order to detect discursive communities, i.e., account groups interacting between each other by retweeting on the same (Covid-related) subjects according to the procedure described in Section 2.3.2. Remarkably, our methodology does not consider the shared texts, being focused on the retweeting activity among users only.

The top panel of Figure 3 shows the resulting network. Hereafter, a network resulting from the projection procedure will be called *validated* network[1]. The network presents a strong community structure, composed by four main subgraphs. When analysing them, we find that they correspond to

1. Media and right/center-right wing parties (in steel blue);
2. Center-left wing (in dark red);
3. Movimento 5 Stelle (*5 Stars Movement*, or M5S; in dark orange);
4. Institutional accounts (in sky blue).

Details about the political situation in Italy during the period of data collection can be found in the Appendix B.1.

While the various groups display a quite evident homophily among their elements, we further examined them by re-running the Louvain algorithm inside each of them, with the same care as above for the node order.

Since the subcommunities structure is extremely rich, we invite the interested reader to consult Appendix Section B.2 for a more detailed description. Hereafter, we will focus on purely political subcommunities, highlighted in the right panel of Figure 3. Starting from the center-left wing, we can find a darker red community, including the main politicians of the Italian Democratic Party (*Partito Democratico*, or PD), its representatives in the European Parliament and some EU commissioners. The magenta group is instead mostly composed by the representatives of Italia Viva, a new party founded by the former Italian Prime Minister Matteo Renzi (December 2014 - February 2016).

In turn, also the dark orange (M5S) community shows the presence of a purely political subcommunity (in orange in the bottom panel of Figure 3), which contains the accounts of M5S politicians, parliament representatives and ministers. Also, we can find some journalists of *Il Fatto Quotidiano*, a newspaper supporting M5S.

[1]The term *validated* should not be confused with the term *verified*, which instead denotes a Twitter user who has passed the formal Twitter verification procedure.

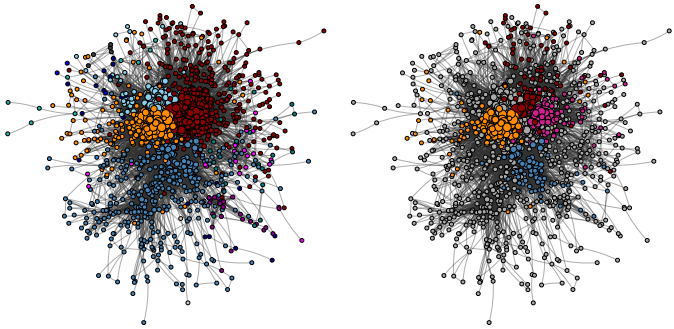


Figure 3: Discursive communities of verified users: ON THE LEFT COMMUNITIES They have been found running the Louvain community detection algorithm on the Largest Connected Component (LCC) of the validated network of verified users. Top panel: In red, top right corner, there are the center-left wing parties; in sky blue (on top), there are the official government accounts; in orange, the M5S-oriented community and in steel blue (on the bottom) the news media and center-right and right wing communities. Other minor communities can be found in the periphery of the LCC. Actually, by rerunning the same community detection algorithm inside these larger communities, it is possible to find *purely* political subcommunities, i.e., communities composed quite exclusively by politicians and official accounts of political parties. This can be seen in the lowest panel: in magenta, Italia Viva, the political party of the former Prime Minister Matteo Renzi; in red, the Partito Democratico, i.e., the Italian Democratic Party; in orange, M5S and in blue the center-right and right wing parties Forza Italia, Lega and Fratelli d'Italia. A more detailed description of the subcommunities of the network can be found in Section 2 of the Supplementary Material. In both panels, the node dimensions are proportional to their degree. The layout used for network visualization is the Fruchterman-Reingold one [81]

Concerning the steel blue community, the purely political subcommunity of center-right and right wing parties (as Forza Italia, Lega and Fratelli d'Italia, from now on FI-L-FdI) is represented in blue in the bottom panel of Figure 3.

Finally, the sky blue community is mainly composed by Italian embassies around the world.

We would to remark that, in Ref. [34], the authors exploited simi-

lar techniques to analyse the Italian debate on Twitter about migration policies. As in the current paper, after cleaning the system from random noise, the authors highlighted a group of coordinated accounts - a *bot squad*- increasing the visibility of a group of human-operated accounts. The division in community resembles the one found here, with few differences. First, in [34], media and center-right/right wing parties appeared in different communities from the very beginning; this is probably due to the fact that, in the present case, the criticism regarding the management of the pandemic by the main leaders of these parties were promptly reported by media. Secondly, in [34], M5S was not distinguishable from the the right/center-right wing discursive community. This is not so surprising, since, at time of data collection of the previous manuscript, M5S was allied with Lega, the main right wing party in Italy. The data showed that M5S and Lega shared the same views on migration policies. In the present work, however, because Lega was no longer governing the country at time of data collection, and probably because of the difference in topics covered (immigration policies *versus* pandemic), M5S manifests its individuality.

Due to the different subject and to the different political scenario (at the moment of the data collection, M5S, PD and Italia Viva are at the government: the interested reader can find more information about the Italian political situation in Appendix B.1), in the present analysis M5S manifests its individuality.

3.1.3 Analysis of domains - verified users

Here, we report a series of analyses related to the domains that mostly appear in the tweets of the validated network of verified users. The domains have been tagged according to their degree of credibility and transparency, as indicated by the NewsGuard[2] (see Section 2.3.3 for further details about methods and notations).

Unlike in our previous analyses, in this study, we chose to use a more detailed classification of domains compared to the one provided by

[2] `https://www.newsguardtech.com/`

NewsGuard. Specifically, to capture nuances within trustworthiness categories, we classify sources scoring between 55 and 65 as quasi-trustworthy ($\sim$T), expanding upon the tags listed in Table 1[3]. Furthermore, some domains in our dataset were not evaluated by Newsguard, prompting our team to assess them using a subset of Newsguard's criteria. Classification discrepancies were resolved through majority voting among annotators, achieving moderate agreement as measured by Fleiss' kappa ($\kappa = 0.63$). The details of this procedure are reported below.

As a first step, we considered the network of verified accounts, whose communities and subcommunities have been shown in Figure 3. On this topology, we labelled all domains that had been shared at least 20 times in tweets and retweets.

Table 2 gives statistics about number and kind of tweets, the number of url and distinct url (dist url), the number of domains and users in the validated network of verified users.

A url maintains here its standard definition[4] and an example is `http://www.example.com/index.html`.

Figure 4 shows, on the left panel, the absolute value of Trustworthy (T), Quasi Trustworthy ($\sim$T), and untrustworthy (N) shared domains per political subcommunity. On the right panel, we can see a similar plot, but the results are given in terms of percentages. As reported in Section 2.3.3 we recall that in this analysis, with 'Others', we denote all domains that do not refer to news sites, e.g., social networking sites, marketplaces, crowd sourcing platforms, etc. Others include also the UNC class, i.e., that of domains appearing less than 20 times in the posts of the validated network of verified users. Indeed, there are many domains that occur only few times; for example, there are 300 domains that appear in the posts only once.

At first glance, the majority of the news domains belong to the Trustworthy category.

Broadly speaking, we now examine the contribution of the different

[3]Hence, only for this analysis trustworthy (T) sources are those with a score that goes from 66 to 100, and untrustworthy sources those that reach a NewsGuard score between 0 and 54

[4]`https://en.wikipedia.org/wiki/URL`

type	#post	#url	#dist url	#domain	#user
tw	46277	37095	32605	1168	1115
rt	17190	9796	7504	1178	1385

Table 2: Posts, urls, domains and users statistics in the validated network of verified users. "Tw" represent pure tweets, while "rt" indicates retweets. The number of tweets sharing an url is much higher than the one of retweets and it is a known results for verified users, from which they appear to drive the online debate.

political parties, as represented on Twitter, to the spread of d/misinformation and propaganda.

Figure 4 clearly shows how the vast majority of the news coming from sources considered few or non trustworthy are shared by the center-right/right wing subcommunity (*FI-L-FdI*). Notably, the percentage of untrustworthy sources shared by the FI-L-FdI accounts is more than 30 times the second community in the N ratio ranking. The impact of N sources is even greater in absolute numbers, due to a major sharing activity of the users in this group (more than twice the value of the M5S subcommunity). Table 29 in Appendix B gives more details on the annotation results.

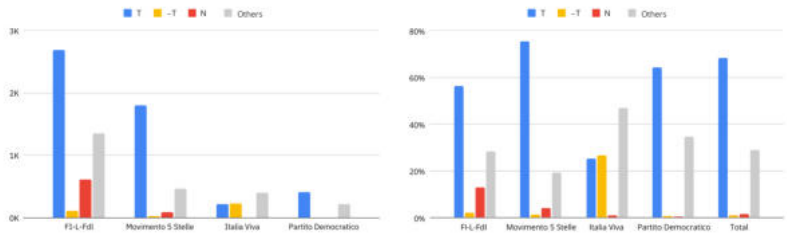


Figure 4: Number (left panel) and percentage (right panel) of Trustworthy (T), Circa Trustworthy (∼T), and untrustworthy (N) news sites shared by the political subcommunities – Validated network of verified users.

Looking at Table 3, some peculiar behaviours can still be observed. Again, the center-right/right wing parties, while being the least represented ones in terms of users, are much more active than the other

Subcommunity	#post	#url	#dist url	#domain	#user
only tweets					
FI-L-FdI	5031	4177	3728	210	62
Movimento 5 Stelle	2406	1839	1742	139	103
Italia Viva	943	458	417	96	69
Partito Democratico	736	370	353	74	60
only retweets					
FI-L-FdI	1587	582	510	151	72
Movimento 5 Stelle	997	546	469	104	103
Italia Viva	1048	399	348	147	82
Partito Democratico	747	273	258	94	88

Table 3: Posts, urls, domains and users statistics per political subcommunities – validated network of verified users: #post is the number of posts (divided in tweets and retweets), #url is the number of shared links, #dist url is the number of distinct urls, #domain is the number of distinct domains contained in all urls. While the number of (validated) verified users in the center-right/right wing subcommunity is lower than any other political group, their activity in writing original posts is at least twice greater than any other group. This difference is not present in the number of retweets.

groups: each (verified) user is responsible, on average, of almost 77.86 messages, while the average is 23.96, 22.12 and 15.29 for M5S, IV and PD, respectively. It is worth noticing that IV, while being a recently founded party, is very active. Finally, the Supplementary Material reports an analysis of the hashtags used by the political subcommunities, in order to study the focus of the narratives within the various political groups.

3.1.4 The validated retweet network

Here, we examine the *effective* retweet network, composed by users that retweet as a reaction to an interesting original tweet. As for effectiveness, we mean to consider the non random flow of messages from user to user. Indeed, it may happen that one tweet is shared either because it is viral, or because the retweeter is particularly active. Also, it could be that the account publishing the original tweet is extremely prolific. Instead, we are interested in the flow that cannot be explained only by the activity of

users or by the popularity of the specific posts. Otherwise stated, our aim is to highlight the non-trivial sharing activity, distinguishing the relevant information from the random noise. We thus define a *directed* bipartite network in which one layer is composed by accounts and the other one by tweets. An arrow connecting a user u to a tweet t represents u writing the message t. An arrow in the opposite direction means that u is retweeting t. To filter out the random noise from this network, we make use of the directed version of the BiCM, i.e., the Bipartite Directed Configuration Model (*BiDCM* [119]), described in Subsection 2.3.1. BiDCM constrains the in- and out-degree sequences of nodes on both layers. In our scenario, these represent the users' tweeting and retweeting activ-

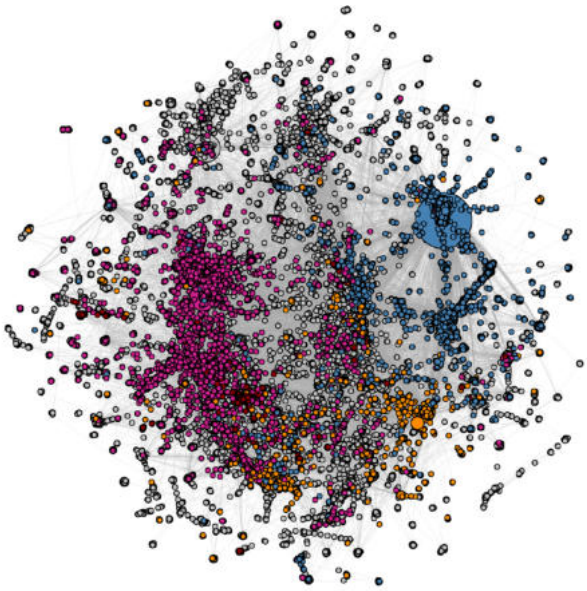


Figure 5: The directed validated projection of the retweet activity network: the communities have been highlighted according to the political discursive groups they take part to. All nodes not belonging to political discursive communities are in grey. Nodes' dimensions are proportional to their out degree. The layout used for network visualization is the Distributed Recursive (Graph) Layout [126].

ity and the virality of posts. In order to detect the non trivial flow of messages from user to user, for every (directed) couple of accounts, we compare the number of retweets observed in the real system with the expectation of the null model. If the amount of retweets cannot be explained by the theoretical model, we project a link from the author to the retweeter in the monopartite directed network of users. Due to the process of validation, we call this network *directed validated projection*. More details about the methodology can be found in Subsection 2.3.1.

The affiliation of unverified users to the various discursive communities is inferred exploiting the labels associated to verified users (see Subsection 3.1.2). The labels are propagated on the validated retweet network using the algorithm proposed in [167].

In Appendix B.5 we show that propagating labels on the entire weighted retweet network, on its binary version or on the validated version is almost equivalent in order to get the labels for the users in the directed validated network.

After applying the label propagation, we obtain the political communities in the validated retweet network, as shown in Figure 5. We can see that the whole scenario changes dramatically with respect to the one of verified users. The center-right/right wing community is the most represented community in the whole network, with 11,063 users (representing 21.1% of all the users in the validated network), followed by Italia Viva users with 8,035 accounts (15.4% of all the accounts in the validated network). The impact of M5S and PD is much more limited, with, respectively, 3,286 and 564 accounts. It is worth noting that this result is unexpected, due to the recent formation of Italia Viva.

As in a previous study targeting the migration debate [34], the most effective users in terms of hub score [115] are almost exclusively from the center-right/right wing parties. Considering, e.g., the first 100 hubs, only 4 are not from these groups. Interestingly, 3 out of these 4 are verified users: Roberto Burioni, a popular Italian virologist, ranking 32nd; Agenzia Ansa, an Italian news agency, ranking 61st; and Tgcom24, the newscast of a private TV channel, ranking 73rd. The fourth account is an

online news website, ranking 88th: this is an unverified account which belongs to a non political community.

Further, 3 in the top 5 hubs were already found in [34]. In particular, a journalist of a neo-fascist online newspaper (unverified user), an extreme right activist (unverified user) and the leader of Fratelli d'Italia, Giorgia Meloni (verified user), who ranks 3rd in the hub score. Matteo Salvini (verified user), who was the first hub in [34], ranks 9th, surpassed by his party partner Claudio Borghi (verified user), ranking 6th. The first hub in the present network is an (unverified) extreme right activist, posting videos against African migrants and accusing them to be responsible of the contagion and of violating lockdown measures.

Domain analysis on the directed validated network

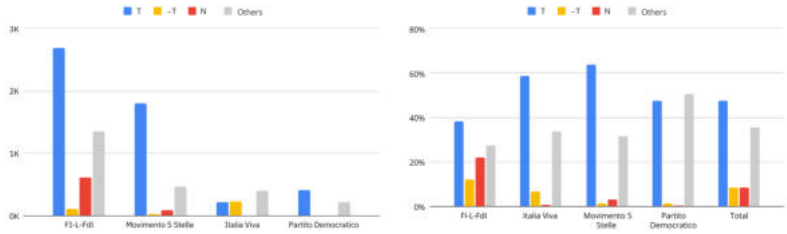


Figure 6: Number (left panel) and percentage (right panel) of Trustworthy (T), Circa Trustworthy (∼T), and untrustworthy (N) news sites shared by the political subcommunities – Directed validated network.

Figure 6 shows the annotation results of all the domains tweeted and retweeted by users in the directed validated network. The annotation was made considering the domains occurring at least 100 times. Similarly to the approach outlined in Section 3.1.3, the same three members of our team have annotated sites that have not yet been evaluated by NewsGuard. We have 100 domains annotated by Newsguard and 53 domains annotated by the three annotators. Also in this case, the annotators showed a moderate agreement for the classification of domains, with $\kappa = 0.57$.

With respect to the annotation results for the network of verified users, the majority of URLs referring to news sources is still considered trustworthy, but its incidence is much reduced. Interestingly enough, the impact of at least nearly trustworthy sources is almost 19% for tweets and 16% for retweets, against percentages around 3% and 2% for the network of verified users.

The incidence of untrustworthy source in the subcommunity of center-right/right wing parties reaches the impressive percentage of 22.1%, which is even greater than what observed in Fig. 4 (i.e., 12.8%). The contribution of unverified users seems to boost the diffusion of unreliable content. It is even more alarming that the percentage of quasi trustworthy source is great too: considering both untrustworthy and quasi trustworthy sources the percentage is 34.2%. Thus, more than one third of the URLs shared in the validated network by FI-L-FdI is at least quasi trustworthy.

In absolute numbers, FI-L-FdI shares the highest number of N URLs, being responsible of the 96% of N URLs shared by all the political subcommunities. This behaviour is not only due to the the greater amount of users: in the FI-L-FdI subcommunity, the accounts sharing N URLs are particularly active. In this group, the average number of (original) N posts sent per user is 32.21, which is almost 6 times the average for the M5S users (which has 5.38 N posts per users); IV and PD have 4.48 and 1.00 as average, respectively. The frequency of accounts retweeting N sources among all users from the same community is extremely high also for FI-L-FdI (57.6% for FI-L-FdI, 23.5% for M5S, 5.79% for IV and 2.5% for PD).

Table 4 reports statistics about posts, urls, distinct urls, users and verified users in the political subcommunities in the directed validated network. Noticeably, by comparing these numbers with those of Table 3, reporting analogous statistics about the validated network of verified users, we can see that now the number of retweets is much higher than that of tweets, and the opposite holds for verified user. Verified users

Community	#post	#url	#dist url	#domain	#user	#verif
only tweets						
FI-L-FdI	176137	95902	63710	3272	6831	56
Italia Viva	82356	33648	25364	2243	4976	56
Movimento 5 Stelle	41838	22940	17747	1536	1974	92
Partito Democratico	3247	1759	1671	277	337	51
only retweets						
FI-L-FdI	959748	361844	54768	4304	10749	48
Italia Viva	379096	121477	37084	3915	7827	52
Movimento 5 Stelle	208195	97304	27692	2647	3135	72
Partito Democratico	11517	4424	3079	683	528	44

Table 4: Posts, urls, domains and users statistics per political subcommunities – directed validated network. Differently from the case of verified users, the number of tweets is nearly one fifth of the number of retweets.

tend to tweet more than retweet, while users in the directed validated network, which comprehends also unverified users, have a greater number of retweets, being even more than ~ 5 times the one of tweets (depending on the community). This behaviour was already observed in [9, 34] and it is essentially due to the preeminence of verified users in shaping the public debate on Twitter. It is also remarkable the fact that verified users represent a minority of all users in the directed validated network.

Fig. 7 shows the trend of the number of posts containing URLs over the period of data collection. The highest peak appears after the discovery of the first cases in Lombardy. This corresponds to more than 68,000 posts containing URLs, but a higher traffic is still present before the beginning of the Italian lockdown, while there is a settling down as the quarantine went on[5]. Interestingly, similar trends are present even in the analysis [42, 82].
It is interesting to note that the incidence of N sources is nearly constant in the entire period.

[5]The low peaks for February 27 and March 10 are due to an interruption in the data collection, caused by a connection breakdown.

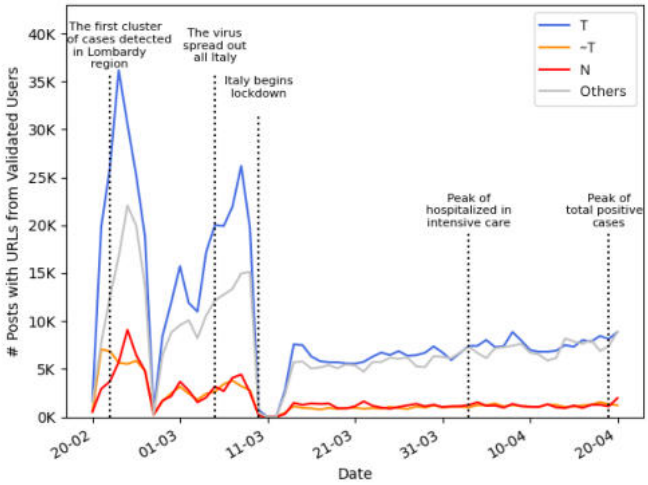


Figure 7: Domains' spreading over time – validated directed network The various main event regarding the pandemic have been reported in the plot. It is interesting to notice that the incidence of N sources in the entire period is more or less constant in time. Interestingly enough, the same reduction of the overall activity after the beginning of the lockdown was detected even in [42, 82].

3.1.5 Untrustworthy domains shared in the effective flow of misinformation

As a final task, over the whole set of tweets produced or shared by the users in the directed validated network, we counted the number of times a message containing a URL was shared by users belonging to different political subcommunities, although without considering the semantics of the tweets. Namely, we ignored whether the URLs were shared to support or to oppose the presented arguments.

Table 5 shows the most frequent (tweeted and retweeted) N domains shared by the political subcommunities; the number of occurrences is reported next to each domain.

The first untrustworthy (N) domains for FI-L-FdI refer to right, ex-

treme right and neo-fascist propaganda. It is the case of imolaoggi.it, ilprimatonazionale.it and voxnews.info, recognised as disinformation websites by NewsGuard and by the two main Italian debunker websites, bufale.net and BUTAC.it.

As shown in the table, some domains, although in different number of occurrences, are present under more than one column, thus shared by users close to different political areas. However, since the semantics of the posts in which these domains are present were not investigated, the retweets of the links by more than one political subcommunity could be due to contrast, and not to support, the opinions present in the original posts. here, we intend to just present the most frequent N domains.

FI-L-FdI		**Italia Viva**		**Movimento 5 Stelle**		**Partito Democratico**	
imolaoggi.it	16041	dagospia.com	315	lantidiplomatico.it	1114	it.sputniknews.com	2
ilprimatonazionale.it	15383	m.dagospia.com	134	m.dagospia.com	286	dagospia.com	2
voxnews.info	9334	imolaoggi.it	109	dagospia.com	266	laverita.info	1
stopcensura.info	8460	lantidiplomatico.it	72	it.sputniknews.com	98	lantidiplomatico.it	1
laverita.info	2647	ilprimatonazionale.it	61	imolaoggi.it	89	m.dagospia.com	1
stopcensura.org	2407	it.sputniknews.com	44	ilprimatonazionale.it	87	-	-
m.dagospia.com	2125	stopcensura.info	28	stopcensura.info	65	-	-
scenarieconomici.it	1647	agenpress.it	25	voxnews.info	46	-	-
it.sputniknews.com	1313	voxnews.info	25	agenpress.it	37	-	-
dagospia.com	1291	laverita.info	19	stopcensura.org	21	-	-
lantidiplomatico.it	1245	scenarieconomici.it	13	laverita.info	10	-	-
agenpress.it	1121	stopcensura.org	8	scenarieconomici.it	7	-	-
lavocedelpatriota.it	986	lavocedelpatriota.it	6	lavocedelpatriota.it	2	-	-

Table 5: List of the most frequent N domains, with relative occurrences, per political subcommunities. The count was made considering all posts for users of the direct validated network.

3.1.6 Discussion

Due to its impact on several dimensions of the society, the online debate regarding COVID-19 was the target of several early studies [39, 42, 44, 82, 103, 149, 170, 181, 212, 218]. In the present section, we examine the presence of misinformation campaigns in the Italian online societal debate about the pandemic, during its peak of the first wave (end of February, 2020 - end of April, 2020). Our analysis is based on a general methodology reviewed in [43, 188] in order to extract both the discursive communities and the effective flow of messages [9, 34]: in particular, we build an entropy-based null-model, constraining part of the information of the real system, and we compare the observations on the real network with

this benchmark.
The discursive communities are extracted focusing on verified users, i.e., public figures whose identity has been checked directly by Twitter. As in other studies [9, 34, 165], we observe that verified accounts lead the debate: their tweets are much more than their retweets. Due to such role, we examine in details the activity of verified users. Furthermore, we focus on the *effective* flow of information in the online debate: by comparing the system with an entropy-based null model, we filter out all the random noise associated to the activity of users and virality of tweets. We highlight all the non trivial retweeting activities and further examine the properties of the filtered network, focusing on the incidence of untrustworthy news sources shared in the debate.

Despite the fact that the results have been achieved for a specific country, we believe that our approach, being general and unbiased by construction, is extremely useful to highlight non trivial properties and peculiarities. In particular, when analyzing the outcome of our investigation, some features attracted our attention:

1. *Persistence of clusters w.r.t. different discussion topics:* In Caldarelli et al. [34], we focused on tweets concerned with immigration, an issue that has been central in the Italian political debate for years. In particular, using the same techniques here adopted to extract the effective retweet network, we highlighted the presence of coordinated automated accounts increasing *effectively* the visibility of users belonging to the same discursive community. In this paper, we discover that the clusters and the echo chambers that were detected when analysing tweets about immigration are almost the same as those singled out when considering discussions about COVID-19[6] This may seem surprising, because a discussion about the pandemic may not be exclusively political, but also medical, economic and social. We can thus argue that the clusters are political in na-

[6] Actually, in [34] the center-right/right wing parties were distinct from the Media community. Here, we found them distinct after launching, on the first community, a further community detection algorithm.

ture and, even when the topics change, users remain in their cluster on Twitter. (It is, in fact, well known that journalists and politicians use Twitter for spreading information and political propaganda, respectively).

The reasons why political polarisation affect so strongly the vision of what should be an objective phenomenon is still an intriguing question.

2. *(Dis)Similarities amongst offline and online behaviors of members and voters of parties:* Maybe less surprisingly, the political habits is also reflected in the degree of participation to the online discussions. In particular, among the parties in the center-left wing side, a small party (Italia Viva) shows a much more effective social presence than the larger party of the Italian center-left wing (Partito Democratico), which has many more active members and more parliamentary representation. More generally, there is a significant difference in social presence among the different political parties, and the amount of activity is not at all proportional to the size of the parties in terms of members and voters.

3. *Spread of untrustworthy news sources:* In the online debate about the pandemic, many links to untrustworthy news sources are posted and shared. Kind and occurrences of the domains vary with respect to the corresponding political subcommunity. Furthermore, the center-right/right wing discursive community is characterised by a relatively small number of verified users that corresponds to a very large number of acolytes which are (on their turn) very active, three times as much as the ones of the opposite communities in the partition. In particular, when considering the amount of retweets from poorly trustworthy news sites, this community is by far (one order of magnitude) much more active than the others. As noted already in our previous publication [34], this extra activity could be explained by a more skilled use of the systems of propaganda – in that case a massive use of bot accounts and a targeted activity against migrants (as resulted from the analysis of the hub list).

While our work contributes to the literature regarding the analysis of the impact of misinformation on the online societal debate, it paves the way to other crucial analyses. In particular, it would be interesting to analyse the structure of the retweet network and how it may contribute to increase the visibility of some of the influential accounts that we detected (this was, in part, the target of the analysis in [5]). In this sense, even the role of automated accounts for the diffusion of N news domains is of utmost importance in order to tackle the problem of online misinformation. Another relevant direction concerns the analysis of highly prolific users in the network, often referred to as *superspreaders*. In the present work, our analysis mainly focused on the characterization of the debate through patterns of interaction and homophily among users. However, identifying and analysing highly active users could provide additional insights into how certain actors contribute to the production and dissemination of information within the detected communities. In particular, studying the role of superspreaders could help clarify how visibility and engagement are concentrated around specific accounts, for instance within politically aligned groups, and how these dynamics may affect the circulation of both reliable and unreliable information in the debate.

3.2 Online disinformation in the 2020 U.S. Presidential election: swing vs. safe states

3.2.1 Problem formulation and contributions

The study presented in this Section of the Thesis, focuses on the Twitter debate during the week leading up to November 4, 2020. Like previous research (see Section 2.1.2), we examine the flow of disinformation and the infiltration of bots into this discourse. However, our work uniquely focuses on two specific aspects of the U.S. presidential election: the presence of swing and safe states and the winner-take-all system. Recent literature comparing online political debates across countries highlights how different electoral systems lead to different structural proper-

ties within online social networks [26, 151, 197, 199].

The term 'swing' refers to states where a landslide victory for either Republicans or Democrats is uncertain, owing to the lack of a clear voting orientation. In contrast, a state is deemed 'safe' when the electoral races are not competitive and are unlikely to be closely contested. Competitiveness is determined by several factors, including the political composition of the state and its counties, the prevailing local and national political climate, and insights from interviews with campaign experts[7]. Therefore, it is important to note that the status of swing and safe states is not fixed. Major swing states may become reliably safe Republican or Democratic states over time, while traditionally solid red or blue states may move into the swing state category. Changing demographics and political realignments within specific regions or demographic groups often drive these shifts[8].

With the exception of Maine and Nebraska, all U.S. states utilize the winner-take-all voting method. Each state has a varying number of presidential electors, determined in part by its population. Following a popular vote, each state allocates its presidential electors based on the candidate with the most votes, due to the winner-take-all system. A major criticism of this system is that it incentivizes presidential candidates to focus their campaigns on a select few swing states, as they hold the key to victory[9]. In particular, certain battleground states, such as Florida, traditionally a swing state with a substantial population and a large allocation of presidential electors[10], have been subjected to more intense electoral campaigns. Transferring this critique to the realm of Twitter, our paper poses and answers a central question: Could it be that the Twitter discourse leading up to the 2020 U.S. presidential election mirrors the electoral system, specifically the distinction between swing and safe states?

Specifically referring to disinformation flows,

[7] https://www.cookpolitical.com/ratings/presidential-race-ratings

[8] https://www.maynoothuniversity.ie/research/spotlight-research/10-swing-states-will-decide-us-presidential-election

[9] https://www.jstor.org/stable/j.ctt1npxbp

[10] https://edition.cnn.com/election/2020/results/state/florida

- Is there a difference in the frequency of tweets containing links to dubious or unreliable news when they are associated with swing states or safe states during the 2020 pre-election season? Is this difference in frequency also related to the political orientation of the account?

- Does the prevalence of automated accounts in online pre-election political debates differ depending on whether the discussion focuses on swing states or safe states? If so, is the difference also related to the political orientation of the account?

To perform the analysis, we collected Twitter data using keyword searches, specifically pairing candidate and state names (see Section 1.5.1). We then processed the data and the users who created and shared it as follows. First, we extracted links to news stories in the tweets and associated those stories with a level of trustworthiness. Second, we classified the users as bots or not. Third, we extracted the main discursive communities and their political orientation, which we used to (i) filter out irrelevant data from the entire dataset, specifically users who were not interested in the political narrative, and (ii) gain insight into the specific political leanings of the accounts.

Contributions Our main contributions are:

- We provide a fine-grained characterization of the Twitter traffic about the 2020 U.S. presidential election, in the week leading up to Election Day, adopting a multidisciplinary approach including complex network analysis, to identify non-trivial communities of users and their political leanings, artificial intelligence (to classify users as bots or not), and human-based annotation (to classify news sources as trustworthy or not).

- To the best of our knowledge, this is the first paper that investigates the links between the U.S. presidential electoral system and the online debate about the election, focusing on automated accounts, the diffusion of low-credible news, and employing a sophisticated

network-based approach to identify the specific political leanings of the users participating in the debate.

- We provide compelling evidence of a correlation between the actual electoral mechanism, which tends to prioritize intense campaigning in swing states, and the online electoral debate. Indeed, we observe that a significant portion of the 2020 election-related online traffic revolves around tweets focused on swing states. Furthermore, the discourse surrounding swing states exhibits a higher concentration of links leading to untrustworthy news sites. Importantly, most of the disinformation content associated with swing states (and Republican supporters) originates from automated accounts, indicating their significant role in spreading such content.

Main results The experiments conducted in this work led us to the following results:

- Tweets associated with swing states account for about 88% of the whole traffic. As a rough measure, the population of the swing states in the dataset represents 66% of the population of the states in our dataset. In this sense, the swing states have received more attention than would have been expected based solely on the number of electoral votes they represent.

- Two main user communities emerge from the data: a homogeneous one, consisting of Republican supporters (hereafter referred to as REP), and a mixed one, comprising journalists as well as both Republican and Democratic supporters (hereafter referred to as REP-DEM-JOURN).

- More than 90% of links to news from untrustworthy publishers are concentrated in the REP community. Each of these links is shared an average of 57 times, a significantly higher number than the average number of shares in the REP-DEM-JOURN community (7).

- The percentage of tweets with URLs pointing to news from untrustworthy publishers is consistently higher for swing states in all

communities.

- Tweets associated with safe states have a higher concentration of URLs pointing to news with trustworthy publishers. Tweets associated with swing states have a higher concentration of URLs pointing to news with untrustworthy publishers.

- Of the total number of tweets associated with swing states and containing untrustworthy URLs, 74% of these are posted or retweeted by accounts classified as bots.

Originality This analysis is neither the first nor the last to address the impact of real-world events on virtual ones, and vice versa. See a review on the relationships between electoral politics and social media in Section 2.1.2 of this thesis. The work of Howard et al. in [104] examines tweets from authors who left some evidence of their physical location in the period leading up to the 2016 U.S. presidential election. The analysis reveals a high concentration of polarized news in tweets associated to swing states with a significant number of presidential electors.

In addition to differences in years (2016 versus 2020) and differences in data collection methods (hashtags versus general keywords), the present study differs from the work of Howard et al. [104] in some important ways. First, our analysis includes an evaluation of automated accounts, and the classification of news sources is based on the annotations of expert journalists.

However, the primary distinction of our study lies in the rigorous filtering process applied to our dataset. This process employs advanced statistical methods specifically tailored for the analysis of complex networks, making them well suited for the study of interactions within social networks. For a complete understanding of these methodologies, the reader is referred to Sections 2.3.1 and 2.3.2. Using this filtering process allows us to gain insight into the political affiliations of users participating in these discussions (i.e., which political party users tend to be more closely associated with).

3.2.2 The selection of swing and safe states

The states were selected based on measures and indications provided in reports by experienced political analysts in the months leading up to the 2020 elections[11]. We chose a balanced list of states, four safe states and four swing states. For the safe states, we chose two pairs that were balanced in terms of political leanings and presidential electors. We took Washington and New Jersey from the solid Democratic states and Indiana and Louisiana from the solid Republican states. This results in 26 electoral votes for the Democratic candidate and 19 for the Republican. For the selection of the swing states, we took the three most important states from the point of view of presidential electors: Florida (29 votes), Pennsylvania (20 votes), and Michigan (16 votes); we also added Arizona (11 votes) because it has been of particular interest in the election debates[12,13]. We should further clarify that our choice was not driven by any "formal" definition based on statistics relative to the results of previous elections, but by the indications of political analysts (especially those in [11]). For example, Arizona gave its electoral votes to the Republicans in the 2000, 2004, 2008, 2012, and 2016 elections. However, the consensus among political analysts, based on various polls, was that Arizona was no longer a safe state for Republicans. In fact, Arizona gave its electoral vote to Biden in 2020. With that in mind, we considered Arizona a swing state, even though the historical data would have placed it in the safe set.

The data was further processed to (i) identify user communities with a vested interest in the political narrative through our filtering process, (ii) classify link domains using NewsGuard, and (iii) map each tweet to its corresponding state type (i.e., swing or safe) using a content-based approach.

The procedure for filtering the data set is described in Section 2.3.2. From here on, we will refer to the product of the filtering procedure as the

[11] https://www.cookpolitical.com/analysis/national/national-politics/latest-cook-political-report-electoral-college-map

[12] https://fivethirtyeight.com/features/how-arizona-became-a-swing-state/

[13] https://www.washingtonpost.com/politics/2022/09/16/senate-control-midterm-elections-2022/

'validated dataset' (to distinguish it from the original dataset). For both the verified and unverified accounts that pass the filtering procedure, we also collect the bot scores via BotometerLite.

For URL classification, we rely on NewsGuard[14] (see Section 2.3.3), which provides a set of $\{domain\_name, tag\}$ pairs (tags are in Table 1). It was therefore necessary to translate all the short-form URLs contained in the text of the tweets, so that we could have the domain names in clear.

We use a content-based approach to establish the association between each tweet and the state type (i.e., swing or safe). In practice, we first check each tweet - or retweet - for the presence of at least one state name from the selected list (e.g., Arizona, Florida, etc.). We then exclude any tweets that contain more than one state name (approximately 1.5 million tweets contain more than one state name). Consequently, each tweet in the resulting dataset contains only one state name, which can be swing or safe. Although it is true that posts discussing the states under analysis have been lost, we prefer such a conservative approach to eliminate possible noise from our data set. Furthermore, we only consider English tweets (non-English tweets number about 422,000). The resulting dataset contains about 3.3 million tweets and about 398,000 URLs (see Table 6).

Finally, we conclude this subsection by noting that our analysis aims to investigate whether the debate about swing states is meaningfully different from the debate about safe states. Therefore, we are not interested in checking the origin of the tweets, i.e. whether or not their authors are located in the states under analysis. In fact, even users outside the US can contribute to the debate and/or the level of disinformation in it.

3.2.3 Bot detection using Botometer

Botometer is one of the most widely used tools for bot detection in the literature [174, 200, 209, 211]. It employs a supervised machine learning approach based on Random Forest classifiers [24] to evaluate whether a Twitter account exhibits automated behavior.

In our studies, we primarily rely on version 4 of Botometer, which

[14] `https://www.newsguardtech.com/`

Table 6: Twitter's statistics by state. The asterisk '$*$' indicates swing states.

State	No. Tweets	No. URL
Arizona$*$	224046	34637
Florida$*$	744006	85373
Michigan$*$	734600	87529
Pennsylvania$*$	1209083	145067
New Jersey	38007	8114
Indiana	17185	988
Washington	342104	36254
Louisiana	6886	633
Total	3315917	398595

was the most recent release at the time of experimentation. Botometer v4 has demonstrated strong performance in detecting both individual bots and coordinated bot networks [174, 211]. Specifically, we make use of the premium lightweight version, *BotometerLite*[15], which is designed for efficient processing of historical data.

Unlike the standard Botometer version that requires direct interaction with the Twitter API, BotometerLite performs bot detection using only metadata from user profiles. It does not fetch or analyze tweets in real time but instead assesses user accounts based on available profile information. This design makes it particularly well-suited for analyzing archived datasets or accounts that are no longer active on the platform. BotometerLite supports batch processing of up to 100 users per request, with a limit of 200 requests per day, allowing a maximum of 20,000 accounts to be processed daily.

The tool returns a bot score $S \in [0, 1]$, where higher values indicate a greater likelihood of automated behavior. Importantly, the score does not represent a direct probability that an account is a bot. Instead, it should be interpreted in a relative sense—useful for comparing and ranking accounts within a specific population or dataset.

It is important to note that, at the time of writing this thesis, Botometer (now rebranded as *Botometer X*) has been placed into archival mode

[15] `https://cnets.indiana.edu/blog/2020/09/01/botometer-v4/`

due to changes in Twitter's ownership and API policies (see Section 1.5.1). As a result, Botometer X now returns precomputed bot scores based on data collected prior to May 31, 2023, and cannot evaluate accounts created after this cutoff date.

3.2.4 Detecting the presence of discursive communities

Table 7: Characteristics of the main discursive communities. The columns *No. Users*, *No. Tweets*, and *No. URL* report absolute values, while all other columns represent percentages. Specifically, *Tweets Safe* and *Tweets Swing* indicate the proportion of tweets related to safe and swing states, respectively, whereas *Left* and *Right* denote the percentage of URLs associated with left-leaning and right-leaning sources.

Community	No. Users	No. Tweets	Tweets Safe	Tweets Swing	No. URL	Left	Right
REP	269019	2083158	12.35	87.65	241488	0.59	**49.74**
REP-DEM-JOURN	213679	919949	10.39	89.61	92412	**16.18**	1.86
JOURN-1	197	1174	4.86	95.14	485	3.30	0.82
JOURN-2	53	404	10.64	89.36	74	22.97	2.70
OTHERS	218880	311232	16.44	83.56	64136	6.62	14.60
Dataset	701828	3315917	12.19	87.81	398595	5.18	32.92

We execute the procedure outlined in Section 2.3.2 to identify discursive communities, which are groups of Twitter accounts that actively contribute to the development of a shared discourse by retweeting among themselves; results are summarised in Table 7.

The giant component of the retweet network includes more than 4.8×10^5 accounts, while nearly 2.2×10^5 accounts belong to smaller clusters: the latter are not going to be analysed in the following since they are not relevant for the entire debate.

To characterize the community structure of the giant component, we conduct a manual analysis *a posteriori* of the various communities, leveraging the presence of verified accounts (as discussed in Section 2.3.2) i.e., authentic public interest accounts like politicians, journalists or VIPs.

To assign labels to the list of verified users within each community, we gave priority to users with higher node degrees, indicating a greater number of connections. The largest community within the giant component primarily consists of Republican supporters and comprises approxi-

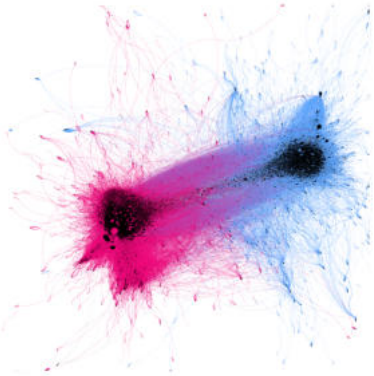
mately 2.7×10^5 users. Some examples of users within this community include '@TrumpWarRoom', '@TeamTrump', and '@TrumpStudents'. This community will be referred to as REP henceforth.

The second most populated community, with around 2.1×10^5 accounts, is a mixed one encompassing Republicans, Democrats, as well as various journals and journalists with diverse political orientations. Accordingly, it will be labeled as REP-DEM-JOURN.

Accounts in the REP and REP-DEM-JOURN communities are responsible for over 90% of the tweets in our dataset (see Table 7). While other communities do exist within the giant component, their size is practically negligible compared to the ones described above or they lack a clear political orientation. Therefore, they will not be considered in the subsequent analysis.

In particular, our analysis will focus only on the result of the entropy-based filtering procedure, i.e. the users belonging to the REP and REP-DEM-JOURN communities. Figure 8 shows these two main communities, which emerge after running the label propagation algorithm in [167] to the retweet network.

Figure 8: Retweet Network after label propagation (547k nodes, 1.8M edges).



The result of the analysis of news domains in the tweets of major

communities is a good indication of their correct labeling. The Left and Right columns in Table 7 represent the percentage of sources identified by NewsGuard as left-wing and right-wing oriented, respectively. Within the REP community, almost 50% of the shared URLs come from right-leaning sources. In the REP-DEM-JOURN community, the prevalence of left-leaning sources is significantly lower, approximately 16.2%. This measure can probably be explained by observing the mixed composition of the REP-DEM-JOURN community.

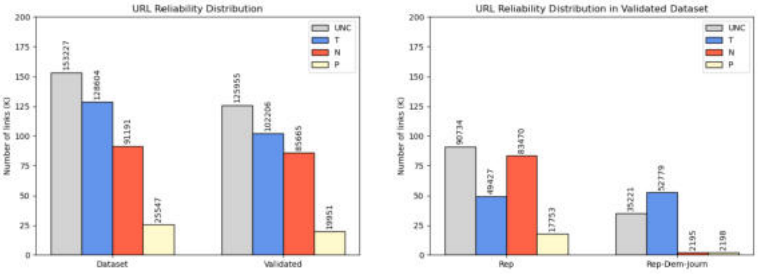
Manual verification of emerging communities We also conducted a manual analysis of a sample of users from the emerging communities to verify the correct composition of the latter. Specifically, we randomly selected a subset of 99 unverified accounts and created a balanced sample representative of emerging discursive communities, including Rep, Rep-Dem-Journ, and those users assigned to other communities or without any community association. We then manually annotated the Twitter users within the sample, taking into account (i) the content of the messages they write or retweet, (ii) the political orientation of the news publishers they share (using the NewsGuard labels), and (iii) the political orientation of well-known Twitter users they retweet.

When comparing the manual annotations with our labeling procedure, 89 out of 99 users showed consistent labels. However, 10 users who were not assigned to any community by our filtering procedure showed inconsistencies in the labels. A content analysis of these 10 users reveals an association with Republican and Democratic political visions (6 and 4 users, respectively). It is important to emphasize that these minor discrepancies do not affect our analysis, which focuses on users belonging to the main communities, i.e., REP-DEM-JOURN and REP.

3.2.5 Reputation of news domains

Figure 9 (left) shows the distribution of URLs found in tweets in the full and validated datasets, respectively. The URLs have been tagged according to the NewsGuard labels. The validation procedure discards $\sim 17\%$

Figure 9: Classification of links



of the tweets with URLs from the full dataset, which is about 64k tweets. Thus, most of the links are distributed within the political communities that emerge from the data.

Figure 9 (right) shows that, with respect to the entire dataset, 93% of untrustworthy links (N) are shared within the two main political communities. In particular, about 91% of the total is shared within the REP community. Furthermore, the links in the Rep community with publishers tagged as N by NewsGuard are mostly right-leaning (i.e., in terms of numbers we found Slightly Left 2, Far Left 6, Slightly Right 4831, Far Right 76161 links).

Figure 10: Distribution of the number of link sharing in REP (left) and REP-DEM-JOURN (right) (see Table1).

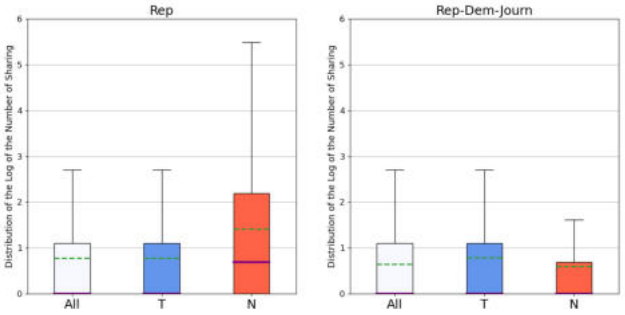


Figure 10 shows the virality of the links, that is, how many times the links in our dataset have been shared. We can see that in REP, links of type N are shared many more times than other types of links. Specifically, in REP, an N link is shared on average 57 times, while in REP-DEM-JOURN it is shared 7 times. These results suggest that untrustworthy links find fertile ground among Republican supporters.

3.2.6 Reputation of news domains in tweets associated to swing and safe states

Table 8: Summary statistics for swing and safe states across the validated dataset and discursive communities. Columns *No. Users*, *No. Tweets*, and *No. URL* report absolute counts, while *T* (trustworthy) and *N* (not trustworthy) represent percentages of URLs associated with each category.

States	No. Users	No. Tweets	No. URL	T	N
VALIDATED					
Swing	451840	**2649642**	299210	**28.30**	**26.06**
Safe	170644	**352863**	34586	**50.66**	**22.25**
REP					
Swing	251615	1825320	218565	18.66	34.72
Safe	112556	257236	22819	37.92	33.25
REP-DEM-JOURN					
Swing	200225	824322	80645	54.45	2.59
Safe	58088	95627	11767	75.37	0.91

Here, we analyze the flow of disinformation in tweets associated with swing or safe states and per discursive community. We recall that a tweet is associated with a state if the name of the state is present in the tweet text. By construction, each tweet in our dataset contains only one state name (see Section 3.2.2).

Table 8 gives statistics on the number of accounts, tweets, and URLs related to the kind of state associated with the tweets and to the two main political communities. We see that the vast majority of traffic is associated with tweets about swing states (about 88% of the total, see row VALIDATED, column No. Tweets). When looking at links pointing to un-

trustworthy news sites (N), the concentration for swing states - 26.06%. - is higher than for safe states - 22.25%. The concentration of trustworthy links (T) is higher for safe states - 50.66% *vs* 28.30% for swing states.

To statistically validate the frequencies of N and T links in tweets associated with swing and safe states, we performed a Pearson chi-square test [144]. The comparison between the distribution of T and N links in these tweets and the distribution observed in the validated dataset reveals a statistically significant difference, with a p-value below 10^{-65}.

At the community level, in agreement with the results in Section 3.2.5, we observe a higher concentration of links N in the REP community; however, we do not observe substantial differences in terms of percentage of links N between swing and safe states for both REP and REP-DEM-JOURN (Table 8, rightmost column). For both communities, the highest concentration of trustworthy links (T) is in tweets associated with safe states (column T).

Figure 11: Distribution of the number of link shared per kind of state in REP.

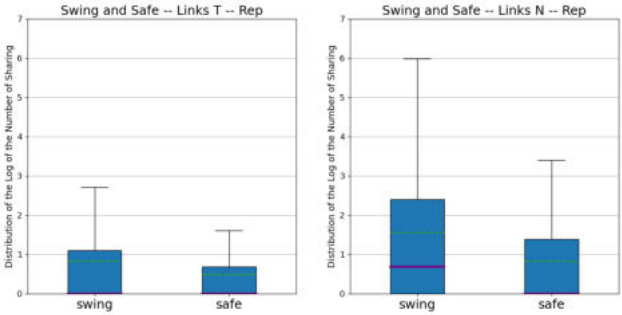


Figure 11 shows that in REP, untrustworthy links (N) are shared many more times on average in the debate associated with swing states. Specifically, the average number of shares is 66 times for *swing* and 22 times for *safe*. For trustworthy links (T), a similar but not as pronounced behavior is observed.

3.2.7 Social bots

In this section, we explore the relationship between disinformation flow and the characteristics of accounts in our dataset. We determine the bot scores of the accounts using BotomerLite, see Section 3.2.3. We recall the the bot score provides a measure of the extent to which an account exhibits bot-like characteristics, on a scale of 0 to 1. The closer the score is to 1, the more likely it is that the account is a bot.

We perform two different analyses. The first analysis aims to determine whether the bots in our dataset exhibit a discernible political orientation. Specifically, we seek to determine whether accounts within the REP community tend to exhibit more automated behavior than those in the REP-DEM-JOURN community. In our second analysis, we aim to examine two critical aspects: 1. the correlation between the type of accounts and their propensity to generate untrustworthy traffic; and 2. exploring potential correlations between automated accounts and traffic associated with swing states.

For the first analysis, we compare the bot score distributions in REP and REP-DEM-JOURN using the Mann-Whitney U [125] and Kolmogorov-Smirnov [184] statistical tests. Both tests are used to determine whether two distributions are different, and if so, in what way. The bot score distributions were created by keeping the bot score of the account that posted each tweet.

Figure 12 shows the distributions of bot scores with respect to total traffic (left) and traffic containing only URLs (right). Since the distributions associated with the two communities have relatively close means (for the total traffic: REP 0.26 and REP-DEM-JOURN 0.23; for URL traffic: REP 0.272 and REP-DEM-JOURN 0.238), we perform the Kolmogorov-Smirnov and Mann-Whitney U statistical tests to assess whether the distributions are statistically different.

The Kolmogorov-Smirnov (KS) test measures the distance between two empirical distributions as the maximum difference in their cumulative distributions. The p-values of the Kolmogorov-Smirnov tests (as shown in Table 9) indicate that the distributions of bot scores in the two

communities are significantly different from that of the entire dataset (p-values are less than 10^{-319}). The Mann-Whitney U (MWU) test evaluates the difference in location between the distributions. Our results, as shown in Table 10, confirm that the distributions are significantly different (again, all p-values are less than 10^{-309}). Furthermore, the values of the bot scores in the REP community are higher than those of the entire dataset, and significantly exceed the scores measured in the REP-DEM-JOURN community. These results suggest that tweets in the two communities are generated by users with different characteristics in terms of bot scores (Figure 12).

Figure 12: Bot scores distributions in both communities (left for each panel), REP (center) and REP-DEM-JOURN (right).

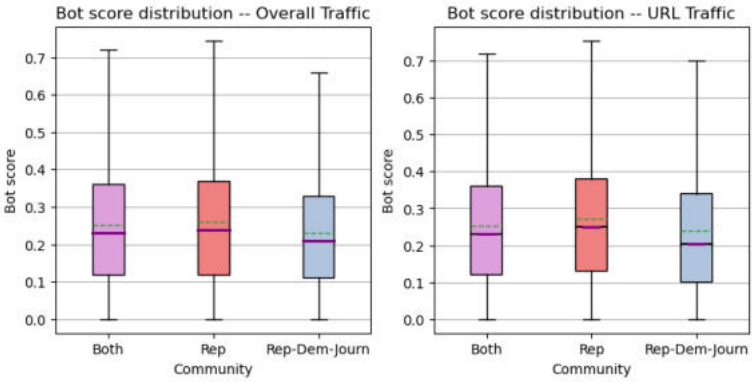


Table 9: Results of the Kolmogorov-Smirnov test about the bot scores distribution in the two main communities.

dist_A	dist_B	KS test (dist_A,dist_B)	p-value$_{KS}$
VALIDATED	REP	0.021	$< 10^{-319}$
VALIDATED	REP-DEM-JOURN	0.047	$< 10^{-319}$
REP-DEM-JOURN	REP	0.068	$< 10^{-319}$

The second analysis aims to detect untrustworthy tweets posted by bot accounts, and the potential correlation between bots and swing-related tweets. To identify which accounts are bots, we take the conservative approach used in [74]: we classify as bots those accounts 'that fall at the

Table 10: Results of the Mann-Whitney U test about the bot scores distribution in the two communities.

$dist_A$	$dist_B$	MWU test ($dist_A$,$dist_B$)	MWU test ($dist_B$,$dist_A$)	p-$value_{MWU}$
VALIDATED	REP	0.486	0.514	$< 10^{-309}$
VALIDATED	REP-DEM-JOURN	0.532	0.468	$< 10^{-309}$
REP-DEM-JOURN	REP	0.453	0.547	$< 10^{-309}$

upper end of the bot score distribution'. This approach has the dual benefit of preventing misclassification of accounts with borderline scores, while focusing on accounts with clear bot characteristics. In practice, we tag each account in the validated dataset using BotometerLite, sort them from lowest to highest bot score, and isolate those with bot scores in the first and last deciles. In the first decile we have real accounts, while in the last decile we have bot accounts. Specifically, the first decile contains accounts with bot scores in the range [0, 0.04], and the last decile contains accounts with bot scores in the range [0.45, 1]. We collect tweets from both real accounts and bots to investigate the source of untrustworthy traffic. We acknowledge that we exclude many accounts from our validated dataset by not including those with bot scores in the range [0.04, 0.45]. Nevertheless, this approach provides us with more reliable guarantees to minimize false positive and false negative predictions.

Table 11 shows statistics for classified accounts. Of the total number of classified accounts, 47.19% are bots. Considering only the REP community, this percentage increases to 54.93%, while in the REP-DEM-JOURN community it decreases to 37.35%. In terms of posting activity, bots appear to be more active than real accounts, being about twice as active in both posting tweets and sharing tweets with URLs. Of the total traffic generated by classified accounts, bots contribute 64.19% of the traffic, reaching 68.07% in the REP community and dropping to 53.94% in the REP-DEM-JOURN community.

Disinformation, bots, discursive communities, and swing states

Here we focus on the role of bots in spreading links to low-trustworthy/non-trustworthy news stories. Table 12 shows the percentages of (i) all, (ii) trustworthy (T), and (iii) non-trustworthy (N) URLs shared by users clas-

Table 11: Genuine and bot accounts in the validated dataset and in the main political communities.

Label	No. Users	No. Tweets	No. URL
VALIDATED DATASET			
human	57797	228378	25422
bot	51648	409449	53017
REP			
human	27624	147772	16156
bot	33663	315065	41383
REP-DEM-JOURN			
human	30173	80606	9266
bot	17985	94384	11634

sified as bots or real. The table also takes into account their membership in a discursive community (REP or REP-DEM-JOURN) and a state category (swing or safe).

Focusing on the *Swing & Safe* column in Table 12, we see that about 73% of the untrustworthy (N) traffic is generated by bots, regardless of the community, while bots are responsible for about $\sim 63\%$ of tweets with trustworthy URLs. If we focus only on the traffic generated in REP, the bots spread 73. 84% of the N and 69.11% of the T links. Focusing only on untrustworthy links, of the 91% of the total in swing states, more than 74% are posted or retweeted by bots. Furthermore, while untrustworthy links associated with safe states are only a small part of the total (8.47%), the vast majority of this traffic comes from bot accounts (68.75%).

3.2.8 Discussion

The study of disinformation within online social networks during election campaigns has yielded a wealth of contributions, exemplified by works such as Becatti et al. [9], Bovet and Makse [21], Budak et al. [28], Ferrara et al. [74], Georgacopoulos et al. [86], Luceri et al. [124], and Mattei et al. [127], to name a few. However, the spread of untrustworthy content has rarely been linked to the specifics of a particular electoral system; most existing disinformation studies tend to focus on individ-

Table 12: Distribution of shared links by reputability category, state type, and discursive community. The column "No. URL" reports the absolute number of URLs, while all other values are expressed as percentages. For each group, we report (i) the proportion of links shared in swing vs. safe states, and (ii) the percentage of links shared by bots and humans, both overall and disaggregated by state type.

				Swing & Safe		*Swing*		*Safe*	
Community	No. URL	swing	safe	bot	human	bot	human	bot	human
All Links									
VALIDATED DATASET	78439	89.92	10.08	**67.59**	32.41	**67.87**	32.13	**65.06**	34.94
REP	57539	90.78	9.22	**71.92**	28.08	72.26	27.74	68.63	31.37
REP-DEM-JOURN	20900	87.54	12.46	55.67	44.33	55.36	44.64	57.77	42.23
Trustworthy Links (T)									
VALIDATED DATASET	23036	83.07	16.93	**62.90**	37.10	62.69	37.31	63.96	36.04
REP	11812	83.46	16.54	**69.11**	30.89	68.84	31.16	70.47	29.53
REP-DEM-JOURN	11224	82.65	17.35	56.37	43.63	56.15	43.85	57.42	42.58
Non-trustworthy Links (N)									
VALIDATED DATASET	20627	**91.53**	**8.47**	**73.69**	26.31	**74.15**	25.85	**68.75**	31.25
REP	20147	91.42	8.58	**73.84**	26.16	74.33	25.67	68.59	31.41
REP-DEM-JOURN	480	96.25	3.75	67.50	32.50	66.88	33.12	83.33	16.67

ual countries. Yet emerging evidence suggests that the electoral process plays a role in shaping the dynamics of online discourse. Limited findings to date [26, 104, 151, 197, 199] suggest that there are differences in how online accounts organize themselves in discussions, either promoting divisive or cohesive structures, depending on whether a country uses majoritarian, proportional, or plurality voting systems.

In our current research, while still focusing on a single country, we direct our attention to two specific aspects: (i) a feature of its presidential electoral system-the presence of swing and safe states-and (ii) whether and to what extent this feature is reflected in the spread of online disinformation.

To elaborate further, each U.S. state is allocated a certain number of presidential electors, and after the statewide popular vote, the faction that receives the highest number of votes claims all of the electors, regardless of the margin of victory. Safe states are those where election outcomes can be easily predicted, while swing states represent fiercely contested battlegrounds that are crucial to securing the presidential election.

With this context in mind, our analysis focuses on the 2020 U.S. presi-

dential election. We focus specifically on the Twitter discourse surrounding eight states, four of which are categorized as safe states (New Jersey, Indiana, Washington, and Louisiana) and the remaining four as swing states (Arizona, Florida, Michigan, and Pennsylvania). We then selected tweets that contained the names of the presidential candidates (either Biden or Trump) and the names of one of the selected states in their text.

Our first result is that 88% tweets in our dataset is related to swing states. This underlines the importance of swing states (as opposed to safe) in the political discussion.

Secondly, from Table 8 we observe that the frequency of untrustworthy URLs shared in the political debate of swing states (26.06) is greater than the analogous of safe states (22.25%). Symmetrically, the frequency of trustworthy URLs is higher in safe states (50.66%) than swing ones (28.30%). In this sense, not only the debate, but also the spread of disinformation is more intense in swing states due to their importance for the election outcome. To summarize, both the total flow of news and the frequency of untrustworthy URLs are higher in swing states.

Thirdly, we investigate the exposure to disinformation of the two main emergent discursive communities: a great community of Republican supporters (the REP community) and a mixed one, including both Democrats and Republicans, as well as various journalists (the REP-DEM-JOURN community). Remarkably, the REP community hosts 91% of the total URLs pointing to untrustworthy news sources. In addition, each untrustworthy URL in the REP community is shared, on average, more than any other type of URL.

Finally, we investigate the contribution of automated accounts in the spreading of disinformation. Let the reader consider Table 12: bots appear to be more active than genuine accounts in posting tweets, both in swing and in safe states, with comparable percentages, i.e. $\sim$67% vs. $\sim$65%, respectively in swing and safe states. Regarding the untrustworthy links shared in swing states, more than 74% are posted or retweeted by bots.

Our analyses were conducted through a careful filtering process applied to the original dataset. We used techniques rooted in Information

Theory and Statistical Mechanics principles related to complex networks, as discussed in Section 2.3.2, to elucidate political communities. In particular, we focused on the bipartite network representing retweet interactions between verified and unverified users.

To validate the projection of the bipartite network onto the verified user layer, we employed the BiCM (Bipartite Configuration Model) as a benchmark. This involved establishing links between verified users if the number of shared unverified retweeters was statistically significant. We then ran a community detection algorithm on the resulting network of verified users. To extend these communities to unverified Twitter users, we leveraged our knowledge of verified users and implemented a label propagation procedure. Our validation approach ensures that we account for interactions that cannot be attributed solely to user degree sequences, which distinguishes our work from similar studies such as that of Howard et al. [104], who analyzed disinformation flows in swing and safe states during the 2016 election but did not employ entropy-based null models.

In sum, our hypothesis that the spread of disinformation is more pronounced in swing states finds robust support in the data. Due to their pivotal role in determining election outcomes, swing states not only attract a higher volume of tweets, but also bear a greater percentage of the brunt of disinformation campaigns compared to safe states. This disparity in the impact of disinformation, coupled with the increased flow of messages, leads to a particularly worrisome spike in disinformation messages.

Limitations and future work While our findings provide compelling insights, it is important to acknowledge certain limitations that invite further investigation. First, our research was limited to a select number of U.S. swing and safe states, providing a specific snapshot of the broader electoral landscape. In addition, our analysis was limited to the 2020 U.S. presidential election. Expanding our study to include comparative analyses with the 2012 and 2016 elections could either validate our conclusions or contextualize them in the context of the 2020 contest.

We also focus on a subset of swing and safe states, rather than all of them. Our choice was a compromise between a number of practical limitations. First, we needed a dataset of manageable size, so we limited our analyses to a subset of swing states, focusing on the four largest. Second, we needed an appropriate benchmark, i.e., a sample of safe states against which to compare our measurements. Such a choice was more complicated because the number of safe Republican states tends to be much larger than the number of safe Democratic states, but they tend to be less populous and thus represent a smaller number of electoral votes. In this sense, a good compromise was to choose four safe states, i.e. the same number of swing states, equally divided between Democrats and Republicans, with almost the same number of electoral votes. In addition, it is important to note that even if the states we chose to study happen to exhibit a flow of misinformation that is significantly different from that of other solid states, we still have evidence of an increased focus by misinformation producers on states that are more likely to influence the final outcome of the national election.

Furthermore, our data collection methodology relied on keyword-based approaches that inherently lack a precise understanding of the exact content of the collected tweets. Although the presence of both state and candidate names in the tweets implies a connection to the election and the state, the specific content remains unknown until examined.

Extending our study to other plurality voting systems with similarities to the U.S., such as the United Kingdom, would provide valuable insights into the presence of analogous disinformation diffusion dynamics within swing constituencies. In addition, examining the presence of disinformation at the geographic level within different electoral systems, including proportional systems (e.g., Germany and Spain), majoritarian systems (e.g., France), or mixed systems (e.g., Italy, South Korea, and Japan), would further enrich our understanding of this phenomenon.

Another aspect that could be further investigated concerns the role of highly active users in the diffusion of information within the debate. As discussed in Section 3.1.6, identifying highly prolific accounts (often referred to as *superspreaders*) could provide additional insights into how

visibility and engagement are concentrated around specific users and how these actors may contribute to amplifying both reliable and unreliable information across different communities.

Finally, we argue that our research contributes to a more detailed examination of the relationship between electoral systems, online discourse, and the spread of online disinformation.

Chapter 4

Disinformation and echo chamber detection during the COVID-19 vaccination debate in Italy

This chapter introduces the first methodological component of the second core contribution of this thesis: a novel approach for detecting echo chambers in online social media. Based on a formal definition of echo chambers as groups of users who are both ideologically aligned and exposed to similar content, the method combines network science techniques with an entropy-based null model to jointly analyze user interactions and content exposure.

We apply this framework to the Italian COVID-19 vaccination debate on Twitter/X, a highly polarized context characterized by the spread of disinformation. The results reveal nine echo chambers composed of densely connected, politically homogeneous, and highly active users who frequently share low-credibility content. Although limited in size, these groups play a key role in amplifying disinformation and sustaining polarized narratives.

Overall, the proposed methodology enables large-scale detection of

echo chambers, provides insights into the mechanisms driving online polarization, and lays the groundwork for the methodological developments presented in Chapter 6.

4.1 Problem formulation and contributions

Although the concept of echo chambers is well-understood, the literature still lacks a universally accepted method for their detection on Online Social Networks (OSNs). A more detailed review of the literature on echo chamber detection in OSNs can be found in Section 2.1.3.

The methodology introduced in this chapter aims to bridge this gap by providing a general and unbiased approach to detect echo chambers solely through social interactions.

This approach starts by considering two key events indicative of echo chamber formation: (i) interaction between users with similar opinions, and (ii) exposure of users to the same news articles. The core idea is to model and verify the presence of these events using entropy-based null models as statistical benchmarks. Essentially, the presence of an echo chamber in OSNs is confirmed by detecting significant overlaps between these two events.

Practically, the process begins by identifying groups of users with *similar opinions* based on the significant similarity of the main content creators: we utilize the so-called Discursive Communities (DiCo) for this purpose (see Section 2.3.2). DiCos encompass users who contribute to the formation of a common discourse. Next, we identify groups of users engaged with the same news articles by proposing a novel methodology to extract what we term News Engagement Communities of users (users' NECs). A non-trivial overlap between these two types of communities (DiCos and users' NECs), along with evidence of existing connections between the users, indicates the presence of an echo chamber.

Contributions: The main contribution of this chapter is a novel unbiased method for echo chamber detection. The procedure is based on

the very definition of echo chambers and involves the application of an entropy-based null model to discard signals assimilated to noise.

Research questions: Keeping in mind that our ultimate goal is to observe if and when well-connected users belonging to discursive communities and news engagement communities overlap, thus forming echo chambers, we organize the structure of the chapter to answer the following research questions (RQs):

- RQ.1: What are the characteristics of the discursive communities (DiCos) and of the news engagement communities of users (users' NECs)? Are there well-connected users in common indicating the presence of *echo chambers*?;
- RQ.2: What is the relation between the emergent echo chambers and the presence of disinformation, if any?

To explore the two research questions, we use the Italian Twitter debate on the COVID-19 vaccination as a case study (see Section 1.5.1); we observed that while users within echo chambers constitute a small minority, they significantly influence the debate, often spreading misinformation. The platform considered in this study is Twitter/X.

4.2 Pipeline for echo-chamber detection

Starting from Online Social Network (OSN) data, specifically tweets, Figure 13 illustrates our pipeline to identify echo chambers. This pipeline comprises two parallel paths: the top path delineates the process for extracting what we term 'discursive communities' (DiCo), which are groups of users uniform in the opinions they support. Conversely, the bottom path focuses on detecting user News Engagement Communities (NECs), i.e., groups of users uniform in the narratives they engage with.

Then, the results from these two paths — the identified DiCos and user NECs — are compared to extract a common subset of users. These subsets shared at the same time (i) similar opinions and (ii) exposure

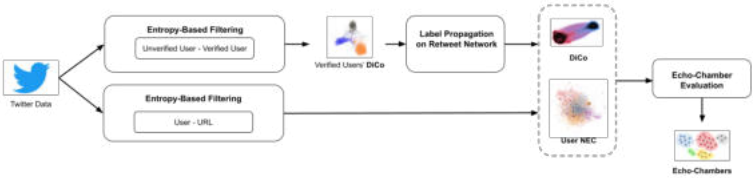


Figure 13: Pipeline for Echo-chamber detection. The upper path focuses on the detection of Discursive Communities (*DiCo*), while the lower one on the detection of News Engagement Communities (*NEC*). Both procedures pass through the statistical validation of empirical data with an entropy-based null model.

to the same news articles. The interactions/connections among users within this intersection are further analyzed to confirm the presence of an echo chamber. Specifically, we look for well-connected users who belong to the same DiCo and the same user NEC, as this combination indicates membership in the same echo chamber.

Further details regarding these methods will be elaborated upon in subsequent sections.

4.2.1 Detecting similar opinions

Assessing the opinions of various accounts is challenging; however, these opinions can be inferred from interactions among the accounts. Recent research highlights that the presence of a discursive community (DiCo) — a group of accounts contributing to a common discourse — can be deduced from users' retweet interactions[1].

In this study, we utilize DiCos to identify groups of accounts with similar opinions (see Figure 13 top path). The process begins by detecting groups of verified accounts perceived as similar, called Verified users' DiCo. These groups are characterized using a manual approach. Given that the identities and owners of verified accounts are known, the affiliations of their members (e.g., political, journalist) are identified. A

[1]A detailed description of the method designed to detect the discursive communities can be found in Section 2.3.2.

community-level label is assigned if a group's characteristics are uniform. For instance, if a verified user's DiCo exclusively comprises members of a specific right-wing political party, that community is accordingly labelled. Subsequently, this community-level label is extended to every unverified user who, during the discussion, demonstrates support for the content produced or spread by such a community. When a user is associated with the label of one Verified user's DiCo, it is concluded that the opinion of such user aligns with that group. This methodology enables the efficient inference of opinions across a large number of unverified accounts.

4.2.2 Detecting exposure to the same news articles

Regarding the exposure to the same news articles, we approach its assessment by analyzing the ties between the users and the URLs present in their tweets and retweets. The bottom path of Figure 13 shows the approach. The idea of leveraging the bipartite network of users and URLs was already considered in Guarino et al. [92] for Facebook: in the present case, we translate the idea therein to Twitter. The procedure goes through a comparison between observations and an entropy-based benchmark: if two users tweeted (or retweeted) the same URLs significantly more than the benchmark, we conclude that the two users share the same information diet in a statistically significant way. We can thus identify groups of users sharing the same URLs. In the following, user communities that passed the validation are called *news engagement communities* of users, for short, user NECs. User NECs contextualize the second event: exposure of users to the same news articles.

4.2.3 Echo-chamber detection

In previous sections we describe how to identify groups of users who share a common discourse (DiCo) and groups of users exposed to the same news articles (user NECs). Users who share a group of the first type and a group of the second type form an echo chamber, provided they interact with each other. The interaction for us is that of retweets

since retweets are considered as a form of endorsement to the content created by others Becatti et al. [9], Conover, Ratkiewicz, and Francisco [48], and Conover et al. [49, 50]. Verifying user interactions is an important step because accounts belonging to the same user NEC may either not belong to the same DiCo or, even in the case where they are in the same discursive community, may not interact with each other. In this sense, only users who (i) belong to the same user NEC and (ii) belong to the same DiCo and (iii) are connected, even indirectly, through retweets (i.e., they form a weakly connected component in the retweet network) can be said to represent an echo chamber.

Measuring connections among users To study how much users in echo chambers are connected, we use the undirected clustering coefficient: ignoring the direction of the edges, it captures the observed frequency of interactions between the neighbors of each node [31].

4.3 Results

In this section, we aim to address RQ.1 by conducting a detailed analysis of the selected case study, namely the online discussion about the COVID-19 vaccine in Italy. We will examine the discursive communities, called Discursive Communities (DiCos), and the News Engagement Communities (NECs) of users that emerge from the data. Furthermore, in the last subsection, we will assess the actual presence of echo chambers.

4.3.1 Discursive community

Figure 14 (top) describes the characteristics of the main discursive communities (DiCos) that emerge from the data. We recall that it is possible to assign labels to verified accounts, as the identity of their owner has been certified by the platform. Starting from the original dataset, we run the community detection algorithm [19] on the validated network of verified users and the label propagation algorithm [167] on the network of

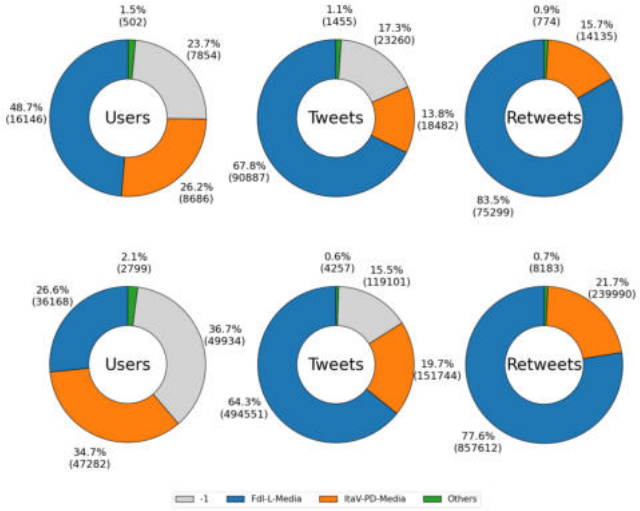


Figure 14: Characterization of the main DiCos in terms of the number of users, tweets, and retweets. Charts at the bottom only consider tweets and retweets that contain URLs.

retweets of the different communities. In our case study, two main discursive communities emerge, associated with political parties and Italian newspapers. Specifically, most of the users who are part of a DiCo belong either to the ITAV-PD-MEDIA community ($\sim 34.7\%$; the community includes journalists and exponents of the Italian parties Italia Viva and Democratic Party) or to the FDI-L-MEDIA community ($\sim 26.6\%$; the community includes journalists and exponents of the Italian parties Fratelli D'Italia and Lega). About 2.1% of users belong to smaller DiCos, while $\sim 36.7\%$ of users do not belong to any DiCo. The FDI-L-MEDIA community posted the most new content (64.3%), although it represents about a quarter of all users in our dataset. The ITAV-PD-MEDIA community is responsible for 19.7% of the new content, while the remaining 15.5% is posted by users who do not belong to any particular community. In terms of retweets, FDI-L-MEDIA is by far the most active community with 77.6% of the retweets. Figure 14 (bottom) characterizes DiCos by focusing only on posts containing URLs. In general, the observations made for the top doughnut charts still hold, with the exception that almost half of the users who post tweets with URLs belong to the FDI-L-MEDIA community (48.7%).

4.3.2 News Engagement Communities (NEC)

Table 13 shows that of all users who have published at least one post with a URL ($\sim 33k$), only 566 are part of a user NEC, which is less than 2%. Accounts in user NECs are proportionally much more active in publishing URLs than users not validated by our procedure (67.7 vs. 5.90 URLs per account).

Table 13: Users in user NEC. Validated users represent a limited minority of all accounts in the debate (less than 2% of users who shared at least one URL). Percentages refer to the proportion of users in each group.

Type	No. Users	(%)	verified	No. URL
Non-validated	32,622	98.29	434	192,334
Validated	566	1.71	1	38,345

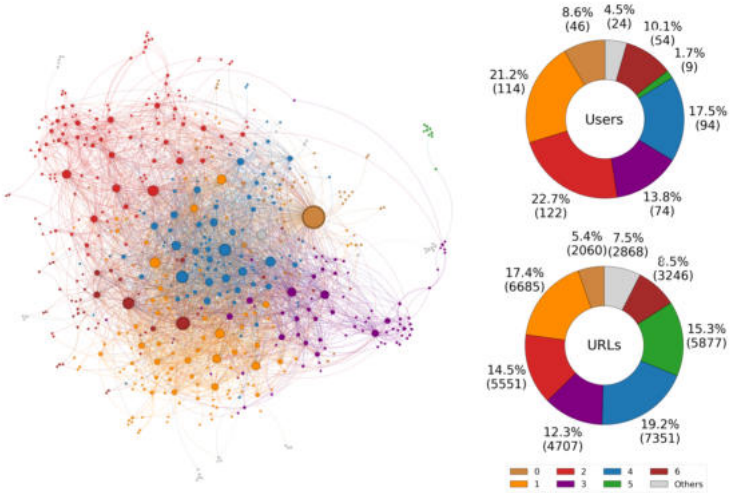


Figure 15: Left: Network representation of user NECs. Right, top: percentage (and number) of user NEC users belonging to each group. Right, bottom: Percentage (and number) of URLs disseminated by users belonging to the various user NECs.

The left panel of Figure 15 shows how the 566 users cluster into different user NECs, while the right panel provides a statistical view of the 566 users associated with the user NEC. On the right, the top doughnut chart illustrates the largest communities based on the number of users. Each of these prominent user NECs (IDs 0, 1, 2, 3, 4, 5, and 6) accounts for at least 95% of the total user population within this type of community. Furthermore, the lower doughnut chart shows that these communities have the highest frequency of tweets containing URLs. Communities 1, 2, 3, 4, and 5 collectively account for over 78% of the total URL traffic generated by all user NEC communities. Further details regarding the News Engagement Communities (NECs) of users identified in this study are provided in Appendix A.1.

4.3.3 Echo-chambers during the Italian COVID-19 vaccination debate

Our analysis shows that all but 1 of the 566 users in the user NECs are also part of the same discursive community, i.e. FDI-L-MEDIA. This is the discursive community with users affiliated with political parties Fratelli D'Italia and Lega, and news outlets showing similar leanings. However, the fact that all users in the user NECs belong to the same DiCo only tells us that users with similar 'information diets' contribute to the formation of the same discourse, but not that they influence each other and reinforce the opinions of their siblings. In other words, users who refer to the same news sources may never meet on the platform. In fact, the information about who interacts with whom is not used to detect user NECs.

As mentioned, users in an echo chamber are users who share a common discourse, are exposed to the same news sources, and are exposed to the same opinions. Being exposed to the same opinions, translated to Twitter, means that they retweet each other. In this sense, if users in the same user NEC form a (weakly) connected component in the same DiCo-induced subgraph of the retweet network (i.e., if there is a flow of influence in the retweet network that is restricted to nodes in the same

discursive community), they form an echo chamber.

The analysis of the weakly connected component shows that 92 users do not belong to it. This leaves 473 users trapped in echo chambers. In particular, all users in user NECs 8, 9, and 10 did not retweet others in the same user NEC on the topic under analysis. Regarding the other user NECs, we observe that for each of them, most of the nodes form echo chambers.

Table 14: Number of users in each echo chamber. Echo chambers inherit the ID of their user NEC. Only echo chambers 1 and 2 include more than 100 accounts. Nevertheless, echo chambers still represent a minority of all users in their discursive community.

EC_0	EC_1	EC_2	EC_3	EC_4	EC_5	EC_6	EC_7	EC_{11}	**Total EC**
24	106	122	66	90	9	28	20	8	473

The numbers of users per echo chamber are summarised in Table 14; in the following, echo chambers inherit the ID of their user NEC. From Table 14, some echo chambers are relatively big: for instance, the ones induced by user NECs 1 and 2 include more than 100 nodes.

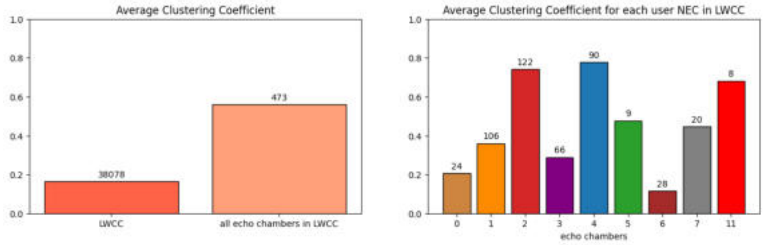


Figure 16: Left: average clustering coefficient measured on the LWCC of the retweet network restricted to users of FDI-L-MEDIA and measured on all users belonging to echo chambers. Right: average clustering coefficient calculated on each echo chamber. Each echo chamber inherits the ID and the color from its user NEC. The number of users in the echo chamber is shown at the top of each bar.

We compare the clustering coefficients (subsection 4.2.3) of the echo chambers with the one measured on the Largest Weakly Connected Com-

ponent (LWCC) of the retweet network restricted to users in the FDI-L-MEDIA DiCo. In this way, we have a benchmark that captures the main contribution to the discourse to which the echo chambers belong. The clustering coefficient associated with users in echo chambers is more than three times as high as that for other users within the LWCC (0.56 compared to 0.16, left panel of Figure 16). We then examine the average clustering coefficient within each echo chamber. The right panel of Figure 16 shows that the average clustering coefficients of echo chambers 2, 4, and 11 are greater than 0.6. High values of the clustering coefficient imply that accounts are highly connected and frequently retweet each other. Therefore, we can conclude that their endorsement activity contributes to the reinforcement of their opinions. Such a conclusion is confirmed by a manual examination (see Section 4.4.2).

4.4 The role of echo-chambers in the common discourse

Figure 17 shows the flow of retweets within an echo chamber and between different echo chambers.

Node -1 represents all nodes in the DiCo that are not part of an echo chamber, and an arrow indicates that tweets published by the source group are retweeted by a certain number of users in the target group. Self-loops represent retweet activity within the same group. The values on the edges indicate the number of retweets associated with each interaction. Although the echo chambers are composed of a small number of users (on the order of 10^2, compared to the total number of DiCo users, on the order of 10^4), they contribute significantly to the DiCo's retweet activity. Echo chambers are involved in generating about 288k retweets, while users not in echo chambers generate about 569k retweets. More specifically, echo chambers 2 and 3 are mainly composed of popular users (in terms of received retweets), while others are mainly composed of retweeting users (0, 1, 4).

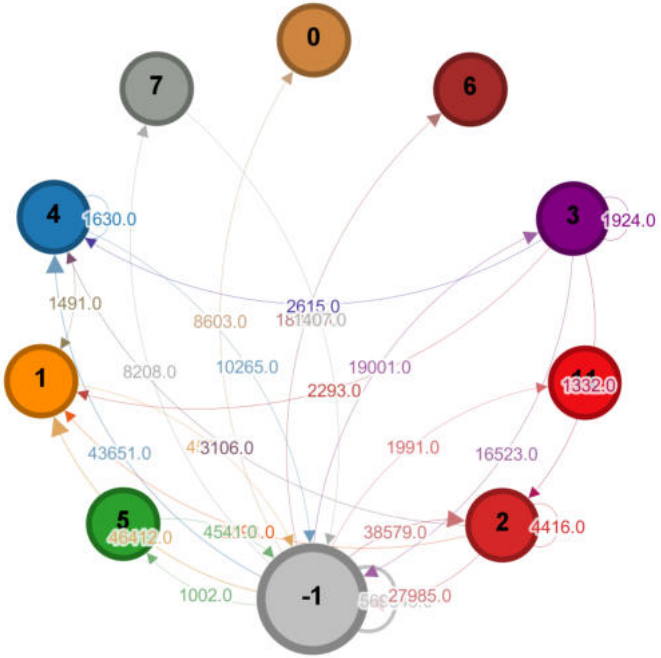


Figure 17: Retweet network for FDI-L-MEDIA **DiCo, aggregated with respect to echo chambers.** Node -1 represents users who do not belong to an echo chamber. Edges indicate the number of retweets between different user groups; weights less than $1k$ have been filtered out.

4.4.1 Exposure of users to misinformation

To quantify the presence of misinformation in echo chambers, we have tagged URLs in our dataset that point to news sites. The labels are those that the NewsGuard journalistic organization has assigned to online media outlets (see Methods 2.3.3 for a more detailed description of the evaluation process)[2].

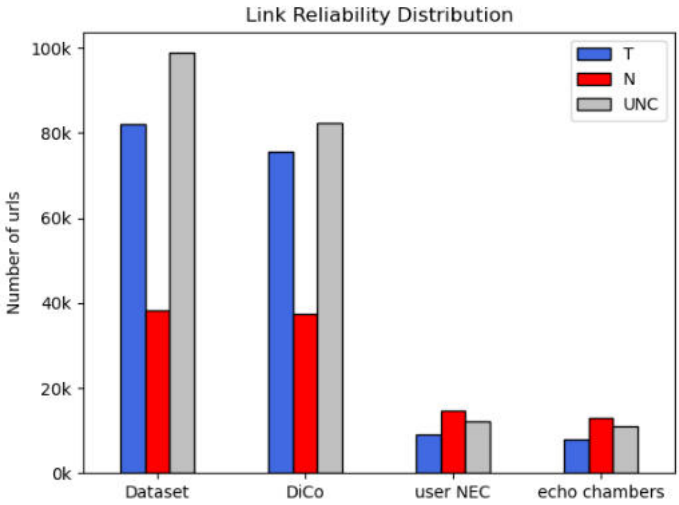


Figure 18: Number of distinct URLs pointing to news publishers tagged as 'Trustworthy' (T), 'Not trustworthy' (N), or 'Unclassified' (UNC) for the entire dataset and for each type of users' community (DiCos, user NECs, echo chambers.)

Figure 18 shows the number of URLs pointing to news from publishers that NewsGuard classifies as 'Trustworthy' (T), 'Not Trustworthy' (N), and 'Unclassified' (UNC) for the entire dataset and for each type of user community. If the same URL is shared multiple times by users in the same group, this multiplicity is taken into account in the analysis.

[2]The use of the labels has been licensed by the authors of Pratelli, Saracco, and Petrocchi [155].

The first observation is that the differences between user NECs and echo chambers are negligible. Second, DiCos cover almost the entire volume of both T and N traffic. Remarkably, while the ratio between untrusted and trusted URLs is around 0.5 for the entire dataset, the ratio is almost reversed for echo chambers: the frequency of N news sources is almost twice that of T news sources.

To better characterise users' exposure to misinformation in echo chambers, we consider the frequency of trustworthy and untrustworthy sources of URLs shared by users in each echo chamber. Also here, we consider the repetitions if a URL has been shared multiple times. The rationale for this is to characterize echo chambers in terms of the extent to which links to news stories from untrustworthy news publishers circulate within them.

If $|\text{EC}_i(\text{URL})|$ and $|\text{EC}_i(\text{URL}; R)|$ count, respectively, the number of messages containing a URL and a R-reputable URL shared by users in echo chamber i, with a little abuse of notation we can define a purity for echo chamber as

$$purity_R(\text{EC}_i) = \frac{|\text{EC}_i(\text{URL}; R)|}{|\text{EC}_i(\text{URL})|}. \tag{4.1}$$

We can define $purity_R(\cup_i \text{EC}_i)$ and $purity_R(\overline{\cup_i \text{EC}_i})$, respectively for all users in echo chambers and for all users outside echo chambers.

The results of the analysis is reported in Figure 19: on the x-axis there are the echo chambers denoted by their ids, on the y-axis the purity values. On the left panel, purities are related to trustworthy URLs. On the right panel, purities are related to untrustworthy URLs. The blue dotted line indicates $purity_R(\cup_i \text{EC}_i)$, the black dotted line indicates $purity_R(\overline{\cup_i \text{EC}_i})$.

Focusing on the $purity_R(\cup_i \text{EC}_i)$ lines, echo chambers on average have a higher purity with respect to untrustworthy URLs (~ 0.377) compared to trustworthy ones ($\sim\ 0.232$). In other words, when a user posts a message containing a URL in an echo chamber, the probability that it points to an untrustworthy news source is close to 0.4; for some echo chambers, this probability is even much higher than this. As in the case

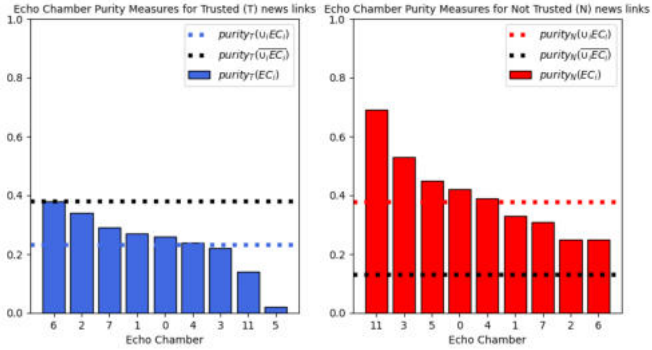


Figure 19: Purity levels of echo chambers. On the left trustworthy URLs, on the right untrustworthy URLs. While the $purity_T(\overline{\cup_i \text{EC}_i})$ value is greater than its counterpart in the echo chamber, $purity_N(\overline{\cup_i \text{EC}_i})$ is lower than the value measured in echo chambers.

of the purity for user NECs, if we compare the $purity_R(\cup_i \text{EC}_i)$ values against $purity_R(\overline{\cup_i \text{EC}_i})$, there is a trend reversal in passing from T to N: the $purity_T(\overline{\cup_i \text{EC}_i})$ value is greater than its counterpart in the echo chamber while $purity_N(\overline{\cup_i \text{EC}_i})$ is lower than the value measured in echo chambers. This finding is worrisome because users in echo chambers are particularly polarized and committed, basing their beliefs on low-quality news. However, it is important to remember that the formation of echo chambers, while alarming in itself, is generally unrelated to the quality of news sources.

4.4.2 An in-depth analysis on the narratives inside major echo-chamber and the persistence of toxic or extreme points of view

To validate our approach and provide a concrete example of the potentiality of using the proposed procedure, we will focus on the content shared in echo chamber 4.

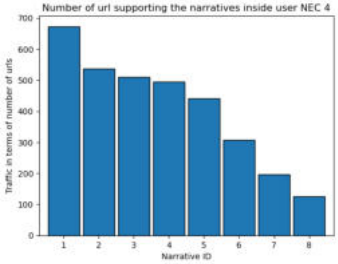


Figure 20: Impact of different narratives on URLs shared by users in echo chamber 4.

As shown in Figure 19, among the greatest echo chambers, 4 tends to share low credibility content.

In practice, we first manually extract the main narratives from the news shared within echo chamber 4, focusing on the users with the highest number of followers at the time of data collection. Then, still focusing on the users with the highest number of followers, we analyze whether there are signals of these narratives in their most recent posts (as of June 7, 2023) and which narratives they currently support.

In echo chamber 4, there are about 1.7k unique news that have been shared about 7.3k times in total. First, we exclude the news with connection errors at the time of this analysis (1k shares) and those that have been shared less than 10 times. Then, we analyze the resulting news narratives, which amount to $\sim$ 3.3k shares and 146 unique URLs from 51 different domains. By classifying only these 146 news stories, we cover about $\sim 45\%$ of the total URL traffic within echo chamber 4.

Figure 20 and Table 15, respectively, show the narratives' distribution and their descriptions: the main 8 narratives are all against vaccination and government regulations.

Table 16 shows the narratives supported by the users in echo chamber 4 with the most followers, almost two years after data collection (June 7, 2023). Users hold extreme views on current controversial issues such as the war in Ukraine, migrants, and LGBT issues. Remarkably, conspiracy

ID	Narratives
1	The discriminatory nature of greenpass. Support to protests by individuals or groups against the greenpass.
2	Several cases of people who died because of the vaccination (aneurysm, meningitis, ...).
3	Several cases of injuries after vaccination (bleeding, myocarditis, pericarditis, neurological problems, excruciating pain)
4	Various statements made by political figures against vaccinations.
5	The obligation cannot be imposed since it is unconstitutional. News about VIPs and governments rejecting the vaccine. Ineffectiveness of vaccines.
6	Mattarella incites to hatred no-vax. Experts reject the third dose, ...
7	(Manipulated) data about vaccine hazard versus efficacy and hospitalisations or infections despite vaccination.
8	COVID-19 vaccines are still too experimental. Police forces were not vaccinated. Support to views of no-vax doctors. VIPs and high-ups pretend to be vaccinated, but actually are not, due to the known dangers of vaccinations.

Table 15: Description of the narrative disseminated by echo chamber 4.

theories about vaccines are still present in their narratives.

4.5 Conclusions and observations

This chapter proposes a novel unbiased method to detect echo chambers. The method is mainly based on two observations. First, echo chambers form when users interact with others who share similar opinions and refer to the same news. Second, a proper null model should be implemented to detect a true signal. This necessity has recently been highlighted in the literature on online social networks and has been shown to be particularly important for the detection of non-trivial phenomena [9, 33, 34, 45, 46, 127]. Our echo chamber detection method is based on the validation of observed structures by comparison with a proper maximum entropy null model; the maximization of entropy guarantees the

User	Followers	Supported Narratives
user_1	36926	no-migrants, no-vax, anti-EU
user_2	6929	pro-Russia, no-vax, no-LGBT
user_3	6335	pro-Russia, no-migrants, anti-EU, conspiracy theories
user_4	4164	no-vax, no-migrants, pro-Russian
user_5	3117	no-vax
user_6	2668	pro-Russian, against the Italian government, no-vax
user_7	2641	suspended
user_8	2448	conspiracy theories, no-vax, no-LGBT, against the Italian government, anti-EU
user_9	2355	religious posts, no-green pass, no-vax
user_10	2316	against Italian government, no-vax

Table 16: Main narrative supported in recent posts (as of June 7, 2023) of users in echo chamber 4 with the highest number of followers. Users are anonymized.

unbiased nature of the benchmark.

We apply the proposed methodology to gain insights from the Italian Twitter debate on Covid-19 vaccination debate: we found that our procedure detects a low presence of echo chambers (just under 0.35% of all users in our dataset belong to an echo chamber). All the echo chambers we detected are part of the same discursive community, i.e. a community of users with similar political positions. Even if their dimension in terms of the number of users is limited, their impact on the shared discourse is remarkable: echo chambers are responsible for almost a third of the retweets in their discursive communities.

The methodology can be extended to other online social networks. In fact, it is based on i) the analysis of the activity of accounts that share URLs to news sources and ii) the detection of discursive communities. While the extension of the former to other online social networks is straightforward, the latter may be more problematic: in the present case, we used the activity of verified users, who are among the main content creators in Twitter [9], but not all social platforms have such certification. Nevertheless, when analyzing other platforms, we can still focus on users who are particularly active in creating new content, such as influential users as defined in [89].

Not unlike other studies, our study has some limitations, which we believe do not affect our final conclusions. First, it may be argued that the validation procedure is quite strict: the validation of multiple p-values leads to the validation of extreme events. While this is true, it is the only way to eliminate random noise from the system and analyze the true

signal (see Appendix A.2 for a detailed analysis). Another aspect that could be further investigated concerns the role of highly active users in the formation and reinforcement of echo chambers. In the present work, echo chambers are detected mainly through similarity patterns among users, namely shared discourse, common exposure to the same news, and retweet-based connectivity. However, identifying highly prolific accounts (often referred to as superspreaders) could provide additional insights into how visibility and engagement become concentrated around specific users, and into how these actors may contribute to amplifying both reliable and unreliable information within politically homogeneous groups. Exploring this aspect therefore represents a promising direction for future work.

Finally, the main idea of echo chambers is that users follow accounts with similar ideas, while in the present study only the retweet network is used, not the information about friendships. Still, the retweet network captures the effective interactions with interesting content as perceived by different users, whether it comes from friends or is suggested by the platform itself: focusing only on friendship will not fully capture the effect of the platform's recommendation algorithm.

Observations During the experiments conducted as detailed in [155], we focused on studying the engaging patterns between online news outlets and social media users.

As reported in this chapter, the detection of echo chambers was guided by identifying well-connected user groups affiliated with the same political group and driven by an interest in the same set of narratives. The results show that the narratives generating the most interest played a significant role in the analyzed discussion.

Based on this observation, the authors considered utilizing the engagement generated by these narratives to identify relevant publishers, namely entities capable of creating compelling and engaging narratives. The idea of the mechanism presented in the next chapter is to exploit the known trustworthiness of relevant publishers (those worth evaluating manually through expert opinion) to estimate the trustworthiness of

other publishers, leveraging on the characterization of users that share it (measuring their tendency to share untrustworthy content).

Furthermore, in the same vein, in [155], it has been observed that groups of articles that engage statistically similar groups of social users typically share similar properties at the publisher level. Particularly, the authors noted that these narratives are homogeneous with respect to their level of quality. Essentially, these groups of news are predominantly published by online newspapers characterized by the same level of quality, be it low or high. In particular, from the characterization of URL NECs, we observe a significant amount of article clusters that are most often judged untrustworthy by NewsGuard. These results suggest that engagement mechanisms can be a good indicator to also capture low-quality domains: these newspapers typically attempt to engage the audience by producing attractive narratives capable of capturing the attention of vulnerable social users, such as those affiliated with a particular political ideology. The implication is that identifying low-quality publishers also involves detecting users who could potentially be interested and engaged by even low-quality producers. These insights are used for the developments presented in Chapter 6.1.6.

Chapter 5

Automatically Evaluating News Publisher and Articles Trustworthiness

This chapter presents a key part of the second methodological component of the thesis's second core contribution: the development of scalable and transparent approaches for the automatic assessment of news article and publisher trustworthiness. Driven by the limitations of manual expert-based evaluations—such as high costs, limited scalability, and subjectivity—this line of research investigates how computational tools, particularly large language models (LLMs) and interaction-based signals, can assist or replicate expert credibility assessments.

The chapter presents two interrelated studies conducted in sequence. We begin, in Section 5.1, with an analysis of real-world procedures for evaluating publisher credibility, focusing on two prominent initiatives: NewsGuard and the Global Disinformation Index (GDI). This analysis serves a dual purpose: (i) to better understand the structure and criteria of current expert-based frameworks, with the goal of identifying elements suitable for automation; and (ii) to assess the degree of alignment between the two procedures, both in methodology and outcomes.

Building on these insights, Section 5.2 introduces a scalable approach

to trustworthiness assessment by automating a selected set of expert-derived criteria using the capabilities of large language models. The goal is to evaluate whether LLMs can approximate expert judgments and offer a viable path toward scalable, consistent, and explainable credibility assessments.

5.1 Comparative analysis of journalistic evaluations: bridging GDI and NewsGuard

5.1.1 Problem formulation and contributions

In today's era of information disorder, many news organizations are moving to verify the veracity of news published on the web and social media. In particular, some agencies are exploring the world of online media and, through a largely manual process, ranking the credibility and transparency of news sources around the world.

In the analysis reported in this Section, we evaluate two procedures for assessing the risk of online media exposing their readers to m/disinformation. The procedures have been dictated by NewsGuard and The Global Disinformation Index, two well-known organizations combating d/misinformation via practices of good journalism.

Specifically, considering a fixed set of media outlets, we examine how many of them were rated equally by the two procedures, and which aspects led to disagreement in the assessment.

Here, the main initial question, which lays the foundation for this analysis, is: *Would it be worthwhile to automate the process?* With an abuse of notation, what we mean is to build a sort of 'automated fact checking on the news outlet', where what is evaluated is not the truthfulness of the single fact, or claim, or post –such as in, e.g., [216, 221]– but the transparency and credibility of the whole online news media. The automation of the process is particularly appealing for what concerns the evaluation of the 'tail' of news sources (brand new ones or with less traffic than those considered by the actuators of the manual verification).

In order to gain a deeper understanding of the processes performed

by specialised organisations and to lay a solid foundation for process automation, in this section we present an evaluation of the procedures and results of the GDI and NewsGuard evaluation methodologies, performed on the same set of media. An assessment of this kind acquires considerable importance if we consider that, in the literature, researchers have conducted analysis in the field of misinformation by leveraging the tags of news sources assigned by such organizations, see, e.g., [3, 33, 91, 127, 180]. Despite the wide use of this approach, no one, to the best of our knowledge, has measured (at the date of publication of our study [152]) the agreement between different evaluation processes (agreement measured both in terms of criteria adopted and in terms of final scores given to the news sources).

Thus, the first research question we would like to answer here is:

RQ1 – *Do the implementation of different decision processes lead to the same results in terms of estimating the reliability of online information sources?*

As introduced above, both organizations evaluate worldwide online media markets, based on a set of criteria inspired by good journalism principles and practices. The considered criteria, while similar to each other, are obviously not the same. Thus, the procedures' assessment must necessarily address the issues that give rise to the second and third research questions:

RQ2 – *Can we establish a conceptual mapping between the two sets of criteria? In other words, is it possible to relate the criteria adopted by one organization to the criteria of the other, and vice versa?*

RQ3 – *Focusing on the criteria that find conceptual mapping, and considering the result of their application on the same set of news sources, what are the criteria that obtain a concordant (resp., discordant) evaluation?*

Hereafter, we illustrate how we intend to proceed to answer the three research questions and we summarize the results.

The authors took part in a study commissioned by GDI, to assess the risk of disinformation exposed by a set of Italian online media, representative of the online media national market by geographical distribution, circulation and political ideology. The methodology, an overview of the contextual scenario and the results of the study are available online [148].

In addition, the authors (at the time of this study) own a valid NewsGuard license, for which they have access to the so-called 'Nutrition Labels' of the news media, outcome of the NewsGuard's evaluation procedure.

By virtue of the above, we can compare two media rankings: the one obtained via our study for GDI, and the one obtained using the NewsGuard's labels[1].

Of the 31 online media analyzed, 7 have a different GDI score than the one obtained according to Newsguard. In particular, all 7 media result reliable for NG and not reliable for GDI. To understand why scores were so different, we went down to the level of the evaluation of the individual criteria. To do this analysis, we initially mapped GDI criteria and NG criteria, which we believe may also be useful for further investigation by Academia.

After the conceptual mapping, we looked at which criteria the two organizations rated differently. These are criteria that concern the policies that the media puts in place to ensure its transparency (e.g., the existence of declarations of editorial independence, declarations on funding sources and ownership structure). So, on the one hand, we can conclude that the concordance on criteria that mainly concern the analysis of news (discordance between title and text, use of sensationalist language, specific words or punctuation in the text...) lays a solid foundation to attempt the automated approach. Section 5.2 of this chapter leverages these insights, particularly the *worthy-to-be-automated* list of criteria, as a foundational set to guide automation. This approach aims to enhance scalability through the application of large language models (LLMs). On the other hand, the discordance on criteria regarding presence of editorial line statements or ownership structure seems very odd, and probably the reason for this dissonance must be sought in the annotation campaign that each of the two agencies carries on. This opens the door for

[1]Because of contractual agreements, we cannot report in plain text the ranking obtained according to the GDI methodology, nor any data that could make it inferable. We can, however, conduct the comparison by reporting the differences found in the two rankings. In addition, the work that led to the actual results was conducted under the terms of the NewsGuard license.

further investigations on the objectivity/subjectivity of annotation and crowd sourcing campaigns, which are currently discussed in works such as [168, 186].

Contributions This work brings the following contributions:

- The realization of a conceptual mapping between the criteria used by two agencies that have the same goal: to evaluate the degree of transparency and credibility of an online newspaper;
- An analysis of the level of agreement in the evaluation of the same set of publications by different agencies;
- The basis for starting to think about which criteria can be calculated automatically, so as to partially automate the evaluation process and embrace a much larger number of news media.

The result of our analysis shows a good degree of agreement, which in our opinion has a double value: it fortifies the correctness of the procedures and lays the groundwork for their automation. Overall, we believe that our study contributes to new directions in the automation of reliability verification, expanding the object of verification automation from the individual post/news to the news source.

5.1.2 Useful notions

Here, we remind the reader the criteria used by the two to evaluate the news outlets, as well as their scoring systems.

5.1.3 GDI and NG criteria

Tables 33 and 32 (see the Appendix C) show the list of criteria considered by the two organizations for evaluating a news site. In particular, Table 33 reports the list of the GDI criteria, as they appear on the GDI website and in many of the GDI reports.

The first column is the criterion name, the second one reports an abbreviation for that name (coined by us), the last column defines the criterion. Sometimes, a criterion is split into more subcriteria, reported in the third column.

Table 32 shows the same information for the NG criteria. Even in this case, the criteria are public available on the NG website[2]. The differences here are that criteria are not split in subcriteria, and the last column reports the score the news media gets when it meets the criterion.

Both organizations consider 2 categories of criteria, one category that is more about the content and presentation of the individual news story, and one category that is about the editorial procedures and policies of the online media.

Regarding GDI, in Table 33 the first 9 criteria are related to content (e.g., the assessment of how well a headline reflects the content of the article, the presence or absence of a fact-based lede, the use of sensationalist terms in the article). The other 6 criteria relate to the media as a whole (e.g., the presence of statements of editorial independence, the attestation of the newspaper's editorial and financial structure, the presence of statements on sources of funding).

With regard to NG, the first 5 criteria in Table 32 concerns the analysis of the content of the single article, while the remaining 4 criteria concern the transparency of the whole newspaper.

5.1.4 GDI and NG scoring systems

The evaluation of the criteria leads both organizations to produce a final score for each news outlet[3]. The score expresses how well the news source meets their criteria and, thus, fulfills good editorial principles and practices. Then, pre-defined threshold values determine the quality of the source under investigation:

[2] https://www.newsguardtech.com/ratings/rating-process-criteria/

[3] Computation of the final score, for both agencies, is described on their websites: https://disinformationindex.org/wp-content/uploads/2019/12/GDI_Index-Methodology_Report_Dec2019.pdf; https://www.newsguardtech.com/ratings/rating-process-criteria/.

- GDI category risk: Each media obtain an overall score as a result of the criteria evaluation. The news sources are then classified on the basis of a five-category risk scale based on the overall score. The risk categories were defined based on the distribution of risk ratings from 180 sites across six media markets in September 2020. This cross-country dataset was standardised to fit a normal distribution with a mean of 0 and a standard deviation of 1. The standardised scores and their distance from the mean were used to determine the bands for each risk level (i.e., minimum, low, medium, high, maximum). These bands are then used to categorise the risk levels for sites in each subsequent media market analysis. On a maximum score of 100, online media with a score $<$40 are labeled as 'high risk of exposure to disinformation', between 40 and 50 a medium-high risk, between 50 and 60 a medium risk, between 60 and 70 a medium-low risk, and those with a score $>$70 pass the assessment positively.

- NG overall score: NG uses 9 criteria to evaluate a news source. The total score the source can obtain is 100. Each criterion is associated with a numerical value, and the value assignment is all or nothing (criterion satisfied = associated value; criterion not satisfied = 0). Obviously, the sum of the values assigned to the 9 criteria is 100. A news site with a score of 60 points or higher receives a positive rating in terms of credibility and transparency. A site with a score lower than 60 points receives a negative rating. A news site that achieves a score greater than 60 can still fail in one or more of the 9 criteria. The Nutrition Label provided by NG details which criteria are met and which are not.

5.1.5 Methods

The methodology of the analysis will lead to two types of comparison regarding the procedures implemented by the two agencies for the evaluation of the same set of news outlets.

The first comparison concern the results of the evaluation, that is,

we measure the amount of news sources judged (i) the same way (i.e., judged reputable by both GDI and NG), (ii) in disagreement (i.e., reputable by GDI but not reputable by NG, and vice-versa). To do this we consider (i) the overall scores of NG and GDI (See Section 'GDI and NG scoring systems') and (ii) the thresholds applied by both NG and GDI to the overall scores for identifying not reliable sources. The first comparison will give an answer to research question **RQ1**.

The second comparison will address the criteria: we will evaluate the amount of GDI criteria that find conceptual mapping with NG criteria, and vice versa. To do this, we will create an association matrix (GDI criteria - NG criteria), based on the specifications provided by the two organizations. This will allow us to answer question **RQ2**.

Finally, for the criteria that find mapping, we will measure the level of agreement, i.e., that is, how much each GDI criterion has been evaluated in the same way as its homonym NG criterion (and vice-versa). This will answer question **RQ3**.

Dataset The list of online Italian media considered in this article is a subset of the 34 online news outlets evaluated in the report 'Disinformation risk assessment: The online news market in Italy' [148]. The 34 news outlets, reported in Table 31 (Appendix C), have been chosen as representative of the Italian online publishing landscape. In particular, GDI proposed a first selection, based on the sites' reach (Alexa rankings, Facebook followers, and Twitter). Then, the set was thinned to arrive to a balanced group in terms of diffusion (either national or local), geographical location (i.e., North, Centre, or South and islands Italy) and political orientation (ultra-right, right, mainly neutral, left, ultra-left). Finally, for this work, we focus on 31 of the 34 outlets, because they were tagged both by NewsGuard and by GDI.

5.1.6 Results

Evaluation results To answer Research Question **RQ1**, we consider the scores obtained by the 31 online media according to the GDI and News-

Guard assessments. It is important to note that NewsGuard applies a sharp division between trustworthy and untrustworthy online media (out of a maximum score of 100 points, those scoring $\geq$60 pass the assessment positively). In contrast, GDI divides media into 5 bands of risk of exposure to disinformation.

In order to compare the two rankings, we simply considered a two-class division for both the organizations. We proceeded as follows: we mapped the 5 GDI risk levels on these 2 NG classes. In particular, the NG untrustworthy class was associated with the maximum and high disinformation risk levels; the NG trustworthy class was associated with the medium, low and minimum risk levels. While the two-class division makes the evaluation less granular, it also makes the two rankings comparable, and we maintain the same thresholds originally chosen by the two associations, i.e., 60 for NG and 50 for GDI (50 is the threshold established by GDI to divide the high and medium risk levels).

	GDI Neg	GDI Pos
NG Neg	4	0
NG Pos	7	20

Table 17: Agreement and discordance in the investigated media ratings. One threshold value (50 for GDI, 60 for NG).

Table 17 reports the number of sites that received concordant/discordant ratings from both organizations. Out of 31 sites, 20 received a positive rating from both, 7 received a negative rating from both, and 4 received a discordant rating. Notably, the latter were rated positively by NG, and negatively by GDI. None of the 31 sites received a positive rating from GDI and a negative rating from NG. As a percentage, 77.4% of ratings are in agreement.

Mapping of criteria We proceed with the analysis by looking for the two agencies' criteria there were evaluated in a convergent/divergent way. To reach the scope, we construct a conceptual mapping between criteria. This task has been carried out by the two of us, aided by our

knowledge of the GDI criteria, which we evaluated on each of the 31 news outlets, see [148]. We initially conducted the task on our own, associating GDI criteria with NG criteria, and vice-versa, based on their textual description and our experience. It took us about three hours to finish the task. Next, we compared our associations, dwelling on those for which we disagreed. For the few discordant associations, we managed to find a common view and, thus, an agreement. This second phase lasted about two hours.

Table 18 shows the result. Labels S and W stand for Strong and Weak. We assigned a Strong connection between criteria when the definitions were such that they were virtually the same. This is the case, for example, of GDI: *Headline Accuracy* and NG: *Avoid Deceptive Headlines*.

A Weak label was instead assigned when only part of the definition of a criterion by one agency is matched by the definition of another criterion defined by the other agency. For example, GDI: *Common Coverage* is, according to its definition, 'indicative of a true and significant event' and, thus, it finds a connection with NG: *Does not repeatedly publish false content*, even if in a less direct way.

As shown in Table 33, some of the GDI criteria find declination in multiple subcriteria. An asterisk in the mapping indicates that a NG criterion has a partial match with some of the subcriteria of a GDI criterion.

Moreover, to give an idea about the strength of the conceptual link found between each criterion of organization x mapped onto one (or more) criteria of organization y, we introduce the concept of *Conceptual Mapping Level* (CML). In particular, we have defined the following four CML levels of mapping, based on the proportion of Strong and Weak mappings identified for each criterion:

- 4, Strong: There is perfect conceptual mapping between 1 criterion of organization x and 1 (or more) criteria of organization y;
- 3, Almost Strong: If a criterion of organization x finds mapping in more than one criterion of organization y, at least half of these mappings are Strong;
- 2, Almost Weak: If a criterion of organization x finds mapping in

more than one criterion of organization y, and more than half of these mappings are Weak;

- 1, Weak: There is no perfect conceptual mapping between 1 criterion of organization x and 1 (or more) criteria of organization y.

	RepFalseCont	InfoResp	ErrCorr	NewsOpDiff	AvDecHeadlines	DiscOwnFin	LabAds	RevConfOfInt	ContCreators	GDI to NG CML
ArtBias		S		S						4
BylInfo		W							S	3
ComCov	W									1
HeadAcc					S					4
LedePres		S								4
NegTarg		S								4
RecCov	W	W								1
SensLang		S			W					3
VisPres		S								4
Attr		S							W	3
CommPol										-
EdPrincPract		S^*		S^*		S^*		S^*		4^*
EnsAcc	W	S^*	S^*							3^*
Fund						S	W^*	W^*		2^*
Own						S		S^*		4^*
NG to GDI CML	1	3^*	4^*	4^*	3	4^*	1^*	3^*	3	

Table 18: Conceptual mapping between GDI and NG criteria (in the left column and header, respectively). The degree of mapping can be 'strong' (S) or 'weak' (W). When a GDI criterion is composed of multiple subcriteria, the S or W assignment is by majority vote. The asterisk indicates that a NG criterion has been partially mapped into a GDI one (i.e., we found a partial conceptual mapping with only some of the GDI subcriteria from which it is composed.). Side numbers summarize the *Conceptual Mapping Level* (CML), explained in the 'Mapping Criteria section'.

The results of the conceptual mapping (see Table 18) can be summarised as follows:

- All NG criteria find at least one mapping to GDI criteria. Seven (*InfoResp*, *ErrCorr*, *NewsOpDiff*, *AvDecHeadlines*, *DiscOwnFin*, *RevConfOfInt*, *ContCreators*) reach at least the 'Almost Strong' CML and three of these (*ErrCorr*, *NewsOpDiff*, *DiscOwnFin*) reach the 'Strong' one. Six criteria have at least a partial mapping; three of these (*ErrCorr*, *NewsOpDiff*, *RevConfOfInt*) with at least half of partial mappings. Only two criteria (*RepFalseCont*, *LabAds*) reach the 'Weak' value of CML.

- Fourteen GDI criteria (out of fifteen ones) find at least one mapping to NG criteria. The only one for which we cannot find a mapping is the criterion regarding the presence, on the website, of policies regulating the user-generated content: the *Comments policies*. Eleven criteria (*ArtBias*, *ByInfo*, *HeadAcc*, *LedePres*, *NegTarg*, *SensLang*, *VisPres*, *Attr*, *EdPrincPract*, *EnsAcc*, *Own*) reach at least the 'Almost Strong' value of CML. Four criteria (*EdPrincPract*, *EnsAcc*, *Fund*, *Own*) have at least half of partial mappings. Only three criteria (*ComCov*, *RecCov*, *Fund*) reach a value of CML less (or equal) than 'Almost Weak'.

Recalling the question **RQ2**, where we wondered whether the criteria of the two organizations could be mapped onto each other, the answer is positive. Some mappings are stronger than others in the sense that we find a precise match in the definition of the criteria. Still, in the end, (i) all the criteria are matched (apart from the GDI criterion on moderation of user-generated content) and (ii) the majority of the criteria (for both GDI and NG) achieves a CML value of at least 'Almost Strong'. These positive signals show that the two agencies are moving along similar guidelines for assessing the credibility and transparency of an online media outlet. In the next section, we will seek to answer the third research question, regarding the evaluation of the criteria agreement.

Evaluation of the agreement on the single criteria Regardless of how a source was classified (low/high reliable by either GDI or NG or both),

here we analyse how much agreement there is between the criteria that find a conceptual mapping, considering all the media sources in the study.

Thus, for each GDI criterion and the NG criteria that find conceptual mapping to it, Figure 21 shows the level of agreement between the evaluation of the GDI criterion and that of its NG 'analog' ones.

On the x-axis, we list each GDI criterion. The y-axis shows the percentage of the investigated online media rated the same under the analogous NG criteria. For example, the NG analogs of the *Article Bias* criterion, which aims to assess whether a news story is written in neutral or biased terms, were evaluated the same way on 30 media out of 31 ones (in percentage terms, is 96.77%).

The GDI criteria whose NG analogues were evaluated more discordantly concern the presence of statements about the ownership structure of the media (*Own*); the implementation of pre-publication fact-checking and post-publication error correction procedures (*EnsAcc*); the presence of independence declarations on the media website (*EdPrincPract*) and the declarations about funding sources (*Fund*). For the last 3 GDI criteria, 20% of the analyzed media only (i.e., 6 media) feature the same evaluation of the analogous NG criteria.

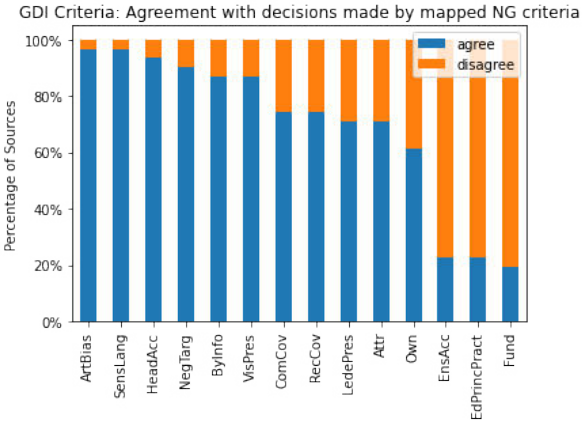


Figure 21: For each GDI criterion, percentage of online media rated the same under the analogous NG criteria

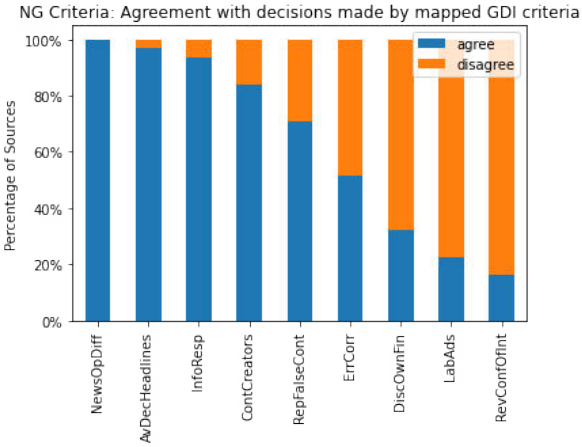


Figure 22: For each NG criterion, percentage of online media rated the same under the analogous GDI criteria.

Figure 22 shows the same analysis, starting from each NG criterion.

What we notice is that the biggest disagreement in the evaluation of similar criteria is always on those that concern the investigation of the whole newspaper, and not on those that concern the evaluation of the single news item. In fact, even in the case of Figure 22, the similar GDIs of ownership, funding sources, editorial structure and guidelines (*DiscOwnFin, ContCreators, RevConfOfInt*) were evaluated in the same way in well under 40% of the considered media.

To give more insights about the agreement values, we also study the relation between the agreement and the level of mapping (see Section 'Mapping of criteria'). Table 19 lists the GDI and NG criteria, the category to which they belong (T=transparency of media/CR=credibility of published news), the level obtained in the conceptual mapping (see also CML values in Table 18), and finally, the degree of agreement with the mapped criteria of the other organization (same values as in Figure 21 and Figure 22).

We observe that the GDI criteria that obtain a highest agreement (i) belong to the category *credibility*, and (ii) have no partial conceptual map-

Criteria	Cat.	CML	Agreement(%)
GDI			
ArtBias	CR	4	96.77
SensLang	CR	3	96.77
HeadAcc	CR	4	93.55
NegTarg	CR	4	90.32
Bylnfo	CR	3	87.1
VisPres	CR	4	87.1
ComCov	CR	1	74.19
RecCov	CR	1	74.19
LedePres	CR	4	70.97
Attr	T	3	70.97
Own	T	4*	61.29
EnsAcc	T	3*	22.58
EdPrincPract	T	4*	22.58
Fund	T	2*	19.35
NG			
NewsOpDiff	CR	4*	100
AvDecHeadlines	CR	3	96.77
InfoResp	CR	3*	93.55
ContCreators	T	3	83.87
RepFalseCont	CR	1	70.97
ErrCorr	CR	4*	51.61
DiscOwnFin	T	4*	32.26
LabAds	T	1*	22.58
RevConfOfInt	T	3*	16.13

Table 19: Relation between the *Conceptual Mapping Level* (CML) and the agreement on the criteria evaluation. Category={CR=credibility, T=transparency, }; The asterisk on the CML value indicates at least one partial mapping.

ping (no asterisk). We recall that partial mapping (presence of asterisk) indicates that the criterion finds mapping but is only partially represented by the other agency. It is, therefore, reasonable to think that, in presence of partial mapping, the agreement in the criteria evaluation is lower, since some specific aspects of one criterion are not taken into account by the set of criteria on which it is mapped.

NG criteria with the highest CML values belong that to the *credibility*

category. Most of the criteria with high agreement find at least a partial mapping: this is true for *NewsOpDiff* and *InfoResp* and means that, although the mapping is partial, these two criteria are still well represented by the mapped criteria. This assumption is also supported by the high CML values achieved by *NewsOpDiff* and *InfoResp* (Strong and Almost Strong, respectively) and a large number of mapped criteria for *InfoResp*.

5.1.7 Discussion

The analysis of this Section born from the question of whether and to which extent it is possible to automate the current journalistic procedures of evaluation of an online media, based on criteria that consider aspects of credibility of the published news and transparency of the media itself.

Moving towards automated classification of news media –a *kind of* automated fact checking, but on the whole news outlet rather than the single piece of information– is useful since, to date, many online media have not yet been manually evaluated. This is certainly true for the long tail of news outlets characterized by low traffic. However, even in considering the 34 Italian media representative of the country's information landscape, 3 of them do not have a NewsGuard rating, at time of writing.

However, in our opinion, it is useless to start with the automatic computation of certain features, if we do not first carry out an assessment of the solidity of the manual procedures implemented to date. Hence, we started by assessing how many, out of a fixed set of media, were considered the same way by two well-known organizations of journalists and media experts working to unveil low reliable news sources.

The ranking of the analysed online media was available because I contributed to a study conducted on behalf of the Global Disinformation Index (GDI) at the beginning of my PhD. As part of this collaboration, I participated in a training phase followed by an annotation phase in which two additional annotators and I assessed both website-level and article-level quality according to the criteria defined by GDI. Due to contractual confidentiality agreements, the full ranking cannot be publicly

disclosed; however, it was available to us during the present study and only aggregated information is reported here. For NewsGuard, at time of the analysis we hold a NG license.

The first analysis carried out was to quantify the number of sites judged in agreement by both organizations. A coarse-grained analysis (i.e., using only 1 threshold to distinguish 'good' news media from 'bad' news media) found that 20 sites were judged negatively by both the agencies, 7 positively, while for 4 sites they gave discordant judgments. This addresses research question **RQ1**.

The second analysis led to the definition of a conceptual mapping between the GDI and NG criteria. Mostly based on the literal definitions of the criteria, we found that the criteria have a correspondence, in some cases strong (the definitions of the criteria are essentially the same), in some cases weak (the definition of a GDI criterion includes or is included in the definition of an NG criterion). In addition, the correspondence may be partial when the definition of a NG criterion includes, or is included in, some of the sub-criteria into which a GDI criterion is divided. The construction of the mapping answers research question **RQ2**.

Finally, we asked which of all the mapped criteria were evaluated the same by the two organizations, on the largest (or smallest) number of online media. The outcome of the analysis was, in our opinion, noteworthy. The criteria regarding the evaluation of the single news item (such as, e.g., the appropriateness of the title, the presence of the author's name, the presence of a fact-based lede, the type of language, the visual presentation, etc.) are the criteria whose evaluation is agreed upon on all, or many, of the analyzed news outlets. There is a greater discordance in the evaluation of criteria concerning the news media as a whole. In particular, one agency, GDI, rated some media more negatively on these criteria than the other. This answers research question **RQ3**, and brings food for thought.

On the one hand, the discrepancy in the rating may be due to how annotators were instructed by the organizations employees. On the other hand, there is an underlying bias on annotation processes. Considering different groups of annotators (or samples of articles) may lead to differ-

ent annotations and, consequently a possible different assessment of the reliability for the same news source. Work in [16] highlights that among groups of experienced annotators with different backgrounds (i.e., journalists and scientists), there is no perfect agreement regarding the credibility assessment of news. Therefore, it is reasonable to think that in case organizations *a la* GDI and NG commission the analysis to people with different skills, the result of the evaluation changes even if the criteria find a perfect conceptual mapping. A promising direction to improve the annotation process is to set up crowd-sourcing campaigns formed by persons with different expertise and then compare the result of the evaluations with the judgment of experienced journalists. Once evaluated which of the campaigns gives a result more similar to that given by the experts, the adoption of ad-hoc campaigns could assist the work of these journalistic organizations, given their difficulty in scaling up. Furthermore, crowd-sourcing based projects have recently been exploited also by technology companies such as Meta and Twitter to review and rate viral misinformation (see, e.g., [132], [196]). Thus, the move to use such campaigns to judge the source of news appears promising.

Summing up: The meta-question we asked in the introduction of this Section was: *Is it worth trying to automate the evaluation process carried out manually so far?* The analyses help answer that. First of all, we have seen how it is possible to construct a conceptual mapping between criteria adopted by two different organizations, also highlighting the degree of correspondence between criteria (i.e., strong, weak, partial). Second, we analyzed which criteria were evaluated the same way by the organizations, on the same set of news media.

Both the existence of the mapping between criteria and the agreement on their evaluations aid the choice of which criteria/features might be worth automating. We believe that a good starting point is to consider those criteria for which i) there is strong mapping (i.e., CML 4 and 3) and ii) the evaluations are more in agreement (e.g., the agreement is at least on 2/3 of the media that, reported in percentage, is equal to approximately 67%.

In Table 19 we have highlighted the GDI criteria which, in relation to the thresholds we have set (CML $\geq$ 3, agreement percentage $\geq$ 67), appear as good candidates to be automatically computed. Obviously, the thresholds were set by us in a rational but arbitrary manner, and different thresholds may be considered in the future.

Criteria matching and agreement in their evaluation has also another implication. Suppose that, using state-of-the-art NLP tools, we automatically compute the values of some GDI criteria such as, for example, the use of sensational language in the news and the accuracy of the headline (SensLang and HeadAcc in Table 19). The fact of having correspondence and agreement in the evaluation with the NG criterion 'Avoid Deceptive Headlines' (see both the mapping in Table 18 and the agreement figures –Figure 22 and 21) means that the value of 'Avoid Deceptive Headlines' can be taken as a reference, even as ground truth, to evaluate the goodness of the result of the automatic evaluation of the corresponding GDI criteria.

Finally, we would like to point out that our choice to start with the automatic computation of GDI criteria is due to the fact that in our previous work on the assessment of the media market in Italy we have gained a good understanding of the meaning of these criteria and how to evaluate them.

5.1.8 Conclusions and limitations

Driven by the curiosity to understand if and to what extent a process of evaluating a news source can be made automatic, in this analysis we compared 2 journalistic procedures currently in use to classify new sites as reliable or not. For the aspect concerning the published content, we found a good correspondence in the evaluations of the criteria taken into consideration by the two procedures. The situation is different for the criteria regarding the transparency of the source as a whole. This opens the way to new investigations, such as, for example, on the possibility of calibrating ad hoc crowd-sourcing campaigns hybrid human-LLMs annotators to better evaluate those aspects that find discordant evaluation.

Limitations: In this work, we have explored the -mostly manual- procedures performed by organizations experienced in journalism and media communication to assess the degree of transparency and credibility of a news source. The analysis was carried out with the ultimate aim of 1) emerging a set of criteria on which to focus a first automatic computation process; 2) better investigating the reasons why on some criteria there is no agreement. Obviously, this work is not without limitations. First of all, the analysed procedures are only two, as well as the news media are Italian and small in number. Given the difference in classification of the two agencies (NG has two rating levels, GDI has 5), we opted for a binary classification in the comparison, losing granularity. However, setting arbitrary thresholds to split the NewsGuard ranking into several parts seemed to us too arbitrary. In constructing the conceptual mapping between criteria, it was not always possible to establish a strong connection. In addition, the lack of agreement we found may be due to: 1) a different process of implementation of the criterion, dictated by the organizations themselves; 2) an inherent bias in the annotators (e.g., the study on the media market in Italy was conducted by computer scientists, the study on the same media by NewsGuard may have been conducted by a group of annotators with a different background, e.g., journalists); 3) the articles chosen to assess the credibility aspects were obviously not the same in the two studies; 4) the study of the news website has been realized at different times (its content may have changed and the two assessments may have been affected by this change).

Aware of these limitations, we believe that it is noteworthy that the two procedures lead to the same outcomes, at least in terms of news content. Furthermore, it is interesting that criteria such as, e.g., the existence of declarations of independence, or declarations on ownership and editorial structure of the media are the criteria for which the two procedures gave opposite results. A good line of inquiry for the future is to study how different annotation campaigns lead to the same results for news source evaluation. We argue that the outcome of that analysis will improve current assessment procedures.

5.2 Evaluation of reliability criteria for news publishers with Large Language Models

In this Section, we investigate the use of a large language model to assist in the evaluation of the reliability of the vast number of existing online news publishers, addressing the impracticality of relying solely on human expert annotators for this task. In the context of the Italian news media market, we first task the model with evaluating expert-designed reliability criteria using a representative sample of news articles. We then compare the model's answers with those of human experts. The dataset consists of 352 news articles annotated by three human experts and the LLM. Examining 6,081 annotations over six criteria, we observe good agreement between LLM and human annotators in three evaluated criteria, including the critical ability to detect instances where a text negatively targets an entity or individual. For two additional criteria, such as the detection of sensational language and the recognition of bias in news content, LLMs generate fair annotations, albeit with certain trade-offs. Furthermore, we show that the LLM is able to help resolve disagreements among human experts, especially in tasks such as identifying cases of negative targeting.

5.2.1 Problem formulation and contributions

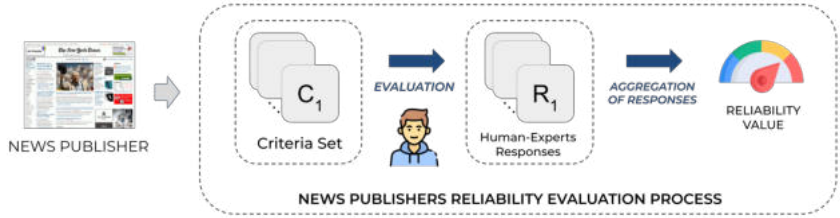


Figure 23: Traditional Approach to Evaluating the Reliability of a News Publisher

As anticipated in Section 2.2, each organisation conducts the eval-

uation process by defining a specific set of criteria and identifying the expert profiles required for their assessment (e.g., journalists or trained annotators). Criteria typically consider elements like the inclination to publish propaganda or politically biased content [8, 113]. The phase of the criteria evaluation process generally lacks automation. This process can be modeled as depicted in Figure 23. The organization in charge of evaluating publishers sets criteria for "good journalism" and assigns experts in the field to manually evaluate the publisher against these criteria. The evaluation results are aggregated at the criterion level, taking into account potential disagreements among experts, to produce a final score - often numerical in nature. This score serves as a holistic measure of the publisher's overall reliability, providing a comprehensive assessment without detailing the specific contributions of individual criteria. Although these scores offer valuable insights [38, 113, 161], the process of evaluating individual news outlets is labor-intensive and time-consuming. Manual assessments, where expert annotators scrutinize ownership information and content, remain crucial yet demanding [80, 146].

In recent years, Large Language Models (LLMs) have excelled at generating text that closely mimics human language [108]. Their capabilities extend to various natural language processing tasks, such as sentiment analysis and text summarization [215]. This capability has paved the way for numerous potential applications, including improving educational processes and providing answers to medical questions [7].

On the one hand, we have an accurate but time-consuming evaluation process: it takes months to evaluate a single news source, starting with the selection and training of annotators, through the actual evaluation of specific criteria, and ending with the calculation of the final score, not to mention that after months the situation of the media outlet itself may have changed and require a new analysis. On the other hand, a generative intelligence whose evolution seems to progress constantly, see, e.g., the recent release of GPT-4o (omni)[142].

Objective and Approach. The main goal of this work is to develop a more scalable and more efficient trust evaluation process by incorpo-

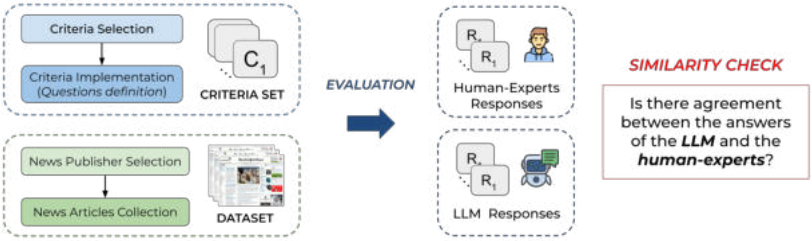


Figure 24: The core idea: Given a set of criteria and a set of news articles, we evaluate the agreement between the manual evaluation of the criteria and the automated evaluation to determine whether automated evaluation is a valid approach to assessing reliability of the articles' publishers.

rating automation into the traditional approach of manually evaluating good journalism criteria defined by specialized organizations, see Figure 23. To the best of our knowledge, this is the first study that proposes to introduce automation into the task of evaluating such criteria. To conduct the experiment, as shown in Figure 24 that represent the core idea of the work, we use a two-step process. In the first step, we perform the two following tasks in parallel: (i) the selection and implementation of a set of quality criteria, and (ii) the collection of a representative sample of publishers and news articles. In the second step, three experienced human annotators evaluate the criteria on each of the sampled articles. Their answers are compared with those of an LLM. The intuition underlying this study is that agreement between human experts and the LLM on individual criteria leaves room for automation of related subtasks. In addition, we explore the potential of the LLM to assist in resolving disagreements among human experts.

Research Questions. Achieving our goal translates into answering the following questions:

- **RQ1:** How to define quality prompts to help LLMs correctly evaluate good journalism criteria to assess the reliability level of a publisher?

- **RQ2:** How good is an LLM at aligning with the responses given by

human-experts when evaluating the good journalism criteria?

- **RQ3:** What are the types of questions for which an LLM can effectively support, dare we say even replace, an expert annotator in the process of evaluating the reliability of a publisher?

Main Contributions. Through a process of refinement, we are able to write prompts that are optimized so that 1) the LLM understands the questions it needs to answer, and 2) the answers are consistent with those of human annotators. The results show agreement between the LLM and the annotators in recognizing the type of the news, the presence of a fact-based *lede* at the beginning of the article and whether the text negatively targets someone or something. Also, the evaluation of labor-intensive criteria, such as checking for the presence of article bias and sensational language, achieves a fair level of agreement, albeit with certain trade-offs. In addition, the LLM can provide valuable support in cases of disagreement between human experts in different evaluation tasks.

Applications. The ability to automatically generate high-quality, criteria-specific responses from the analysis of news articles opens up at least two potential avenues of application:

- On an organizational level, automated evaluation of criteria can be integrated into existing workflows for assessing the reliability of news publishers. This is particularly effective in resource-constrained environments where automation provides scalability. For example, assessing labor-intensive criteria on a large volume of news articles is often not feasible due to limited expertise or funding. Automation, on the other hand, enables large-scale, near-real-time monitoring and analysis, significantly improving the efficiency and reach of reliability assessments.

- At the user level, criteria-specific responses can increase readers' awareness of the content they consume in real time. The goal is to highlight potential deviations in specific aspects (e.g., sensational language) from established quality standards of good journalism. This approach translates expert practices into tools accessible to

end-users, fostering critical thinking and enabling more informed and reflective engagement with news content.

5.2.2 Framework overview

This section presents our methodology. The goal is to define appropriate prompts for the LLM to evaluate journalistic criteria used to assess the reliability of online publishers, and measure the agreement between the LLM's answers and those of experienced annotators.

We start with the selection (Section 5.2.3) and implementation (Section 5.2.4) of a set of criteria for evaluating the reliability of online news publishers (upper left blue dotted box in Figure 24). As explained in 5.2.4, in the context of this study, criteria implementation means defining a list of questions that can be evaluated by both a human expert and an LLM-based annotator. Each question aims to highlight a specific aspect of a news item according to the corresponding criterion. Second, as described in Section 5.2.5, we proceed to select a set of publishers to evaluate and extract a sample of news articles from those publishers (lower left green dotted box in Figure 24). Then, in 5.2.6, we ask three experienced annotators and the LLM to answer the questions representing the criteria. Finally, in the *Results* section, in case of agreements between human experts we evaluate the agreement between the experts' answer and the LLM answer (right red box in Figure 24). In cases where there is disagreement between experts, we explore the potential of the LLM to facilitate resolution of these conflicts.

5.2.3 Criteria selection

The goal of this section is to select a *worthy-to-automate* list of criteria. As mentioned earlier, these criteria must be both widely recognized and labor intensive to evaluate manually.

The starting point for this analysis is the list of criteria proposed by [152], presented in the previous Section. This work identifies, from the complete list of GDI criteria [148], a preliminary set of *worthy-to-automate* criteria. A detailed explanation of these criteria can be found in Ta-

ble 33. These criteria are selected because, although they have different names, and their definitions do not match NewsGuard's definition, there is a strong conceptual correspondence between them [152]. The NewsGuard list of criteria is particularly noteworthy, as it is widely used [101]) too. This consistency in definition across both organizations assures that these criteria are widely accepted as effective for conducting real-world reliability assessments of news publishers.

The tests in [152] were conducted on the Italian media market, which is also the focus of this study. For the scope of this paper, however, we have excluded some criteria from the original set. In particular, we will not consider in our analysis the criteria that do not involve the analysis of a written text (like *VisPres*) and the (trivial-to-evaluate) features that can be extracted using heuristic methods (*ByInfo*). Also, *Common Coverage* is excluded due to their reliance on multi-source analysis. The last six criteria in Table 33 cover broader editorial and operational aspects of news outlets, such as editorial independence, funding sources, and ownership structure. These require access to meta-information that is not typically part of the article text, making them less amenable to automated evaluation by an LLM.

The resulting set of *worthy-to-automate* criteria is as follows:

- *Headline Accuracy*: Assesses the congruence between a headline and the corresponding article content.
- *Lede Presence*: Checks for the presence of a factual summary at the beginning of an article.
- *Negative Targeting*: Evaluates whether the article negatively targets specific individuals or institutions.
- *Article Bias*: Gauges the fairness and balance of the article's content.
- *Sensational Language*: Identifies emotionally charged or exaggerated language that could mislead readers.

5.2.4 Criteria implementation

In this paper, implementing a criterion means defining a clear and concise question, along with potential responses, that can be effectively evaluated by both human experts and LLM-based annotators, using the text of a news article as input. Table 20 details how we have implemented the criteria selected in Section 5.2.3. Most questions closely align with the descriptions of the criteria provided by GDI. For example, GDI criterion *Negative Targeting*: "*Rating of whether the story negatively targets a specific individual or group*" is made concrete by the question: "*Does the article negatively target individuals or groups?*". We also instruct the model to specify the reason why an individual or group is negatively targeted (see "NegTarg (Identification)" in Table 20). This refinement allows us to assess the ability of the model to identify specific groups or minorities that are negatively targeted in the news. Throughout the manuscript, the general term "NegTarg" refers to "NegTarg (Detection)", which denotes the ability to detect textual content that negatively targets groups or individuals. In Section 5.2.8, we explicitly distinguish between Detection and Identification responses.

In addition to the selected criteria reported in section 5.2.3, we also introduce a meta-criterion (*Type*) to assess whether the LLM can accurately determine the type of article it is analyzing, e.g., distinguishing between straight news and other journalistic forms such as editorials, satire or soft news [206].

5.2.5 Articles Dataset

This study focuses on the Italian media market. To ensure an unbiased analysis, it is essential to gather textual data from articles that best represent the Italian online media landscape. The approach adopted involves collecting articles from a list of representative Italian news publishers. Specifically, a past study about the Italian online news market [148] proposed a balanced list of Italian news publishers, considering: (i) the media outlet distribution (national or local), (ii) the geographic location of the media outlet, (iii) the political orientation, and (iv) the disinformation

Table 20: Selected criteria with their short names and implementations

Criteria Name	Short name	Implementation	Answers' Options
Headline accuracy	HeadAcc	How accurate is the news's headline with the content of the news?	Inaccurate/Quite inaccurate/Quite accurate/Accurate
Lede present	LedePres	Does the article start with a summary of the main facts?	Yes/No
Negative targeting	NegTarg	*(Detection)* Does the article negatively target individuals or groups?	Yes/No
		(Identification) Indicate what issue the group or individual is negatively targeted on.	issue (Politics / Gender / Religion / Other)
Article bias	ArtBias	How much biased is the article?	Biased/Quite biased/Quite unbiased/Unbiased
Sensational language	SensLang	How sensational is the tone of the news?	Sensational/Quite sensational/Quite neutral/Neutral
Meta-criterion			
Type of news	Type	What kind of news are you reading?	Straight news/Editorial/Investigation/Satire/Soft News

risk scores. Table 31 (Appendix C) shows the online publishers selected in [148]. We have included these 34 publishers in this study because they have already been selected to be representative of a country's news market.

Furthermore, for the current analysis we have collected news articles that were published at the same time as the articles in the original study. This is because the quality of a newspaper can vary over time, and we wanted to try to replicate the same distribution of news articles, both by source and risk of exposure to disinformation. Thus, for each of the 34 news publishers, we retrieved a mean of 10 articles published within a 7-month period, from April to October 2021 for a total of 352 news articles, which is consistent with the timeframe and sample size used in the GDI analysis.

We would like to emphasize that the primary goal of this study is not to evaluate the reliability of a single publisher. Instead, our focus is on assessing the ability of a large language model to evaluate specific journalistic quality criteria applied to a curated collection of article texts.

The 352 article texts were collected using an automated pipeline combining several tools. First, we employed the Selenium library [178] to gather URLs from publishers' websites. Next, we utilized the GNU Wget command [205] to download each article's HTML page. Finally, XPATH

Table 21: Criteria implementation refinement

Short Name	Question-Level Modification	Response-Level Modification	Description
HeadAcc	Rephrased for clarity	None	-
LedePres	Rephrased for clarity	None	Introduced the definition of lede into the question
NegTarg	None	None	-
ArtBias	None	Response options simplified to Biased/Unbiased	-
SensLang	None	Response options simplified to Sens/Neutral	-
Meta-criterion			
Type	None	Response option changed from *"Editorial"* to *"Opinion"*	-

queries were used to extract the textual content, including the title and main text of each article.

5.2.6 Criteria evaluation

This section introduces the human annotators and explains their annotation process. In addition, the section presents the design of the LLM-annotator prompts.

Human Expert Annotations. Initially, we sanitized the text of the articles to ensure anonymity of publishers and authors to prevent biasing the annotators' judgments. Then, we asked two experts from the Media and Communication Unit of our department, and a Ph.D. student in Data Science, to annotate the articles. Articles, questions, and options for answers were uploaded to the Google Drive platform. Each article was annotated by two of the three annotators. The annotators were able to consult with each other when in doubt: This is because we want their annotation to be a ground truth against which we can later compare LLM responses.

Prompt Design. The prompt design process ends with the formulation of the questions shown in Table 20. For each of the articles, we ask the LLM the questions listed in the table (column *Implementation*) and give some options for the answers (column *Answers' Options*). It is important to note that since our dataset consists of Italian-language news, the questions we pose to the LLM and the answers we expect are in Italian. In Table 20, we have provided the English translation for the convenience of the reader. The validity of the answers is compared to the answers of the expert annotators, which we consider to be ground truth. The agreement, measured by Cohen's Kappa [47], between the model's

answers and those of the experts provides a measure of the quality of the prompts.

During the prompt design phase, we analyze the cases with the most disagreements and identify three main motivations: (i) the prompt wording is not adequate for correct processing by the LLM, (ii) there are errors in the experts' annotations, and (iii) the LLM has difficulty interpreting certain contexts or questions correctly.

For prompts categorized under (i), we reviewed the initial versions of questions and answers until the latter were consistent with one of the answer options. Consistency was evaluated on a selected subset of news articles.

In modifying the prompts, both at the *question* level and the *answer* level, we preserved the semantics so that the validity of the answers given by human annotators remain unchanged, while improving the model's interpretive capabilities. Table 21 shows how we refined the prompts.

The *HeadAcc* criterion was rephrased to enhance clarity. Similarly, the question for the *LedePres* criterion was revised to explicitly incorporate the definition of an article's *lede* within the question itself.

Some response options remain unchanged. Specifically, *HeadAcc*, *LedePres*, and *NegTarg* continue to be binary responses. For *ArtBias* and *SensLang*, as detailed in Section 5.2.8, converting the original responses into binary formats enhances the consistency and reliability of the scoring process. Regarding the *Type* meta-criterion, modifying the term *"Editorial"* to *"Opinion"* has proven to improve the quality of the responses.

After testing the new prompts on a subset of the news articles and manually checking the consistency of LLM's responses, we repeated the experiment on the entire dataset. In the cases of *ArtBias* and *SensLang*, to implement the changes outlined in Table 21, we reduced the original four-class responses to two classes. This remapping allows us to maintain both datasets without the need for additional data collection.

During the prompt refinement process, we addressed the cases that fell under category (i). The improvements achieved through this process and the ability of the LLM to align with human-annotators are detailed

in section 5.2.8. For cases categorized under (ii) and (iii), the observed issues stemmed from either human annotation errors or the intrinsic limitations of the model. In Section 5.2.8, we investigate scenarios where the LLM can support in resolving disagreements between experts.

Collection of LLM annotations: The model is instructed to assume the persona of an experienced journalist to ensure that the answers are appropriate for that role. Each question was asked to LLM three times, this was to check that it was not answering inconsistently, perhaps as a result of hallucinations. The answers given by LLM were always the same for the three rounds, thus, a single answer was produced for each article and criterion.

Annotated dataset of human and LLM responses: At the end of this process, we obtained a total of 6,081 annotations across the six evaluation criteria (see Section 5.2.3) for the dataset of 352 news articles (see Section 5.2.5). In addition, as described in Section 5.2.4, we conducted a more detailed analysis of one specific criterion—negative targeting—by introducing target identification. This refinement resulted in an additional 1,011 annotations (see "NegTarg (Identification)" in Section 5.2.8).

5.2.7 Experimental setup

Agreement Metric: We use Cohen's Kappa [47] to measure the agreement between annotators. The formula to calculate Cohen's Kappa is as $k = (p_o - p_e) \div (1 - p_e)$ where p_o is the observed probability of agreement between the annotators and p_e is the expected probability of agreement if the annotators assigned the labels randomly [6]. We calculate this metric using the implementation available in the scikit-learn library[177]. As for the interpretation of values, we refer to the descriptions by Mary L. McHugh [129], which are based on Jacob Cohen's original interpretations. A kappa value between 0 and 0.20 indicates no agreement among raters, beyond what would be expected by chance. Values from 0.21 to 0.39 represent minimal agreement. A weak level of agreement is observed for values between 0.40 and 0.59, while values between 0.60 and 0.79 indicate a moderate agreement. A strong agreement is denoted by

kappa values from 0.80 to 0.90. Finally, values above 0.90 suggest an almost perfect agreement between the raters.
Model selection and configuration: As LLM, we selected ChatGPT-4o, the most recent model released by OpenAI at the time of the experiments [99]. We configure the model's hyperparameters to optimize the quality of its output. The temperature is set to zero because we need consistent and predictable responses. The *Maximum Length* parameter has not been used because article lengths vary widely and a fixed limit would not be appropriate in all cases. Both *Frequency Penalty* and *Presence Penalty* are set to their default values (0) because avoiding repetition is not a priority for our task.The *Stop Sequences* parameter is set to *None* because we do not need to stop text generation at a particular end sequence.

5.2.8 Results

First, we see how well the LLM's responses match those of the human annotators, considering the original prompts in Table 20. Then, in cases of strong disagreement, we see if using the refined prompts (Table 21) makes things better. As a third analysis, we look at the cases where the human annotators disagree to see if the LLM could provide support.

Agreement between experts and LLM

In Figure 25, we report the agreement between LLM and human annotators *for only those cases where human annotators agree with each other*. Essentially, for each article and criterion, the single annotation value determined by human experts (equal to the two concordant annotation values) is compared with the response provided by the LLM.

For 3 of the 6 criteria analyzed, the agreement between LLM and experts exceeds (or equals, in the case of LedePres) the minimum acceptance threshold set at 0.4. On the other hand, when assessing the degree of sensationalism in the article or the presence of bias, LLM disagrees with the expert annotators, who let us remember that it is the ground truth for us. Specifically, we obtain $k = 0.2064$ for *ArtBias* and $k = 0.1732$ for *SensLang*. We get the highest value of the agreement for *NegTarg*, $k =$

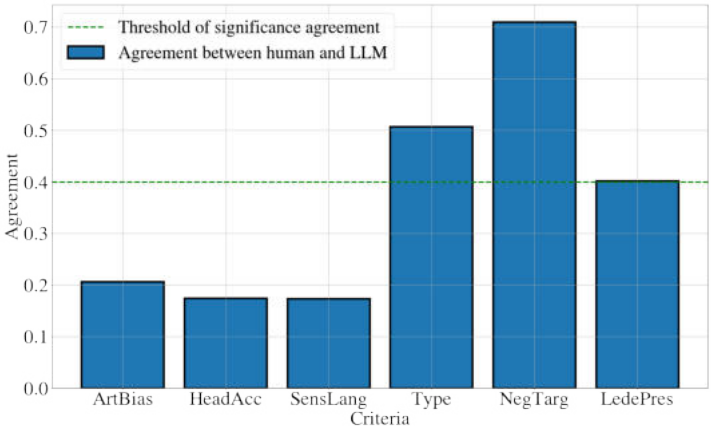


Figure 25: The average agreement between experts and the LLM (computed considering only cases in which the experts themselves reached a consensus).

0.7089. Without prompt modification, the LLM is very careful to detect whether a news text is negatively targeting a group or a person.

We went on to analyze in more detail what happens to the LLM when it has to evaluate the criteria *SensLang* and *ArtBias*.

Analyzing the confusion matrix of *SensLang* (Figure 26) and *ArtBias* (Figure 27), we see that the LLM is more sensitive (i.e., it detects sensational language and bias where human annotators do not); however, it makes errors mainly in neighboring classes. In both confusion matrices, 1 means maximum degree of sensationalism/ bias, and 4 means no sensationalism/no bias. Looking at the case of sensationalism, for example, we see that for experts, the articles read are mostly neutral (10+96+82), while for LLM, only (82+2) have no trace of sensationalism. This is generally true for both matrices: LLM finds higher levels of sensationalism and bias than humans.

Figure 28 shows the gain in agreement when we use the refined prompt from Table 21 for the same two criteria, *ArtBias* and *SensLang*. The ques-

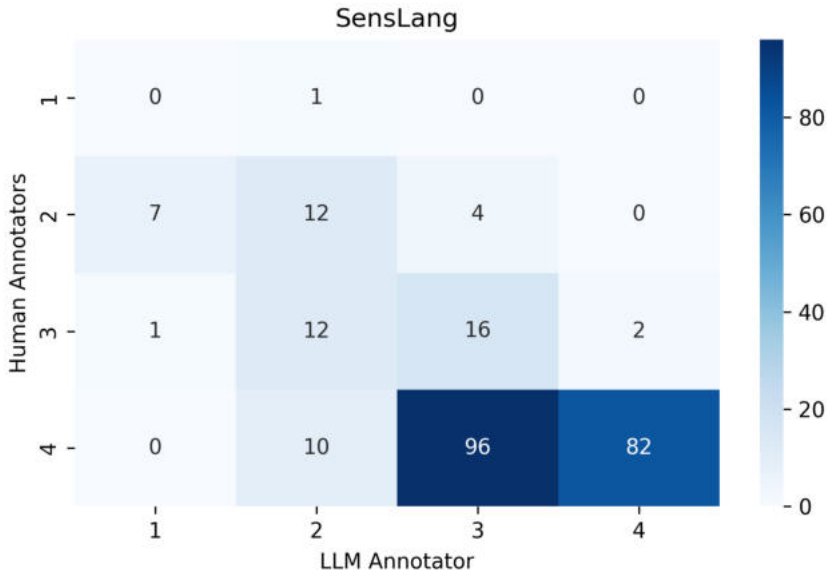


Figure 26: Confusion Matrix for Criterion: Sensational Language. 1: Sensational; 4: Neutral

tion remains unchanged, while the responses have become binary: Biased/Unbiased and Sensational/Neutral. What happens with the refined prompt is that k goes from 0.2064 to 0.4750 for *ArtBias* and from 0.1732 to 0.5486 for *SensLang*.

Resolution of disagreements between human experts

In the previous section, we focused on cases in which human annotators agreed in their evaluations. In this section, we instead analyze cases where annotators disagree, in order to investigate whether an LLM can assist in resolving such disagreements.

To address these cases, we conducted an ex-post human evaluation aimed at establishing a reference ground truth. Specifically, all instances of disagreement between the two original annotators were independently re-evaluated by a third expert annotator. The judgment provided in this

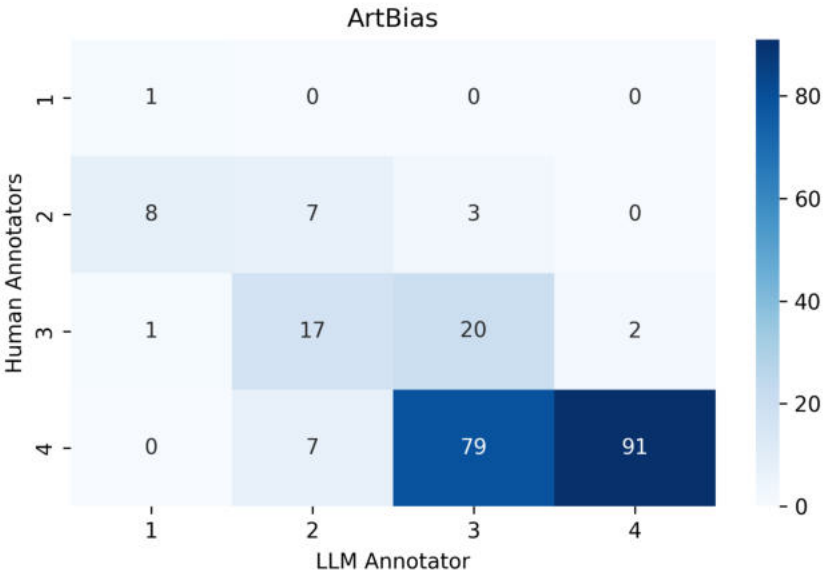


Figure 27: Confusion Matrix for Criterion: Article Bias. 1: Biased; 4: Unbiased

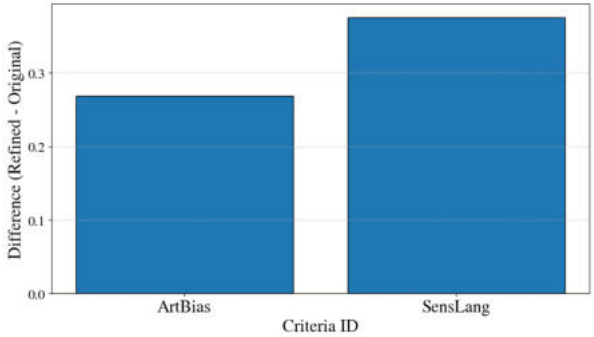


Figure 28: Comparison of agreement levels between the *refined* and *initial* implementations of the criteria, considering only ArtBias and SensLang.

ex-post evaluation is used as the ground truth for the analysis. Cases in which even this additional evaluation did not lead to a definitive judg-

ment are labeled as borderline cases.

We focus on the criteria *ArtBias*, *SensLang*, *NegTarg*, and *HeadAcc*, as these are the criteria for which the annotators disagreed most frequently. Furthermore, we only consider cases of relevant disagreement between annotators. Specifically, relevant disagreements correspond to cases where annotators provided clearly different evaluations: for binary criteria (e.g., *NegTarg*), this corresponds to cases where one annotator answered yes and the other no; for degree-based criteria (e.g., *ArtBias* or *SensLang*), this corresponds to cases where the two evaluations differ by at least two levels on the scale (e.g., one annotator selects Inaccurate while the other selects Quite Accurate).

Table 22 reports, for the considered criteria: (1) the total number of articles, (2) the number of articles where the two annotators disagreed, (3) the number of relevant disagreements, (4) the number of cases where the LLM prediction coincides with the ex-post ground truth, and (5) the number of borderline cases. In this context, we say that the LLM provides the correct answer when its output coincides with the evaluation given in the ex-post ground truth.

Overall, 226 articles present disagreement between the original human annotators on at least one criterion. The table shows that, excluding borderline cases, the LLM resolves 100% of disagreements for *ArtBias* and *NegTarg*, 81% for *HeadAcc*, and 63% for *SensLang*. The 100% result for the two *NegTarg* criteria is due to the fact that the LLM always provides the same answer as the ex-post ground truth in the non-borderline cases.

These results highlight the potential of LLMs to assist human annotators in resolving disagreements when evaluating these criteria.

For the degree-based criterion *ArtBias*, we observe that the LLM often resolves disagreements by selecting an intermediate evaluation between those provided by the human annotators. For example, if one annotator selects Biased and the other Quite Unbiased, the resolution corresponds to the intermediate evaluation Quite Biased, which in several cases coincides with the ex-post ground truth.

For *NegTarg*, the LLM shows a strong ability to identify whether a

Table 22: Statistics about disagreement between expert annotators. For each criterion we report: (1) the total number of analyzed articles, (2) the number of articles where the two annotators disagreed, (3) the number of relevant disagreements, (4) the number of cases where the LLM prediction coincides with the ex-post ground truth, and (5) the number of borderline cases.

Criterion	Articles	Disagreements	Relevant	LLM = GT	Borderline
ArtBias	340	79	**4**	4	0
HeadAcc	340	108	**11**	9	0
NegTarg (Detection)	340	30	**30**	18	12
NegTarg (Identification)	340	47	**47**	32	15
SensLang	340	72	**11**	7	0

group or individual is negatively targeted within an article. However, there are several cases where even human experts find it difficult to reach a definitive judgment, as reflected by the presence of borderline cases. The LLM shows similar proficiency when tasked with identifying the specific subject of negative targeting.

5.2.9 Discussion

In this study, we investigated whether and to what extent one Large Language Model can automatically evaluate criteria that journalistic organizations use in real-world evaluations to estimate reliability of online publishers. We adopt a subset of criteria used by the Global Disinformation Index (and very similar to some used by NewsGuard) to assess the risk of an online publisher exposing its readers to misinformation. The first of the research questions, **RQ1**, is whether it is possible to define quality prompts to help LLMs correctly evaluate good journalism criteria to assess the reliability level of a publisher. The second of the research questions, **RQ2**, is how good is the LLM at aligning with the responses given by human-experts, in the same context as before. To answer these questions, we implemented an iterative refinement of the prompts. This process allowed us to improve some of the initial prompt definitions for better processing by the LLM. As described in Section 5.2.8, among the six criteria analyzed, the evaluation of three of them was already in good agreement with that of human experts. Specifically, in recognizing the

nature of a news story, in recognizing the presence of a lede, and in understanding whether a text is negatively targeting something or someone. When it came to recognizing sensational or biased language, the initial prompts resulted in very low levels of agreement between the humans and the LLM. Upon deeper analysis of these results, we observe that LLMs show increased sensitivity in detecting bias and sensationalism compared to human experts. In addition, most errors are confined to neighboring classes, which mitigates the severity of prediction inaccuracies. In particular, the use of refined prompts (offering binary responses instead of four options) significantly increases agreement, especially in detecting sensational language.

RQ3 asks what types of questions LLMs can answer correctly and thus support the human annotator in the publisher's evaluation process. To address this third question, we first note that in scenarios where increased annotation sensitivity is acceptable or a reduction in predictive granularity is warranted (e.g., considering binary instead of multiclass responses), the LLM-based annotator provides good quality annotations. As anticipated, this is particularly evident for criteria requiring significant effort, such as detecting the use of sensational language or assessing article bias in the text. Even better when it comes to judging whether the article is negatively targeting someone or something (for this criterion, *NegTarg*, Detection, we got the highest k of 0.7089).

In addition, an LLM-based annotator can offer valuable assistance in resolving disagreements among human experts. Specifically, the LLM demonstrates promise in supporting decisions concerning *ArtBias*, *HeadAcc*, and *NegTarg*. This includes detecting negatively targeted groups or individuals and accurately identifying them —tasks that are often complex and may not lead to definitive conclusions. In particular, accurate assessment of negative targeting of individuals or minority groups is critical to combating the spread of adversarial narratives [63]. The significant alignment achieved in a zero-shot experimental setting underscores the potential of the LLM for further improvement through tuning and adaptation.

5.2.10 Conclusions and limitations

We explored the use of one LLM in zero-shot settings to evaluate a set of criteria designed to assess online journalism standards. The LLM achieves substantial levels of agreement with human experts on complex evaluations, such as understanding whether the text negatively targets someone or something. In addition, the LLM is highly effective at resolving disagreements among human experts in tasks such as detecting negative targeting and determining bias and headline accuracy in articles.

While our study provides valuable insights into the use of Large Language Models (LLMs) to support human judgment in evaluating news quality, it is subject to certain limitations. However, we believe that these limitations do not compromise the core objective of our work, which is to demonstrate that LLMs can be a promising tool for this task. A primary limitation concerns the scope of our dataset, which is relatively limited in terms of both size and linguistic coverage, as it only includes Italian news articles. An expansion of this dataset would allow a more comprehensive evaluation of the proposed methodology in different linguistic contexts. Furthermore, although we have considered a number of relevant reliability criteria, our evaluation does not cover all possible aspects of news quality assessment. For example, visual elements such as the likelihood of an image being automatically generated or having sensationalist characteristics were not included.

Furthermore, while our experiments focused on a single LLM, GPT-4o, to demonstrate the feasibility of our approach, a broader investigation of the performance of different models—including both proprietary and open-source alternatives of different sizes—would be beneficial. Moreover, alternative prompting and retrieval strategies, such as Retrieval-Augmented Generation (RAG) [118], one-shot, and few-shot learning, could further enhance the robustness and adaptability of the methodology. Finally, in our experiments the prompts were provided only in the original language of the analyzed articles. Future work could explore the robustness of the approach by testing prompts in different

languages (e.g., English), which could further strengthen the validity and generalizability of the results.

Building on this foundation, future work will focus on extending the scope of our experiments in several key directions. First, we plan to extend the evaluation criteria to include visual aspects, thereby capturing a more holistic representation of news quality. Second, we aim to increase the number of news outlets and articles analyzed in order to improve the robustness and generalizability of our findings. In addition, we will explore the applicability of our approach across multiple languages to evaluate its effectiveness in different linguistic environments. Finally, we intend to experiment with a wider range of LLMs and integrate advanced techniques such as RAG and few-shot learning to refine our evaluation framework and further improve performance.

Potential data contamination A limitation of studies involving proprietary large language models concerns the possibility of training data contamination, i.e., situations in which evaluation data may have been present in the model's training corpus. Since GPT-4o was trained on large-scale web data, it is possible that some of the news articles analyzed in this study were part of its training distribution. However, the annotations used in our experiments were produced specifically for this study—either by human annotators or through synthetic generation based on the selected criteria—and were not publicly available. Consequently, it is highly unlikely that the model had access to the ground-truth labels associated with the articles. As a result, even in the presence of potential exposure to the article text, the model would still need to infer the quality assessments from the content and the provided evaluation criteria. Nevertheless, we acknowledge that prior exposure to similar articles or domains could influence the model's performance. Future work could further mitigate this risk by evaluating models on newly collected or temporally separated datasets that are unlikely to appear in training corpora.

5.2.11 Ethical considerations

Our research involves the use of LLMs to evaluate the reliability of online news publishers, a task which includes a step of annotating news articles according to journalistic criteria, usually done by human annotators. While the ability of LLMs to assist in such evaluations offers some advantages, it is imperative to recognize and address the potential ethical implications of this technology.

- Bias and Discrimination: Given that LLMs are trained on large datasets, they may inherently carry biases present in their training materials. In the context of news evaluation, such biases could affect the impartiality of trustworthiness assessments, potentially perpetuating stereotypes or unfairly targeting certain publishers based on biased training data.

- Transparency and Accountability: The decision-making process of LLMs is typically complex and not fully transparent, making it difficult to understand how or why certain scores are generated. This lack of clarity can be particularly problematic in scenarios where decisions need to be justified or contested.

- Environmental Impact: The computational demands of training and running LLMs are substantial, contributing to significant energy consumption and environmental impact. This is a critical consideration given the global push to reduce carbon footprints. In addition, when using proprietary models such as GPT-4o through API access, the underlying computational infrastructure remains opaque, making it difficult to estimate the actual GPU usage and energy consumption associated with the experiments. While the financial cost of API usage can be measured—in the case of the experiments conducted in this study, approximately 72 USD—the environmental cost of the computation cannot be directly assessed. For this reason, future research will also explore the use of open-source language models. Although open models do not inherently guarantee lower environmental impact, they provide greater trans-

parency and control over the computational infrastructure used to run them. This allows researchers to more accurately monitor resource consumption and to experiment with more efficient deployment strategies, such as model compression, quantization, or the use of smaller architectures. Such transparency may facilitate the development of evaluation pipelines that are not only effective but also more environmentally responsible.

In using LLMs to support the evaluation of news publishers' reliability, it is important not only to harness their potential to improve efficiency and accuracy, but also to rigorously address these ethical concerns. By actively working to mitigate bias, increase transparency, and reduce environmental impact, we can better harness the power of LLMs in a way that is consistent with ethical standards and societal expectations.

Chapter 6

Social interactions analysis for Trustworthiness rating and the development of a first SW prototype

This chapter presents the final component of the second methodological contribution of the thesis. Building on Chapter 4, Section 6.1 introduces a novel approach to estimate the trustworthiness of news publishers without relying on textual content. Instead, the method exploits social interaction patterns and a set of pre-evaluated sources, offering a scalable and cost-effective alternative to content-based and manual evaluations, particularly suitable for early-stage and low-resource scenarios. To demonstrate its applicability, Subsection 6.1.6 presents TROPIC, a software prototype designed to support expert annotators by providing preliminary trustworthiness indicators based on interaction signals. Finally, Section 6.2 introduces a complementary content-based approach for source credibility estimation at the article level, enabling a more fine-grained analysis and enriching the overall framework developed in this thesis.

6.1 Social interaction patterns as trustworthiness indicators

In this section of the thesis, we introduce a novel framework designed to assess the trustworthiness of online news publishers by analysing user interactions on social media platforms. This approach enables us to circumvent the principal challenges typically associated with the manual process of estimating publisher trust (see Section 2.2). While this method involves some trade-offs in terms of accuracy and cost, it offers a versatile solution that fulfils a dual purpose: (i) identifying verifiable online publishers and (ii) automatically providing an initial trustworthiness assessment for previously unclassified online news outlets.

6.1.1 Problem formulation and contributions

The research presented in this Section seeks to enhance the evaluation process by leveraging interactions on social media. Given the pivotal role of online platforms in shaping public discourse, our methodology is designed to bypass the limitations inherent in traditional (manual) evaluation methods. Through a detailed analysis of the dynamics of social media interactions associated with news publishers, we aim to attain a deeper understanding of trustworthiness within the digital information landscape.

The central question of our study, detailed in [157] and discussed in this thesis section, is: "Can we leverage social interactions to estimate publishers' trustworthiness?". In other words, can we design a framework that uses social interactions to discern between low- and high-quality publishers?

With the aim to shed light on several further aspects connected with this central question, we designed three research questions to answer:

- **RQ1:** Can we use the social interactions between online news publishers and consumers to assign a trustworthy label to the publishers?;

- **RQ2:** If the answer to RQ1 is yes, how many of the publishers with whom consumers interact can we assign a trust label? (In other words, does the proposed methodology provide sufficient coverage?);

- **RQ3:** If the answer to RQ1 is yes, what is the classification performance?

Contributions: In addressing these questions, we have developed a novel framework for assessing the trustworthiness of online publishers based exclusively on the interactions observed between users and URLs in data harvested from active social media discussions. Our study primarily focuses on Twitter/X as the reference platform. The main contributions can be summarized as follows:

- **Automatic Classification of Publishers' Trustworthiness:** We present a novel method for automatically classifying the trustworthiness of online publishers, exploiting the interactions within social media discussions between two central actors: publishers (news producers) and users (news consumers). The method is designed to work effectively in real-world applications, where the delicate balance between classification performance and cost is crucial.

- **Recognition of Influential Users Supporting Online News Flow:** We contribute by identifying users within social discussions who play a critical role in supporting the flow of online news articles. These users, called Discussion Supporters, are particularly active in sharing common narratives and publishers. This identification not only improves our understanding of user behavior, but also helps to formalize the publisher classification problem in the context of ongoing social media discussions.

- **Identification of Worthy-to-be-ranked Domains:** Our work includes the identification of a list of *worth-to-be-ranked domains*. This curated list serves as a valuable guide for directing traditional annotation processes towards unknown publishers that may be relevant

for classification, and thus may be useful for organizations such as NewsGuard, MediaBias Fact Check, the Global Disinformation Index, etc.

As we clarify in the course of this chapter, our approach is built on three main pillars:

- Engagement Analysis Using an Entropy-Based Null Model: This model assists in identifying URLs that significantly engage active users within discussions. As discussed in this chapter, leveraging this data allows us to (i) streamline the complexity of constructing a process that maintains high accuracy even in resource-limited contexts, thereby achieving an optimal balance between accuracy and cost, and (ii) efficiently identify publishers that warrant priority evaluation, which, in our case study, often includes low-quality sources known for producing narratives that substantially engage the online public.
- Identification and Utilization of *voters*, defined as specific social media users who, following a characterization phase, aid in assessing the trustworthiness of previously unclassified publishers;
- Knowledge Base Tagging Using External Sources: Rather than evaluating each piece of online content individually, we adopt a source-based approach (see Section 2.2). Specifically, we utilize the high-quality annotation at the domain level provided by NewsGuard. Additionally, we leverage NewsGuard's domain quality assessments to validate our proposed methodology's efficacy.

The use of engaging narrative in our framework. Independent of the intent and editorial quality behind their creation, we have opted to incorporate the most engaging narratives—and by extension, their publishers—into our evaluation process. This strategy facilitates not only the selection of publishers to classify but also the selection and characterization of voters. Our rationale for adopting this approach is grounded in the following reasons:

- **Prioritization of Publishers**: We have chosen to primarily evaluate publishers capable of generating the most engagement, as we believe they are more effective in creating compelling narratives and, consequently, potentially more dangerous. Furthermore, Our observations have shown that narratives produced by low-quality publishers are captured by our benchmark. Specifically, these are narratives capable of attracting the attention of the active audience in the considered discussion.

- **Selection and Characterization of Voters**: In our research, we investigated the use of social media users who actively shared the most compelling narratives. We propose characterizing these users by focusing exclusively on the engaging content they distribute. This approach simplifies the complexity of our analysis, allowing us to (i) focus our attention on a specific subset of all active social media users and (ii) assess the propensity of these *voters* to distribute untrustworthy content, again using only a subset of the total URLs present in their timelines. Our findings demonstrate that selecting these users as 'voters' and characterizing them based on a limited set of engaging content leads to an efficient publisher-level trust evaluation process. This strategy strikes a favorable balance between cost efficiency and the accuracy of the results

  Additionally, examining the publishers with whom these *voters* engage allows us to infer the trust level based on the tendency of voters to interact with untrustworthy sources. Our experiments support the hypothesis that users frequently engaged with sources of a certain trust level are likely to interact similarly with other sources of comparable trust. For example, users who disseminate narratives from low-quality sources are often found sharing other, possibly less engaging but equally low-quality narratives. This observation is corroborated by [214], where the authors demonstrate that 'the adoption of information sources often mirrors users' prior exposure to sources with comparable credibility levels'.

- **Limitations and Context**: We acknowledge that in a context where

there are no publishers capable of generating engagement, our approach is not applicable. In such cases, it might be more sensible to evaluate sources by prioritizing their visibility (for example, the number of hits on Alexa Rank). Our selection method is designed to integrate, rather than replace, a visibility-based selection process.

6.1.2 Methods

Dataset and News Publisher Classification

Our dataset consists of approximately 1.87 million tweets in Italian posted by about 136,000 users, of which nearly 220,000 contain URLs. The data were collected using Twitter's streaming API between September 1 and September 24, 2021 (see Section 1.5.1 for details on the collection procedure). To characterize the news sources shared in the dataset, we rely on the reliability assessments provided by NewsGuard (see Section 2.3.3). After extracting the domains from the URLs contained in the tweets, we identify a total of 5,749 distinct news domains. Among these, 381 are labeled as T, 116 as N, while the remaining 5,252 are classified as UNC, indicating domains that are not evaluated by NewsGuard.

An overview of the proposed approach

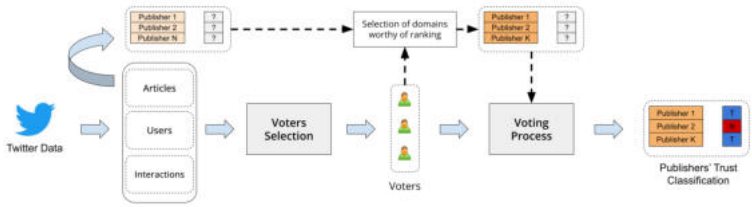


Figure 29: Schematization of the procedure for classifying the online publisher trustworthiness

Figure 29 illustrates the proposed methodology. As previously men-

tioned, the ultimate objective of this method is to estimate publishers' trustworthiness, categorizing them as either trustworthy (T) or untrustworthy (N). The process is structured into three main steps.

In the first step, we analyze posts published on Twitter/X regarding the Covid-19 vaccination campaign. We extract: (i) URLs in the posts that link to online articles, (ii) the retweet/tweet interactions through which these articles are disseminated, and (iii) the involved social users in circulating this content. Concurrently, we gather the domains of these articles, listing the media outlets involved in the online discourse.

The second step, termed 'Voters Selection,' aims to identify a set of key social accounts—*voters*—from the articles, users, and retweet/tweet interactions. These voters are pivotal in ultimately estimating the trustworthiness of publishers. Using a bipartite network of users and URLs, similar to configurations in [92, 155], we compare real-world interactions against an entropy-based benchmark. If two URLs frequently co-occur in the posts of the same users, significantly more than the benchmark predicts, we infer a strong associative link between them based on shared information diets. Using a community detection algorithm, we observe and delineate the network of URLs, forming what we term as *News Engagement Communities* (NECs). We emphasize that NECs and their analytical framework were first introduced in [155]. As established in previous studies [92, 127, 155], this validated projection highlights communities of URLs from domains that are relatively homogeneous in terms of trustworthiness. Moreover, URLs that meet the statistical significance criteria and thus integrate into the NECs typically represent a smaller subset of the original dataset, simplifying further examination, such as tagging URL domains through entities like NewsGuard.

In the concluding phase of this process, we aggregate all users who have interacted with at least one URL from a NEC on the social network. These users, termed *Discussion Supporters*, have demonstrated a vested interest in narratives that significantly contribute to a collective discussion. We aim to utilize these *Discussion Supporters* as voters to evaluate the trustworthiness of all supported online publishers through a classification approach. The publishers endorsed by the *voters* will be included

in a list of domains deemed 'worthy of ranking' and assessed accordingly.

In the final phase, the 'Voting Process,' we leverage the selected voters to determine the trustworthiness of publishers: specifically, we propose measuring and utilizing the voters' propensity to share untrustworthy content to assign a numerical value to each publisher on the 'worthy of ranking' list, which will guide the assignment of a trust label i.e., trustworthy (T) or untrustworthy (N).

Problem Definition

Estimating the level of trust of the publisher is a general problem: our work aims to provide a possible implementation. In this section, we provide some definitions that help to better define and understand the problem and the proposal of this work.

Let be A a set of online news articles whose links are shared in a specific social media discussion, with $n = |A|$. Let be P a set of online news publishers, with $m = |P|$. Each publisher p is assigned a $trust_p$ score, an integer in the range $[0, 100]$, indicating its trustworthiness. From these scores, we define a set L of distinct levels of trustworthiness associated with publishers, where $q = |L|$ is the number of trustworthiness levels. The set P_l contains the publishers associated with the trustworthiness level l. Each article a is associated with a single trustworthiness level l, determined based on the $trust_p$ score inherited from the publisher p of the article.

In this work, we will leverage social interactions between online users and publishers to characterize the trustworthiness level of the latter. We will follow different strategies, by changing the set of users employed in this procedure and the way we choose to characterize them. We will call this set *voters* and denote it as V, disregarding the adopted strategy.

Publisher coverage: Given the sets V, the **publisher coverage** PC_V of V is defined as the quantity of publishers posted by users in V.

Publisher coverage wrt l: Given the trustworthiness level l and the sets V and $P_l \subseteq P$, the **publisher coverage wrt** l of V, called PC_V^l, is

defined as the quantity of publishers posted by users in V also contained in P_l.

Classification task: We will develop a classifier for online publishers, named C_{trust}, which associates the level of trustworthiness $l \in L$ for each publisher p posted by the voters in V. C_{trust} will leverage the characteristics of voters in V.
Since the classification will look at voters and their characteristics, we will need to reach two intermediate milestones before proceeding with the evaluation of publishers:

How to select the voters: Given P, we need to select the set V such that the articles published by the voters in V ensure the maximum publisher coverage on P. The result of this task is the *list of publishers worthy of ranking*.

How to characterize the voters: Given V, we need to characterize each voter $v \in V$ to extract relevant information for the classification task.

Detection of News Engagement Communities

As introduced in [155], the concept of URL News Engagement Communities (URL NECs) plays a key role in the proposed methodology. The main idea is to detect groups of URLs that tend to be shared by the same users more frequently than expected by chance. Rather than relying on the content of the URLs, this approach focuses on patterns of user engagement.

To do so, we model the relationship between users and URLs as a bipartite network, where one layer represents users and the other URLs. A link is established when a user shares a specific URL via tweet or retweet. We then count how often each pair of URLs is co-shared by the same users and assess whether such co-occurrences are statistically significant, comparing them to a null model. Specifically, we employ the Bipartite Configuration Model (BiCM), which serves as an unbiased benchmark by preserving the activity levels of both users and URLs (see Section 2.3.1).

If two URLs are co-shared significantly more often than expected, they are linked in a projected, validated monopartite network of URLs. This results in a similarity network in which connections reflect statistically significant patterns of user engagement. On this network, we apply the Louvain community detection algorithm [19] to identify clusters of closely connected URLs. These clusters represent groups of URLs that were promoted or engaged with by overlapping communities of users.

This procedure parallels the approach used to detect user-based discursive communities (see Section 4.2.2). However, in this case, the focus is shifted to the content being shared rather than the users themselves. The resulting URL clusters—validated through statistical testing and filtered from noise—are referred to as *News Engagement Communities* of URLs (URL NECs). They provide insight into how audiences co-engage with specific sources or narratives, offering a complementary perspective to user-centric analyses.

Purity measure

Aimed to measure the homogeneity of URLs within a URL News Engagement Community (NEC) in terms of publisher trustworthiness, we introduce a metric termed 'purity'. This metric quantifies the proportion of URLs that originate from reputable versus non-reputable publishers within a URL NEC, analogous to the approach described in Section 4.4.1.

Let be URL NEC$_i$ the i-community and l_k, k=1,2, the trustworthiness levels T and N. We define $purity_{l_k}$(URL NEC$_i$) as the frequency of URLs belonging to l_k, i.e.:

$$purity_{l_k}(\text{URL NEC}_i) = \frac{|U_i^{l_k}|}{|U_i|}, \tag{6.1}$$

where $U_i = \{URL_1, \ldots, URL_n\}$ is the set of all the URLs in the i-community and $U_i^{l_k} \subseteq U_i$ is the subset of U_i that contains only URLs with trustwothiness level l_k. The purity defined in Eq. 6.1 can be interpreted as the probability of extracting an l_k-trustworthy URL from the i-th URL NEC. If m is the number of different URL NECs, we can define $purity_{l_k}(\cup_i$URL NEC$_i)$

as the frequency of URLs from l_k domains in all URL NECs:

$$purity_{l_k}(\cup_i \text{URL NEC}_i) = \frac{\sum_{i=1}^{m} |U_i^{l_k}|}{\sum_{i=1}^{m} |U_i|} \tag{6.2}$$

To have a benchmark for the purity of URL NECs, we also consider a purity measure for URLs that do not belong to any community:

$$purity_{l_k}(\overline{\cup_i \text{URL NEC}_i}) = \frac{|U_{-1}^{l_k}|}{|U_{-1}|}, \tag{6.3}$$

where the set of URLs that do not belong to any community is denoted as U_{-1}.

Voters selection

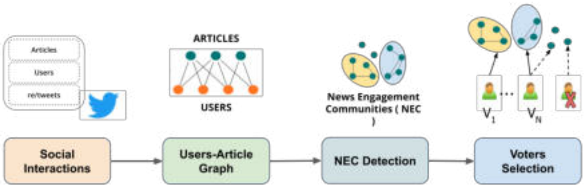


Figure 30: The Voters Selection Task

As anticipated in section 6.1.2, the primary challenge we face is the voter selection problem. This task aims to identify an optimal set of voters V that can effectively contribute to evaluating relevant publishers in the social media discussion under consideration, with a particular focus on identifying potentially low-quality publishers. Figure 30 shows the procedure we propose to implement: starting from the social data (gathered from Twitter/X in this case), we leverage a set of network science techniques to retrieve special communities of articles that prove to engage the public significantly (i.e., the *News Engagement Community* of URLs). Using that knowledge, we filter out all social users who don't interact with these communities and select the rest as voters.

In this study, we prioritize evaluating what we define as *relevant* publishers—entities adept at crafting compelling narratives that engage and captivate like-minded groups. In alignment with this goal, we designate as *voters* the social media users who engage with these publishers. More specifically, we select voters who support articles that notably engage diverse groups within the broader discourse. To facilitate this process, we utilize the concept of URL News Engagement Communities (NECs), introduced in [155]. These communities provide insights into the sets of news articles that generate significant engagement in the current online discussions. We specifically retain those news items that are shared more frequently than would be expected by their occurrence alone (for a detailed description, see Section 6.1.2).

Table 23: Statistics for URL NECs

ID	No. users	Distinct URLs	No. publishers	No. URLs
4	7422	223	71	38731
11	1064	21	4	1681
6	876	87	5	2019
1	674	79	9	2234
9	584	58	1	1238
7	521	64	6	1613
5	311	23	6	562
13	253	3	3	304
12	175	28	4	408
0	161	55	1	365
8	149	25	1	308
3	101	79	1	393
10	65	27	9	1557
14	42	4	2	59
2	23	3	3	32

In the considered context of the Italian COVID-19 vaccination debate, 51,504 URLs fall into URL NECs, about 22% of the total. More details can be found in Table 23, which shows some information about the different url NECs: ID is the individual community identifier, No. users is the

number of users who have shared URLs in the community, No. publishers is the number of online media outlets that have published the URLs. Url NEC 4 is the largest in terms of both size (consisting of 223 nodes) and impact on the overall dataset, as measured by the number of shares ($\sim$ 39k shares). The remaining url NECs can be distinguished based on the order of magnitude of the shares: we have 6 communities whose URLs were shared thousands of times, and other communities whose URLs were shared hundreds or dozens of times. Remarkably, in all but 4 of the url NECs, the number of different sources is quite limited (where source means the online news outlet that published the news to which the URL points).

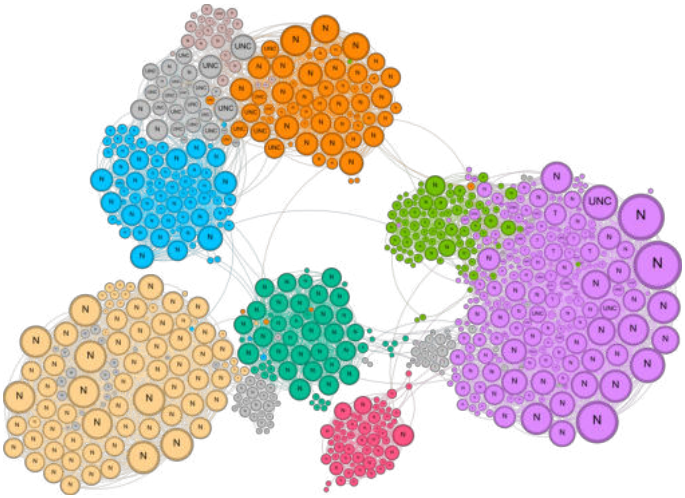


Figure 31: URL NECs: each node represents an single URL. The reliability tags, inherited from the source, is also reported.

Exposure to low-credibility sources. To characterize and observe the properties of emergent URL NECs (see Figure 31), we analyze the homogeneity of a URL's community in terms of publisher trustworthiness. To perform this task, we use the 'purity level' since it expresses how often trustworthy (and untrustworthy) publishers are associated with URLs in

a URL NEC.

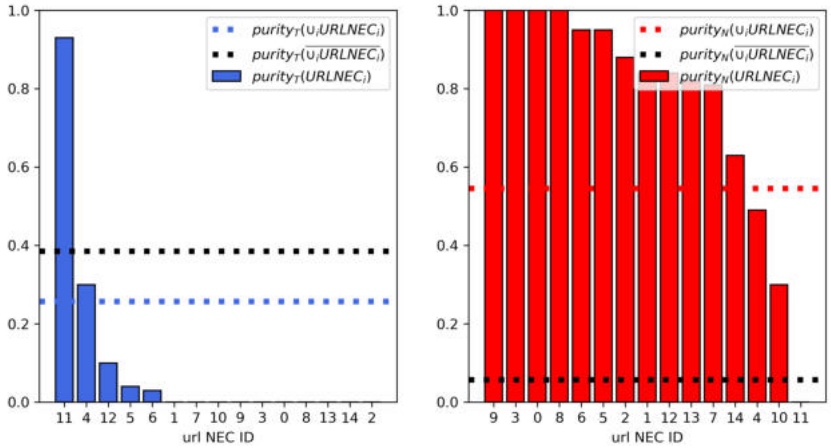


Figure 32: Purity levels of URL NECs. Trusted URLs on the left, untrusted URLs on the right. The size of a node is related to the degree

Figure 32 shows the purity level of each community in terms of links trustworthy (T) and untrustworthy (N). As noted in previous work [92, 127, 155], the URL validated projection has the property of producing communities of URLs belonging to publishers that are almost homogeneous from the point of view of (un)trustworthiness.

In summary, URL NECs tend to capture news articles that (i) are shared by a common audience (similar like-minded people, as evidenced by their statistically significant support for the same news) and (ii) have a homogeneous level of trustworthiness defined at the publisher level. It is important to emphasize that similar results, specifically homogeneous NECs in terms of publisher trustworthiness levels, can also be obtained using other community detection algorithms, such as InfoMap [169]. Since each URL considered in the present analysis is linked to a news article in a one-to-one correspondence, and since we are ultimately interested in using the news articles to infer the reliability of the users who share them, we will call A_{val} the validated monopartite network of URLs/news articles obtained above. Note that $A_{val} \subseteq A$, defined in 6.1.2:

the validated articles of the URL NECs represent a subset of all articles in our dataset.

We will use the validated news articles in A_{val} to characterize voters. Specifically, we can think of selecting as voters those users who have engaged with validated articles at least once. We call these users *Discussion Supporters* (DS): users who have played a central role in supporting the spread of the set of URLs that characterize the current online discussion. In the next section, we will better investigate how to characterize voters, whether *Discussion Supporters* or other sets of users.

Endogenous characterization of voters

Explicitly modeling the endogenous preferences of users based solely on their social network information is a non-trivial task. Building on similar methods used in previous work [65], we propose to use the messages posted by each user to characterize them. Specifically, our approach consists of approximating each voter's tendency to be a spreader (or non-spreader) of low-quality content. We measure this propensity using the average trustworthiness score of news articles shared by the user. To achieve this, we take a publisher-based approach and consider the tag assigned to the news publisher by expert annotators. Despite the approximation inherent in our measure, we argue that the use of expert annotations can ensure high quality ratings.

In the characterization procedure, we do not consider quoted tweets. We exclude them from the analysis because we cannot know for sure whether they are supportive or critical of the original tweet, since we do not address the content added by the user.

We propose four different strategies for characterizing voters, depending on whether and how we adopt the URL NECs.

- DS-URL-NEC: According to this first strategy, the voters are the Discussion Supporters, $V = DS$. We characterize the Discussion Supporters by considering only those articles that *they have shared on the social network in a specific discussion and that are part of at least one URL NEC*. We denote these articles as $A_{v_i}|_{\text{DS-URL-NEC}}$. Follow-

ing this first strategy, the URL NECs are used both to detect the set of voters $V = DS$ and to identify the subset of news articles for their characterization.

- DS-ALL: According to this second strategy, the voters are the Discussion Supporters, $V = DS$. However, for their characterization, we consider *all the news articles that the voters have shared in the social network in a given discussion*. We denote these articles as $A_{v_i}|_{\text{DS-ALL}}$. Thus, in this case, the URL NECs are only used to detect the set of voters V.

- DS-ALL-WO-USR-NEC: According to this third strategy, the voters are the total set of users U minus the Discussion Supporters $V = U - DS$. Each voter v_i is characterized by considering all the news articles $A_{v_i}|_{DS-ALL-WO-USR-NEC}$ shared by the voters on the social network in a given discussion.

- USERS-ALL: This strategy does not consider the notion of URL NECS. Here, the voters are the entire set of users and the news articles are those shared in the discussion: $A_{v_i}|_{USERS-ALL}$.

For all strategies, we characterize the articles shared by voters by the average of the trustworthiness scores NewsGuard assigned to the article publishers. As an example, consider the articles that each voter has shared in the social network using one of the strategies described above. We assume that the $i-th$ voter has shared 10 news articles, 5 from publisher x and 5 from publisher y. We also assume that according to our external source NewsGuard, publisher x has a trust score of =60 and publisher y has a trust score of =90. Then the $i-th$ voter is assigned a characterization value equal to the arithmetic mean $(60 * 5 + 90 * 5)/10 = 75$.

We also consider an additional characterization of voters by considering the variability of their 'information diet,' i.e., how many different publishers a voter has shared in the social network. The goal of this additional characterization is to evaluate whether variations in voters' information diets can influence the reach and performance of our proposal.

Figure 33 shows the number of voters *wrt* the chosen strategy for their selection and the minimum number of shared publishers.

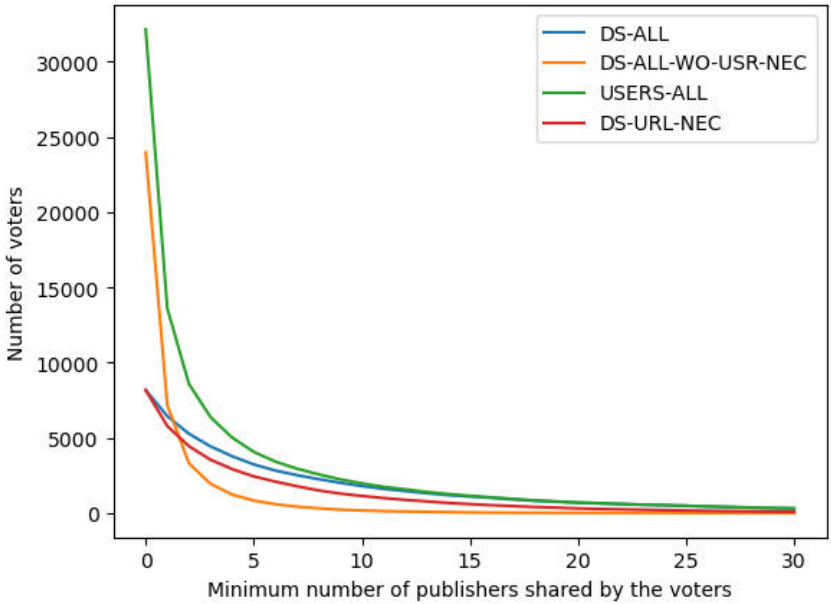


Figure 33: Number of voters *wrt* the minimum number of shared publishers and the adopted strategy

Classification of online publishers via voters

Here, we propose an implementation for the classification task defined in Section 6.1.2. We consider two levels of trustworthiness, such that $L = \{T, N\}, (|L| = 2)$. The label T corresponds to trustworthy publishers, while N denotes untrustworthy publishers. These labels correspond to those used by NewsGuard (see Section 2.3.3). The classification task is binary.

To determine whether a publisher posted by the voters in V belongs to the T or N class, we take the characterizing values of all voters who have interacted with articles from that publisher, and again perform an

arithmetic mean. For example, if we have 10 voters who have interacted with publisher p, 5 with characterizing value =75 and 5 with characterizing value =60, it simply follows that publisher p is associated with a value of $(75 * 5 + 60 * 5)/10 = 67.5$. We can then proceed with the implementation of a simple decision tree model with a maximum depth of one, due to its simplicity and interoperability. What we will show as the result in the following sections comes from a classifier to which we give as input, for each publisher, its score (e.g., 67.5) as the only feature and the label available from Newsguard. We use the scikit-learn Python library, which provides implementations for (i) the decision tree and (ii) a 10-fold stratified cross-validation (preserving the percentage of samples for each class in the folds).

The classifier is intentionally designed to rely on a single feature, namely the average score of the voters interacting with a given publisher. While additional features could be incorporated—such as the distribution of voter scores, network-based properties, or user-level metadata—and more sophisticated models could be employed, the goal of this study is to assess whether social interaction patterns alone can provide a meaningful early signal of publisher trustworthiness.

For this reason, the classification step is deliberately kept simple and interpretable, using a decision tree with maximum depth equal to one. This choice allows us to isolate the predictive contribution of voter behavior while avoiding additional modeling complexity. If reliable classification performance can be achieved under these minimal conditions, richer feature sets and more complex models could further improve the results in future work.

In the following, we will present the classification results for those publishers annotated by NewsGuard. Obviously, *unclassified* publishers cannot be used to evaluate our proposal, as we lack ground truth information for them.

6.1.3 Results

In this section, we validate the proposed method using Twitter/X data related to the Italian vaccination debate (see Section 1.5.1), and present the main findings of our analysis. Initially, we demonstrate the coverage of publishers with respect to the total dataset achieved by the proposed procedure. Subsequently, by utilizing the knowledge provided by NewsGuard on the set of annotated publishers within our dataset, we assess the performance of our method.

Publishers' coverage

Table 24: Percentage of publisher coverage

Strategy	T	N	UNC
DS-ALL / DS-URL-NEC	70.87	90.52	29.74

In Table 24, we show the publisher coverage PC_V^l when the voters in V are the *Discussion Supporters* and the minimum number of publishers shared by the voters is 0. Given this configuration, we note that both DS-ALL and DS-URL-NEC contribute equally to the coverage calculation, since both consider the same set of voters. We consider the trust levels $L = \{\text{T, N, UNC}\}$. We capture $\sim 71\%$ of the trusted publishers. We are also able to reach 90.52% untrustworthy publishers. Thus, we can say that, especially for publishers of type N, the coverage is particularly large. The coverage also includes $\sim 30\%$ of unclassified publishers[1]. In terms of helping organizations like NewsGuard identify news media for analysis, these UNC publishers are promising because they are publishers shared by voters who were also interested in peculiar news (i.e., news whose links end in URL NECs).

In Figure 34, we show how many T and N publishers we can reach, considering all the strategies defined in Section 6.1.2 and the minimum

[1]Unclassified publishers are those publishers for which we do not have a trustworthiness label assigned by an external source (NewsGuard in this study).

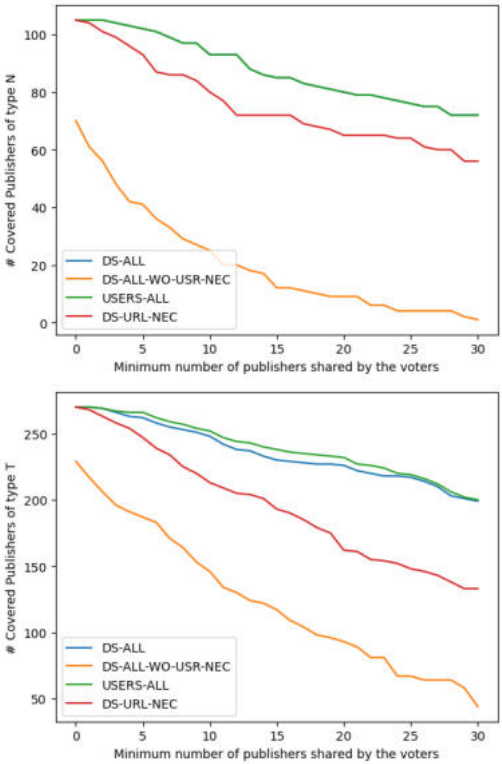


Figure 34: How many publishers can we reach given the original number of publishers, the set of voters, and the minimum number of publishers shared by the voters?

number of publishers a voter must engage with to be considered. On the y-axes, the top panel shows the number of N-type publishers reached, and the bottom panel shows the number of T-type publishers reached (in our case study, we have a total of 116 N-type publishers and 381 T-type publishers, see Section 6.1.2).

In the top panel, we can see that strategies such as DS-ALL, DS-URL-NEC, and USERS-ALL consistently provide the best coverage, even as the threshold for the minimum number of publishers increases. On the contrary, the DS-ALL-WO-USR-NEC strategy shows limitations,

providing the least coverage for untrusted publishers. This underscores the effectiveness of relying on URL NECs to construct an efficient list of publishers worth annotating in terms of coverage. In the bottom panel, we observe a similar pattern for the coverage of trusted publishers.

Classification performance

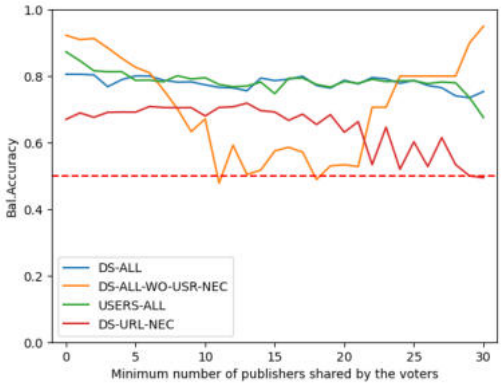


Figure 35: Publisher classification performance

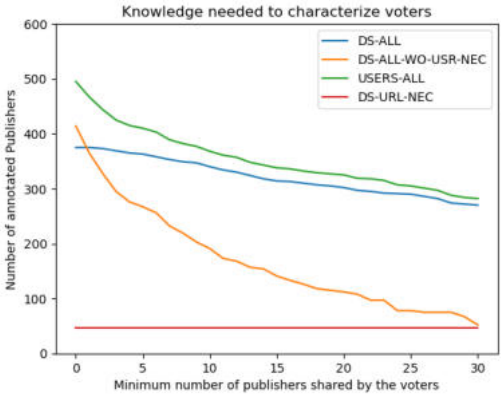


Figure 36: Initial knowledge in terms of publishers annotation

Figure 35 shows the classification results obtained on the covered publishers, using a simple decision tree (Section 6.1.2). The baseline, set at 0.5, serves as a reference point.

Quite intuitively, if the voters are all the users in the initial dataset, and their characterization is calculated from all the news they have shared (strategy USER-ALL), the classification results are the best, with a peak in Balanced Accuracy at 0.806.

Let us now analyze what happens when we consider strategies that consider $V = DS$, namely DS-ALL and DS-URL-NEC. Both strategies outperform the baseline. Noteworthy is the fact that the strategy where we characterize Discussion Supporters with all their shared URLs, DS-ALL produces results comparable to the USER-ALL strategy. As for the strategy that considers only the ground truth of the publishers in the URL NECs, i.e., DS-URL-NEC, it leads to worse results than the DS-ALL strategy, but still achieves peaks of Balanced Accuracy = 0.719, and at a lower cost in terms of initial labeling, see Section 6.1.2.

Finally, considering the classification of publishers only on the basis of users who did not share anything in the URL NECs, we underline that these users cover a very small part of the T and N URLs in the original dataset of URLs, as can be seen in Figure 34. Thus, while it is true that in some cases, depending on the number of shared publishers, the accuracy is comparable to the other strategies, it is also true that these users are not useful for our purposes, given the small coverage.

6.1.4 Discussion

In our quest for the optimal classification solution, we must make a trade-off between three crucial aspects: publisher coverage, classification accuracy, and the cost of acquiring the knowledge required to characterize voters who contribute to publisher classification.

Knowledge vs. Cost. Figure 36 shows the number of annotated publishers required to characterize voters and thus classify additional publishers. Regardless of the minimum number of publishers shared by the voters, if we follow the strategy DS-URL-NEC, the knowledge is con-

stant, equal to the number of publishers shared by DS whose URLs are also part of a URL-NEC. In the specific case of our case study, we need the trust labels of 46 publishers. If we follow the USER-ALL and DS-ALL strategies, we need the annotation of at least 300 publishers (in case the voters have a rather diverse information diet, at least 25 shared publishers per voter). Since the manual process required to annotate even a single publisher is time-consuming and tedious, if we can afford a classifier with lower accuracy (see Figure 35), DS-URL-NEC is obviously preferable.

Accuracy vs. Cost. Still referring to Figure 35, the strategies USER-ALL and DS-ALL give the best results in terms of balanced accuracy. However, as shown in Table 36, both strategies require a considerable amount of initial knowledge compared to DS-URL-NEC. Although DS-URL-NEC achieves a Balanced Accuracy below USER-ALL and DS-ALL, it outperforms the baseline (see Table 35) and can provide acceptable performances; we remember that for the (cheaper) DS-URL-NEC strategy we reach a peak of 0.719, while with the most expensive one (i.e., DS-ALL) we achieve 0.806.

Coverage vs. Knowledge. In real-world scenarios, prioritizing a clear list of relevant publishers for classification is challenging, leading to potential gaps in coverage and the risk of losing the evaluation of low-credible publishers. To address this challenge and ensure coverage of low-credibility publishers, we propose to generate a list of publishers worth annotating using an entropy-based procedure (see Section 6.1.2 for details). In Section 6.1.3, we observed that two strategies proposed in this work, namely DS-ALL and DS-URL-NEC, provided the best publisher coverage with different trade-offs in terms of knowledge.

Twitter's Change of Ownership and the Rise of X. In late October 2022, the American social media company Twitter, Inc. underwent a significant transformation as it came under the ownership of Elon Musk[2]. This transition ushered in a series of radical changes and reforms that included both managerial and technical aspects.

[2]https://www.nytimes.com/2022/10/27/technology/elon-musk-twitter-deal-complete.html

A key development for the scientific community was the elimination of Twitter's free API tier by February 2023, to be replaced by a 'basic paid tier'[3]. For researchers and developers, this shift meant that Twitter content would no longer be available for research purposes without subscribing to a significantly different paid plan. It also posed a challenge in terms of re-hydrating the datasets currently in use.

Although our dataset was collected during a period of free access (from September 1 to September 21, 2021), the policy appears to be unchanged at the time of writing this manuscript[4]. Thus, we acknowledge the potential obstacles to the reproducibility of the experiments presented here.

However, we maintain that our methodology remains highly adaptable to other online social networks. Indeed, it relies on a core principle: the analysis of account activity related to the sharing of news publishers' URLs. We argue that extending this approach to alternative social platforms is feasible.

6.1.5 Conclusions

The study presented in this Section introduce a novel method for automatically classifying the trustworthiness of publishers, starting from limited initial knowledge. Our method exploits social media interactions between two key actors: publishers (news producers) and social users (consumers). The primary goals of our approach are twofold: 1) to identify a list of publishers worthy of annotation, and 2) to assemble a pool of voters capable of making informed judgments based on their historical behavior, thereby quantifying their tendency to spread trustworthy or untrustworthy news.

Our approach shows commendable performance even in scenarios with limited prior knowledge. It effectively navigates the tradeoff between effectiveness and cost, demonstrating its adaptability in contexts where extensive knowledge is limited. Furthermore, our observations

[3] https://twitter.com/XDevelopers/status/1621026986784337922

[4] https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api, consulted February 27, 2024.

highlight the method's ability to achieve robust coverage of publishers, including those classified as untrustworthy. These results underscore the potential of our approach in real-world applications where the trade-off between performance and cost is critical.

Limitations: Like other methodologies, our approach has some clear limitations. First, it depends on the context analyzed: in this case, the focus was on the acceptance of the COVID-19 vaccination in Italy and the news sources analyzed were the most active in covering this topic. Changing the topic would have resulted in a different data set, in which the interactions between the different accounts would have been different, and therefore the procedure would have yielded a different list of news sources to annotate. Such a point can be turned into a potentiality of our methodology: the present annotation strategy is topic-specific. In other words, our process identifies the most relevant news sources to annotate for the specific discussion.

Including different online debates in the dataset would provide a finer description, where the relevance of different news sources is characterized at the topic level. However, it is important to emphasize that when different topics are present in the same dataset, multiple situations may occur. If the topics are somewhat disconnected from each other, such as the US presidential election and *buccellato di Lucca*[5], we expect that our methodology will be less effective since the Italian receipt could break the "bubble" of US election propaganda and make the signal weaker. However, if the debates captured by the dataset correspond to arguments that are part of the same narrative - for example, migration from North Africa to Europe and COVID-19 vaccination, about which the Italian far-right conspiracy theorists are particularly active - then we expect the characterization of the cluster of domains to be finer. This topic will be the subject of further research[6].

Lastly, our proposal is currently limited to single social platform. Testing our procedure on different social platforms will be the subject of future research.

[5] `https://en.wikipedia.org/wiki/Buccellato_di_Lucca`

[6] We are grateful to the reviewer who gave this suggestion.

Finally, in the current version of our procedure, the list of features that characterize discussion supporters is extremely limited. Future works can enrich the description of DSs.

6.1.6 TROPIC, a tool for trustworthiness rating of online publishers

As discussed in Section 2.2, existing methods for assessing the trustworthiness of online news publishers are often resource-intensive and difficult to scale. Specialized organizations such as [1, 106, 130, 136, 192] provide valuable reliability ratings, but these rely on costly and time-consuming manual evaluations, for instance assessing bias or propagandistic content [8, 113]. Consequently, a large number of online publishers remain unlabeled, creating a significant coverage gap.

To address this limitation, the tool presented in this Subsection supports and complements the efforts of such organizations by providing a scalable and data-driven solution. Building on the framework introduced in [157], the system leverages user interactions in online discussions to (i) estimate trustworthiness scores for previously unclassified publishers and (ii) offer an interactive platform to guide annotation efforts and improve the robustness of the resulting evaluations.

Proposal: Based on [157] (Section 6.1), we present TROPIC– Trustworthiness Rating of Online Publishers through online Interactions Calculation (available at `https://tropic.iit.cnr.it`), an interactive tool to improve current methods for assessing an online publisher trustworthiness and to avoid potential problems of limited coverage. TROPIC outputs include:

- **The automatic classification of the level of trustworthiness of an online news publisher:** TROPIC provides an assessment of the trustworthiness of unclassified online news publishers by analyzing social media interactions between news producers and consumers. The evaluation starts with the analysis of an online social discussion and leverages (i) a *base-knowledge* consisting of a subset

of annotated news publishers, (ii) the concept of *News Engagement Communities*, NECs for brevity, which are communities of news articles that received the most attention from discussion participants, introduced in [155, 157] (presented in Chapter 4 and Section 6.1) and (iii) the propensity of social users to share low-quality information. This approach provides (i) good accuracy-cost trade-off and (ii) good coverage of *unclassified* and *untrustworthy* news publishers [157].

- **An interactive guide for base-knowledge extension:** The robustness of the classification can be continuously improved by gradually extending the current *base-knowledge*. The intuition is that an extended knowledge allows for more accurate social user profiling, and, consequently, more accurate predictions about the trustworthiness of not yet annotated news publishers. To guide the extension of the *base-knowledge*, TROPIC provides the end user with a set of functionalities that suggest the "best" next publisher to annotate. This should guide where to focus the base knowledge extension by taking the best advantage of the prediction system.

Context: Manual methods of assessing the trustworthiness of online newspapers are costly, both in terms of the time required to perform the assessment and the cost of finding experienced reviewers. These barriers limit the reach of assessable newspapers. The low coverage also stems from the constant emergence of new, lesser-known online newspapers. TROPIC aims to streamline the rating process and expand the reach of rated newspapers.

Audience: Designed for journalists, media professionals, and researchers evaluating the quality of online information, TROPIC provides estimates of the trustworthiness of online publishers. The tool guides the user in the selection of online news publishers to manually annotate, so that the subsequent automatic evaluation of news publishers not yet evaluated is effective in terms of prediction accuracy and coverage of the largest

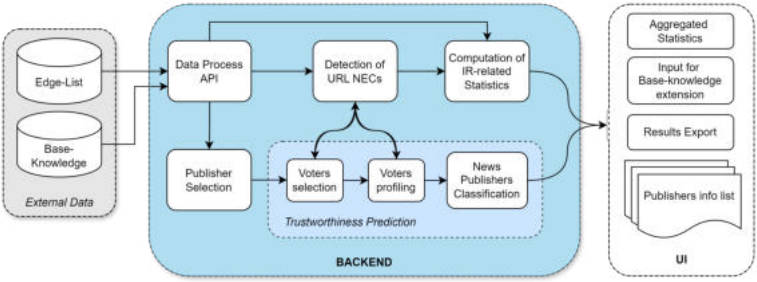


Figure 37: System Overview

number of news publishers evaluated.

System overview

External Data: Figure 37 shows the tool's architecture. TROPIC begins by processing an online discussion in *edge-list* format, which links the URLs - pointing to an online news - in the posts to the user IDs of the social users who shared those URLs. The tool is adaptable to data from any social media platform[7]. A second (optional) input for TROPIC is the initial base-knowledge, which is a list of online publishers with associated trustworthiness scores (integers from 0 to 100). Since TROPIC is designed to support organizations specializing in evaluating news publishers' trustworthiness, the initial knowledge base can consist of publishers already assessed by these organizations[8].

Backend: The TROPIC backend consists of independent software components that work together to provide (i) a predicted trustworthiness score (with a corresponding confidence level) for each publisher not in the base-knowledge, and (ii) metrics to guide users on which publishers to annotate in case they want to manually expand the base-knowledge. The edge list is used both to build NECs communities (which we recall

[7]URL preparation, such as resolving short URLs is left to the TROPIC user.

[8]As an example, NewsGuard [136] uses the 0-100 scale, allowing for easy integration. Conversely, other entities such as MediaBias Fact Check[130] use descriptive labels and should convert them to numerical values for system compatibility.

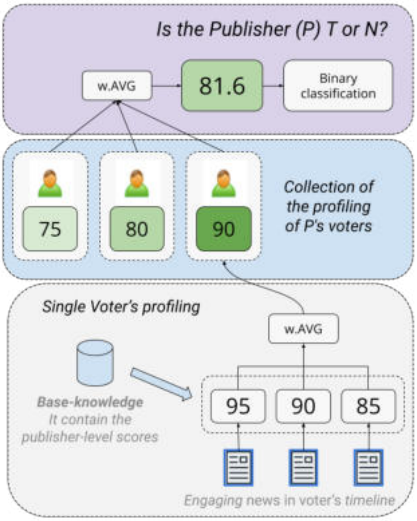


Figure 38: Publisher Trustworthiness Classification

are communities of news URLs that have been very successful among online users in terms of engagement [155, 157]) and to calculate some quantity related to the online discussion we are studying (IR-related statistics). Once we have selected the publishers whose URLs have been shared in the online discussion, we move on to selecting a special group of users, called voters, and profiling them. After this profiling phase, these voters are asked to classify the publishers they interact with. In our current implementation:

1. We select as voters those users who have shared at least one engaging news (i.e., a news whose URL belongs to one of the NECs).

2. For profiling voters - estimating each voter's propensity to share low-quality content -we follow the approach in [157], which consists of a two-step method to downscale the problem. First, instead of considering the complete set of URLs in a voter's timeline, we only select engaging URLs, specifically those belonging to NECs. Second, we evaluate their trustworthiness level by taking

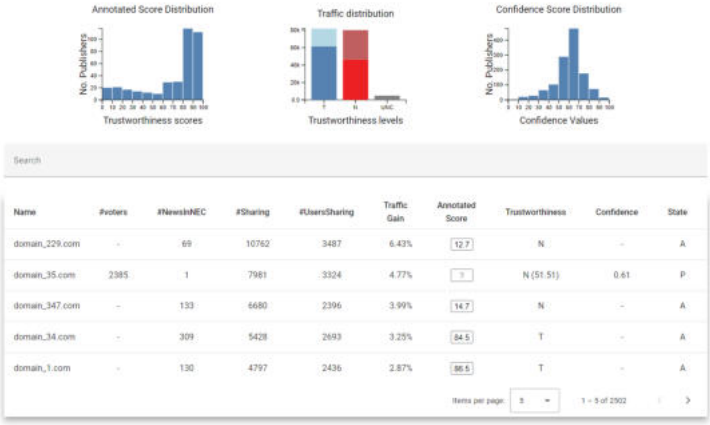


Figure 39: An implementation of the User Interface (UI)

their publisher-level score (if available from the *base-knowledge*) instead of evaluating every single online news.

Finally, to compute the trustworthiness score for an unclassified publisher, we select all voters who share content from that publisher, profile them, and aggregate their scores into a single (average) number that expresses the trustworthiness score of the considered publisher (Figure 38).

User Interface: Figure 39 shows the User Interface (UI). The toolbar allows to upload the edge list (via "File Upload" button) and, optionally, the *base-knowledge*, and to initiate the computation via the "Process data" button. Calculation results are displayed in three plots and a table. The table lists each publisher's name, IR-related statistics, an editable Score column (this allows the user to manually annotate other domains, if desired, to extend the base knowledge), a Prediction column (showing both binary labels - trustworthy T/ untrustworthy N) and prediction scores, a Confidence column, and a State column (indicating if the rating comes from human annotation or from TROPIC). Users can sort the table by

IR-related metrics, such as the number of voters or the number of news URLs shared within NECs. This way, manual annotations can be prioritized based on a metric of interest. Above the table, plots provide information on the number of manually annotated publishers and the distribution of their scores (left plot); the number of publishers annotated by human experts (dark color at bottom) and those whose score are calculated by TROPIC (light color at top) (middle plot); and the confidence level of the scores calculated by TROPIC (right plot). This information guides the user to minimize the number of publishers that remain unclassified and to improve the confidence level of the predictions (by adding new manual annotations if necessary). The "Export Actual Knowledge" button allows to export the calculation results in CSV format.

Implementation and demonstration: The user interface is built with Angular [4], and the backend uses FastAPI [71]. NEC extraction employs the bicm Python library [18]. Both UI and backend are containerized separately with Docker [64]. For practical testing, we provide the tool in a DEMO version. In this configuration, the waiting times for calculating the NECs—which could be several minutes for a very large edge-list—are pre-calculated. To utilize this setup, the end user is supplied with a DEMO edge-list, which must be selected before pressing "Process Data". For completeness, a DEMO base-knowledge is also provided, featuring randomly generated trust scores. If the end user wishes to upload her own edge-list, the number of edges that can be uploaded in the demo version is limited to 50,000 entries.

Conclusions

In this Subsection, we presented an interactive web interface designed to streamline the annotation process for assessing trustworthiness level of an online publisher. TROPIC predicts trustworthiness scores for publishers that have not yet been manually labeled, improving the coverage and reliability characterization of news domains traffic exchanged in an online discussion.

6.2 Evaluating trustworthiness of online news publishers via article classification

The proliferation of low-quality online information in today's era has underscored the need for robust and automatic mechanisms to evaluate the trustworthiness of online news publishers. In this work, we analyse the trustworthiness of online news media outlets by leveraging a dataset of 4033 news stories from 40 different sources. We aim to infer the trustworthiness level of the source based on the classification of individual articles' content. The trust labels are obtained from NewsGuard, a journalistic organization that evaluates news sources using well-established editorial and publishing criteria. The results indicate that the classification model is highly effective in classifying the trustworthiness levels of the news articles. This research has practical applications in alerting readers to potentially untrustworthy news sources, assisting journalistic organizations in evaluating new or unfamiliar media outlets and supporting the selection of articles for their trustworthiness assessment.

6.2.1 Problem formulation and contributions

Disintermediation, or the phenomenon of reducing intermediate flows, is a term coined back in 1983, when author Paul Hawken called by this name the set of processes by which consumers could directly manage financial investments in securities, rather than leaving their money in savings accounts [98].

Over time, many industries have experienced disintermediation. In tourism, the Internet provides users with access to a wealth of information, allowing them to seamlessly assemble various tourism services and create unique travel experiences on their own. Similarly, the growing trend of self-publishing places increasing responsibility on authors to oversee the entire process of producing and distributing their work [77]). Journalism has also experienced changes that reflect the evolving landscape of direct access to information and news dissemination. These shifts indicate a broader societal trend toward greater autonomy and

control in various industries. Particularly with regard to journalism, the emergence of new web technologies and social networks has diminished the essential role of traditional journalists as the prevalence of participatory journalism facilitated by blogs and social networks continues to grow [22, 111]. In this regard, a recent UNESCO report[9] on the existential threat posed by social media to traditional news claims that online 'news outlets often struggle to get the clicks from readers that determine advertising revenue' and job cuts in journalism have resulted in a noticeable void in the information landscape.

The erosion of the mainstream journalism system in recent years, coupled with challenges such as understaffing and the pressure to publish attention-grabbing news to re-engage readers, has raised concerns about the quality of information provided by online media. Various journalism organizations and indices, including NewsGuard[10], the MediaBias Fact Check[11], the Iffy index of unreliable sources[12], the Global Disinformation Index[13], the Ad Fontes Media[14] that conduct studies on the transparency and trustworthiness of online news sources, including their tendency to produce propagandistic and/or politically biased content.

Although different organizations use different criteria to determine the trustworthiness of an online media outlet, recent work has found excellent convergence in the labels each assigns to individual media outlets, confirming the degree of trustworthiness of the judgments [120, 153].

Unfortunately, the process of evaluating each news outlet is very cumbersome, especially in terms of time. For example, the procedure followed by the Global Disinformation Index is to select annotators who are experts in the online information system of a particular country. After training, the annotators select a group of newspapers that accurately reflect the country's information landscape. They then manually analyze these newspapers to find information on aspects such as ownership

[9] `https://news.un.org/en/story/2022/03/1113702`. All of the URLs in this document were last accessed on December 22, 2023.

[10] NewsGuard: https://www.newsguardtech.com

[11] MediaBias Fact Check: https://mediabiasfactcheck.com/

[12] Iffy Index: https://iffy.news/index/

[13] The Global Disinformation Index: https://www.disinformationindex.org/

[14] Ad Fontes Media: https://adfontesmedia.com/

and funding sources. This is followed by a manual content analysis of a sample of articles per media outlet to check for unreliable, sensational and/or propagandistic content. The GDI then processes the results of the study, which are summarized in a score between 0 and 100 that indicates the risk that the media outlet is misinforming its readers[15].

In this work, we aim to label the trustworthiness level of a news source from the classification of the news itself. Operationally, we start with a dataset of 4033 news stories from 40 online news outlets, which we have collected and to which we have attached labels regarding both the main topic and the trustworthiness score of the news outlet. The labels are collected by NewsGuard, which is licensed to the authors of this work at the time of writing [17]. Through qualified journalists, NewsGuard rates all news sources, which account for 95 percent of online engagement[16]. Each site is analyzed according to nine accepted journalistic criteria. Based on these nine criteria, the site receives a trustworthiness score from 0 to 100. The trustworthiness levels are 5, ranging from 'high credibility,' the best rating, to 'proceed with extreme caution,' indicating a site with a very low level of transparency and credibility.

On the one hand, tagging the articles in our dataset with NewsGuard labels relieves us of the tedious task of annotating the data and gives us a solid ground truth based on the work of specialized journalists. On the other hand, before moving on to the main goal of the work, which is to derive the level of trustworthiness of the news source from the analysis of individual articles, we will test the goodness of the dataset by classifying articles by topic and making sure that the predicted topic matches the topic assigned in the label.

We combine 3 standard topics, i.e., *Sports*, *Political News* and *Health*, with an escalating one, in the age of the internet and pandemics, vaccines and wars, namely *Conspiracy Theories* [29].

Main Contributions.

Our models have proven to be highly effective in both classification

[15]The second and third authors are familiar with the procedure, having participated as annotators in the GDI country study on the Italian online media market [147].

[16]https://www.newsguardtech.com/solutions/newsguard/

tasks. The results are summarized as follows:

- **Trustworthiness Detection Task:** This is a multi-class classification task at the article level. Our model successfully predicts the level of trustworthiness of news sources based on the article text alone. Specifically, we obtain an average F1-macro of 0.843 and an average F1-micro of 0.882 for this task (see section 6.2.5).

- **Topic Detection Task:** This is a multi-class classification task at the article level. Our model achieves an average F1-macro of 0.925 and an average F1-micro of 0.929. This gives us an additional level of confidence in the original NewsGuard labeling of the data set. Wrong predictions arise when distinguishing between *'Conspiracy'* and (*'Health"* or *'Political'*), leading to some misclassifications (see Section 6.2.5).

6.2.2 Possible applications

The ability to predict a publisher's level of trustworthiness from the classification of individual articles suggests at least three possible applications, one to assist the user, the others to assist journalistic organizations (e.g., NewsGuard and GDI):

1. at the user level, the classifier can be used as a tool to alert the reader by displaying a meaningful visual signal, such as the classic red flag, while the reader is viewing an article from an unfamiliar news outlet. The red flag could say something like 'the article you are reading is similar to those produced by untrustworthy newspapers. Supplement your reading with other readings of articles produced by trustworthy newspapers'.

2. at the organizational level, let the reader assume that the media outlet is new or completely unknown to the evaluator (which is very common these days given the constant proliferation of alternative online media outlets [190]). An initial idea of its level of trustworthiness can be obtained by collecting a number of articles

and applying the trustworthiness ranking model to them. At a later date, if the evaluator deems it necessary, a more comprehensive investigation can be conducted using traditional journalistic analysis.

3. still at the organizational level, the selection of articles to analyze to assess the publisher's trustworthiness are typically the most shared articles on social media and/or articles containing a set of representative keywords.[17]. Unfortunately, this method may not be sufficient to select a sample of articles that is truly representative of the publisher. For example, if we relied on the most shared articles on social media, we might select a sample consisting only of straight news stories (e.g., traffic accidents, robberies, etc.). Therefore, we argue that the models presented in this work can be used to process a selection of articles from the target media for a more balanced sample that can be manually analyzed. This approach ensures a balanced assessment in terms of both trustworthiness levels and topics.

6.2.3 A more formal problem definition

In this section, we formalize the main problem addressed in this Section, called *Trustworthiness Level* Detection, and present the performance metrics used to evaluate the resulting models. The *Topic Detection* task is also formalized, since we use the results of this classifier to experimentally evaluate the quality of the labeling procedure obtained from NewsGuard.

Let be A a set of articles. This set represents a collection of articles characterized by their textual content. Formally, $A = \{a_1, a_2, \ldots, a_n\}$, where n is the total number of articles. Each article a has attributes $text_a$ representing the textual content of the article and $newspaper_a$, the newspaper from which the article originates.

[17] https://www.disinformationindex.org/country-studies/2023-06-08-disinformation-risk-assessment-the-online-news-market-in-thailand/

The set of Newspapers N comprises various newspapers, each associated with a specific level of trustworthiness. Formally, $N = \{n_1, n_2, \ldots, n_m\}$, where m is the total number of newspapers. Each newspaper n_i has $trust_{n_i}$ that represents the level of trustworthiness associated with it.

Notice that different levels of trustworthiness associated with newspapers belong to the set L. Formally, $L = \{l_1, l_2, \ldots, l_p\}$, where p is the total number of trustworthiness levels. The level of trustworthiness is typically a continuous value, but we prefer to aggregate values in bins that correspond to our levels of trustworthiness L. We provide more details about the bins in Section 6.2.4.

Notice that each article a_i is associated with one and only one level of trustworthiness ($trust_{a_i}$) from the set of trustworthiness levels (L) inherited by the newspaper it originated from.

Therefore, given the sets A, N, and L, and the constraints mentioned above, we formalize the problem as follows: we aim to develop a text classifier C_{trust} that associates the level of trustworthiness $trust_a \in L$ for each article a based on its textual content $text_a$. The Classifier C_{trust} is a machine learning model that addresses the *Trustworthiness Level* Detection task, and it relies on the NewsGuard annotation process described in Section 6.2.1. Given this, we can reasonably assume that the level of trustworthiness associated with each newspaper is reliable, a notion further supported by considering the primary topic associated with each newspaper.

To this end, we define the *Topic Detection* task as follows. Each newspaper n_i has an additional attribute associated with it: $topic_{n_i}$ that indicates the topic it primarily covers. This information is gathered by NewsGuard. The set T defines the possible categories or topics into which articles can be classified. Formally, $T = \{t_1, t_2, \ldots, t_k\}$, where k is the total number of topics. As for the previous classification task, each article a_i is associated with one and only one Topic ($topic_{a_i}$) from the set of Topics (T) inherited by the newspaper it originated from. Thus, given the sets A, N, and T, and the constraints mentioned above, we formalize the problem as follows: we aim to develop a text classifier C_{topic} that associates a $topic_a \in T$ for each article a based on its textual content $text_a$.

The Classifier C_{topic} is a machine learning model for *Topic Detection* task. As in [150], which addressed a similar challenge to ours, the performance of the classifiers is evaluated using the F1 Micro and F1 Macro metrics, defined as follows:

$$\text{F1 Micro} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.4}$$

$$\text{F1 Macro} = \frac{1}{|M|} \sum_{i=1}^{|M|} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{6.5}$$

Using these two metrics allows us to apply a consistent measurement approach, as established by [150], and thereby illuminate two different facets of classification performance. The F1 Micro (6.4) considers all predictions comprehensively, while the F1 Macro (6.5) treats each class independently by calculating a weighted average of the F1 scores for each class. In Equation 6.5, we denote a general set of classes M. In our tasks, the classes are defined by the set L representing levels of trustworthiness for the *Trustworthiness Level* Detection and the set T representing topics for the *Topic* Detection.

6.2.4 Dataset

This section describes the procedure used to construct the dataset of articles employed in our experiments. The process consists of three main steps: (i) selecting a representative list of online media outlets, corresponding to the set N of newspapers introduced in Section 6.2.3; (ii) retrieving and cleaning the textual content of the articles published by these sources, forming the set A of articles; and (iii) tagging the collected articles according to the reliability information provided by NewsGuard.

Online media outlets selection

The goal of the source selection process is to obtain a representative set of online media outlets that reflects real-world conditions and allows for a meaningful evaluation of the proposed models.

To this end, we start from the NewsGuard dataset of annotated online news sources, which is available to the authors under the NewsGuard license. This dataset contains news outlets evaluated by expert journalists according to a set of journalistic criteria (see Section 2.3.3) and represents sources responsible for approximately 95% of online news engagement.

From this dataset we consider two source-level attributes: *topics* and *trustworthiness scores*. We focus on four topics, defined as the set T introduced in Section 6.2.3:

- *Political news or commentary*
- *Conspiracy theories or hoaxes*
- *Sports and athletics*
- *Health or medical information*

Starting from the original dataset, we first select sources associated with exactly one topic.

Our objective is to obtain a list of 10 news outlets per topic while preserving the distribution of NewsGuard trustworthiness scores observed in the original dataset. To achieve this, we adopt a stratified sampling strategy based on the trustworthiness levels L defined by NewsGuard (see Table 25).

For example, consider sources associated with the topic *Sports and athletics*. Suppose that their trustworthiness score distribution in the NewsGuard dataset is as follows: 30% of sources have a score of 100, 50% fall within the range 75–99, and 20% fall within the range 60–74. To preserve this distribution, we randomly select 3 sources with $l = 100$, 5 sources with $l = 75$–99, and 2 sources with $l = 60$–74.

Assuming that the original NewsGuard dataset accurately represents the online media landscape, this stratified sampling procedure yields a set of 40 sources whose trustworthiness distribution closely matches the original dataset across the four selected topics.

Finally, we exclude sources that require a paywall to access their content and sources that do not publish in English. The resulting collection constitutes the set N of newspapers used in this study.

Table 25: NewsGuard trustworthiness levels

Score	Description
100	High credibility
75–99	Generally credible
60–74	Credible with exceptions
40–59	Proceed with caution
0–39	Proceed with maximum caution

Articles collection and cleaning

The next step consists in collecting a sample of articles for each selected source in order to construct the set A of articles defined in Section 6.2.3. In particular, we collect the textual content of the most recently published articles on each source's website.

The data collection process follows several steps. First, we identify the webpage that lists the published articles for each media outlet (often referred to as the *news history* page). Starting from the homepage addresses provided in the NewsGuard dataset, these pages are manually identified.

We then use the Selenium library[18] to develop a script that automatically scrolls through these pages and collects the URLs of individual news articles. For instance, on the website of *The Sun*[19], the news history page is available at:

https://www.the-sun.com/news/us-news/page/1/

The application systematically retrieves additional pages by increasing the page index (e.g., *page/2*, *page/3*, etc.), extracting the URLs of all articles listed on each page.

Once the list of article URLs is collected, we retrieve the HTML content of each article page. Specifically, we use the GNU Wget[20] command to download and store the full HTML pages in WARC-format archives. This ensures the availability of an offline copy of the webpages, facilitating subsequent content extraction.

[18] `https://www.selenium.dev/`
[19] `https://www.the-sun.com/`
[20] `https://www.gnu.org/software/wget/`

Table 26: Number of Articles per trustworthiness level, broken down by topic

Levels	Political	Conspiracy	Sports	Health
0 - 39	204	900	0	183
40 - 59	162	0	0	79
60 - 74	51	0	83	146
75 - 99	579	0	849	269
100	59	0	314	155
Total	**1055**	**900**	**1246**	**832**

The textual content of each article is then extracted using custom parsers developed for each news source. These parsers rely on XPATH-based rules tailored to the structure of each website.

We discard articles containing fewer than 200 words. For each news outlet we collect at least 40 articles, with a maximum of 294 articles per outlet. Overall, we identify 5,006 article URLs, of which the textual content of 4,033 articles is successfully extracted.

Because XPATH-based extraction may also capture extraneous text fragments, the resulting texts are further cleaned. In particular, two types of unwanted elements are removed: (i) repetitive phrases such as signatures, disclaimers, or slogans; and (ii) miscellaneous textual elements such as dates or navigation artifacts.

The cleaning procedure combines automated processing using the *spaCy*[21] library (specifically the `en_core_web_lg` model) with manual verification.

Article tagging

After the collection (performed between May 4 and May 15, 2023) and cleaning phases, each article is tagged according to the reliability score assigned by NewsGuard to its publisher. In other words, we adopt a source-based labeling strategy in which the reliability level of each article

[21] `https://spacy.io/`

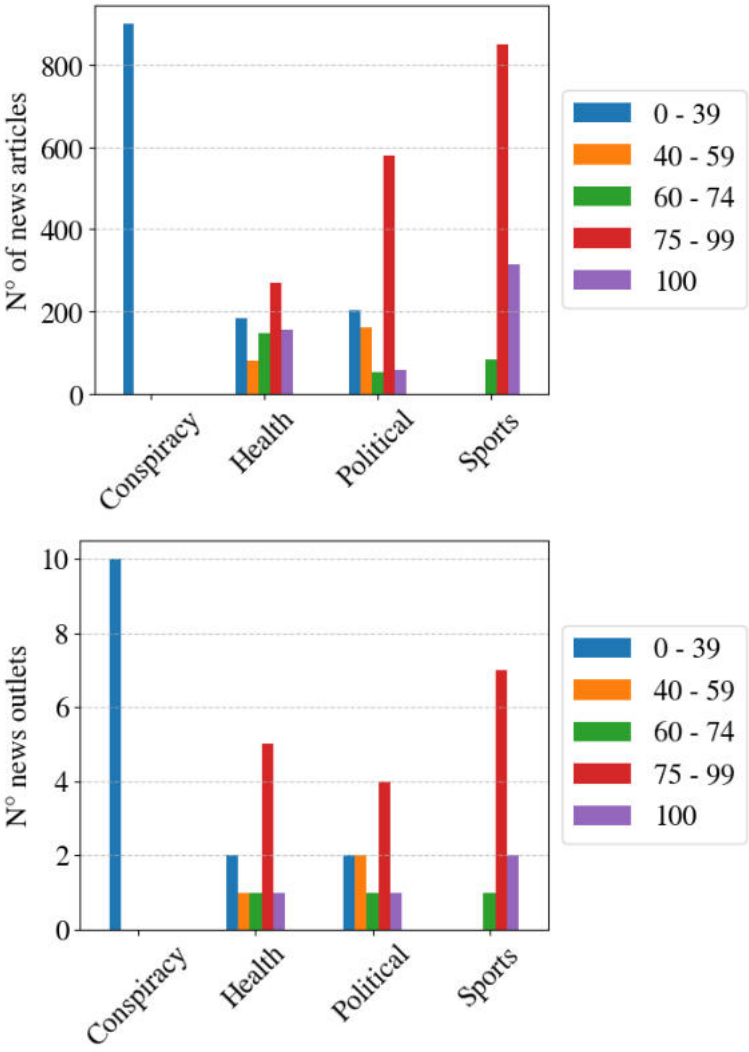


Figure 40: Number of articles (top) and news outlets (bottom) per trustworthiness level, broken down by topic.

is inherited from the NewsGuard score of the corresponding publisher.

The resulting dataset A, illustrated in Figure 40 and summarized in

Table 26, contains articles extracted from the selected media outlets N.

The figures show the distribution of trustworthiness levels l across the four topics. The *Conspiracy* topic is dominated by sources falling in the lowest credibility range. In contrast, the *Health or medical information* and *Political news or commentary* topics display a more balanced distribution across trustworthiness levels, with a prevalence of sources in the $l = 75$–99 range in the latter. The *Sports and athletics* topic contains the largest number of articles (1,246), with a strong concentration in the $l = 75$–99 interval.

Additional statistics describing the final dataset are reported in Table 26. This dataset constitutes the set A of articles used in the problem definition (see Section 6.2.3).

6.2.5 Results and discussion

We present the experimental setup and results for the tasks defined in Section 6.2.3, specifically, *topic* and *trustworthiness* detection. We use BERT [61], the well-known state-of-the-art pre-trained language model. We use the *transformers*[22] library in Python to deploy and fine-tune BERT, as well as to compute performance metrics. In particular, we adopt *BertForSequenceClassification* because it combines the capabilities of a highly trained language model with the adaptability to address specific tasks. For the validation, we adopt the 10-fold stratified cross-validation implemented by *scikit-learn*[23], thus using 10% of the dataset as test and 90% as training in each step. This choice guarantees an even distribution of the target class across each fold, which ensures a more robust evaluation.

Topic detection

Before we continue with the multi-class classification task, we show the results of the topic detection task. We remind the reader that the execution of this task and the evaluation of its results should not be considered

[22] `https://github.com/huggingface/transformers`

[23] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html`

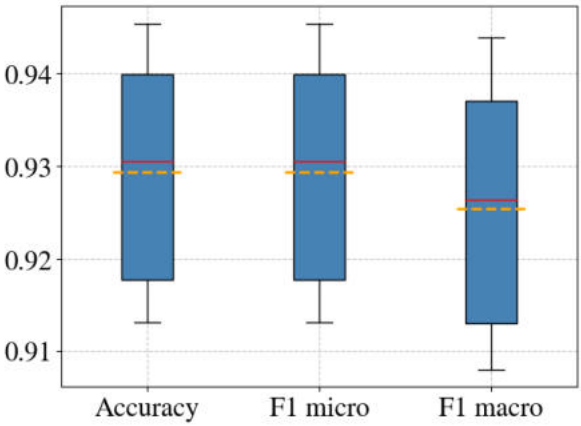


Figure 41: Evaluation results for topic detection. The yellow dotted line represents the average, while the red line represents the median.

as the main result of this article. The task is valuable for understanding the quality of the dataset annotations, which we did not provide ourselves.

Figure 41 shows the evaluation results on the 10 stratified folds. We achieve an average F1-macro of 0.925 (min = 0.908 and max = 0.943) and an average F1-micro of 0.929 (min = 0.913 and max = 0.945). For completeness, we also report the accuracy, precision and recall values in Figure **??** in the appendix, which are quite high for each topic.

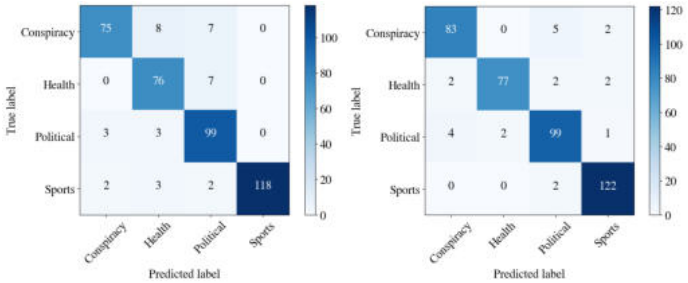


Figure 42: Topic: Confusion matrix for the fold with the lowest (left) and the highest (right) F1 macro

Figure 42 (left) shows the confusion matrix for the fold with the worst F1 macro score. The errors made by the classifier are mainly due to the misclassification of *Political news or commentary* articles as *Conspiracy theories or hoaxes*. This result is not surprising, since the lines between legitimate political news and conspiracy theories can be blurred by information manipulation strategies [66]. Also, some conspiracy articles are mistakenly categorized as related to health or medical information. This is not surprising since conspiracy theories often touch on topics related to public health, a phenomenon that has become more pronounced during and after the COVID-19 global pandemic [163]. This level of misclassification is not observed when examining the best-performing fold (Figure 42 right), where errors still exist, albeit to a lesser extent. Despite some inaccuracies, we argue that these results are satisfactory in that there is a very good match between articles and assigned topics.

Trustworthiness detection

This section presents the results of detecting articles' trustworthiness level as defined in Section 6.2.3. The experiments are performed on the final dataset described in Section 6.2.4, following the methods described at the beginning of this section.

The primary goal of this task is to develop a classifier capable of assigning a level of trustworthiness (trust_a) to each article (a) based on its textual content (text_a). These trustworthiness levels, which include five different categories identified and assigned by NewsGuard, provide a nuanced characterization of publisher-level trustworthiness (see Table 25). This is a multi-class classification task at the article level.

Our results, shown in Figure 43, demonstrate the strong capability of the model to accurately associate the article to one of the five trustworthiness levels. The model achieves an average F1-macro of 0.843 (min = 0.753 and max = 0.901) and an average F1-micro of 0.882 (min = 0.816 and max = 0.930).

We analyze the confusion matrices associated with the best and worst Macro F1 scores to gain deeper insight into the results. Figure 44 (top) shows the confusion matrix for the fold associated with the lowest Macro

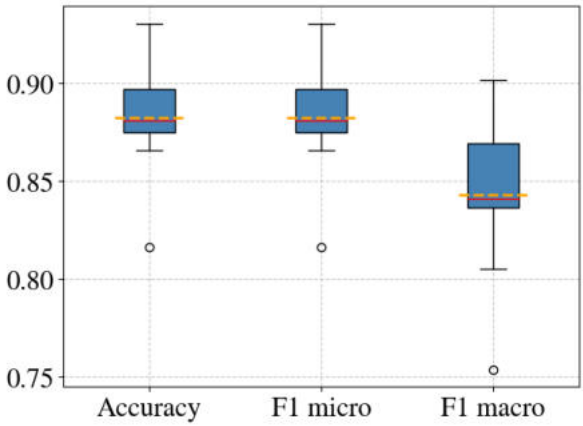


Figure 43: Evaluation results for trustworthiness level detection. The yellow dotted line represents the average, while the red line represents the median.

F1 score. Here we can observe two errors: 49 items assigned to neighboring classes and 25 items assigned to classes significantly different from the true ones. In the best case, as shown in Figure 44 (bottom), the situation is characterized by lower values, with 17 items assigned to adjacent classes and 11 to distant classes.

The importance of the two types of errors can vary depending on how we want to use the model. As mentioned earlier, the trustworthiness levels represent a nuanced characterization of trustworthiness. There may also be situations where a coarser classification is desired. For example, we might consider redefining NewsGuard's thresholds to produce only two levels of trustworthiness: the $0 - 59$ and $60 - 100$ ranges to identify untrusted and trusted publishers. Redefining the thresholds to create coarser levels of trustworthiness can improve the performance of our model, thereby increasing its practical utility in real-world scenarios.

6.2.6 Conclusions

In this Section we examine the quality of the online news landscape, with particular reference to the level of trustworthiness of the news source.

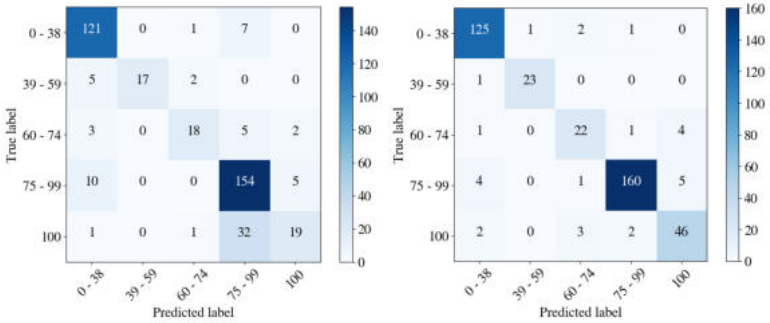


Figure 44: Trustworthiness level: Confusion matrix for the fold with the lowest (left) and highest (right) F1 macro

Many organizations, often formed by journalists and communication experts, have been trying for years to guide readers to read online media more trustworthyly by assigning trustworthiness ratings to various online newspapers. Of course, this process requires experienced annotators and is time-consuming. In this analysis, we have tried to speed up the work of these organizations by evaluating the quality of an automatic ranking of an article's trustworthiness. The results are very promising when compared to the few existing related works.

Our approach is not intended to replace the careful procedures of journalistic organizations that invest much time and manpower in ranking online news media. Instead, as introduced at the beginning of the Section, we believe that the proposed article ranking can provide such organizations with initial guidance, both in selecting articles for human annotators to analyze and gaining insight into completely unfamiliar media outlets. In addition, our model can suggest to users the similarity - or otherwise - of the news they are reading to news from less reputable sources.

Our work contributes to real-world applications to combat the spread and impact of low-credibility content. However, the proposed work can be extended in three main directions: i) enriching the dataset with more sources, including non-English ones, and adding more articles per

source; ii) exploring models other than BERT; and iii) integrating eXplainable Artificial Intelligence (XAI) techniques to understand textual differences between articles with different levels of trustworthiness and whether there is a specific reason why some sources are not classified correctly.

Chapter 7

Conclusions

This thesis addresses one of the most pressing and complex challenges of the digital age: the proliferation of disinformation and misinformation in online environments. Adopting a multidisciplinary approach at the intersection of network science, artificial intelligence and computational social science, it has contributed to both the theoretical understanding of disinformation dynamics and the development of scalable methodologies for identifying and mitigating low-credibility content on social media platforms and online news outlets.

Following a thorough review of current research (see Chapter 2), the thesis progresses along two main lines of investigation: (i) empirical analysis of the diffusion of disinformation, and (ii) development of computational methods for assessing the reliability of online news articles and outlets.

7.1 Disinformation dynamics across contexts

The first core contribution consists of an empirical investigation into how disinformation spreads across different geopolitical and socio-political contexts. In Chapter 3, large-scale analyses of Twitter/X data reveal how both user-level factors (e.g., political affiliation, automated activity) and contextual conditions (e.g., electoral systems, public health crises) shape

the dissemination of unreliable information.

Our analysis of two case studies, focusing on the spread of disinformation during the infodemic of the 2020s and the 2020 U.S. elections, demonstrates that this spread is systematically influenced by pre-existing political structures, user polarization, and coordinated behavior. They build upon existing literature and emphasize the interaction between digital platforms and real-world dynamics.

7.2 Methodologies for disinformation mitigation

The second major contribution is the development of original methodologies to automate the identification and mitigation of disinformation. These methods address the need for scalable and resource-efficient approaches to reliability assessment.

- **Echo Chamber Detection:** In Chapter 4 we propose a novel, entropy-based method for detecting echo chambers through the joint analysis of user interactions and content exposure [155]. This approach can identify ideologically polarized communities that play a key role in amplifying disinformation, particularly during politically sensitive events.

- **Reliability Assessment of Publishers and Articles:** Two complementary frameworks are introduced:

  - The first, in Chapter 5 leverages LLMs to replicate expert-based reliability assessments using structured journalistic criteria. The approach evaluates LLMs' ability to emulate human expert judgments, based on frameworks such as NewsGuard and GDI, and to resolve inter-annotator disagreements. While results show strong alignment, they also reveal model sensitivities (e.g., to political framing), prompting reflection on the limitations of automated reliability evaluation. A further contribution [17] (joint work with another PhD student) applies a BERT-based classifier at the level of individual articles to

assess the reliability of publishers. This method allows for a quick and effective reliability profiling of the news outlet.

– The second, in Chapter 6 infers the reliability level of an online publisher using user–publisher interaction patterns. It identifies a subset of users, called *voters*, whose behavior acts as a proxy for news reliability. This model-independent method, which requires no access to article text or labels, forms the basis of the `TROPIC` software tool [156], which has been publicly released for research purposes.

These methodologies can be integrated into a modular and extensible framework for reliability assessment that jointly leverages social interaction signals, textual analysis, and model-based reasoning.

The overarching goal of this integration is to promote cohesive, multidisciplinary automated approaches to combating online disinformation, thereby avoiding the fragmentation of existing methods [189].

Such a framework supports real-world applications including content moderation, platform governance, and media literacy initiatives, while simultaneously reducing reliance on manual annotation.

7.3 Meta-contribution: a unified framework for disinformation mitigation

Beyond the individual methods and findings, this thesis advances a *unifying framework* for the study and mitigation of online disinformation. This framework integrates empirical evidence and methodological tools into a model designed to capture the socio-technical mechanisms that drive the circulation of false, distorted, or biased narratives.

Conceived as a foundation for developing tools to counter disinformation, the framework emphasizes the interplay of four key dimensions:

- **Content quality:** reliability of news articles and outlets, used as indicators of information quality;

- **User-related factors:** ideological alignment (e.g., political orientation), engagement patterns, and individual susceptibility to misleading information;

- **Contextual factors:** divisive socio-political events (e.g., elections, public health crises) and structural elements (e.g., electoral systems);

- **Account characteristics:** differences between human-operated and automated accounts.

These dimensions are realized through two complementary and interacting components:

1. A **network science component**, which investigates social media interactions within specific online debates to: (i) derive group-level structural indicators of user-related dimensions (such as ideological alignment) by identifying communities in echo chambers and discursive groups; (ii) detect influential narratives, engagement dynamics, and key spreaders.

2. An **automated reliability evaluation component**, which assesses the credibility of accounts, news content, and outlets by combining expert-defined criteria with artificial intelligence techniques. This component enables continuous, scalable, and automated assessments of information quality.

Taken together, these two components provide a scalable mechanism for monitoring disinformation. Methodologically, the framework integrates network science and artificial intelligence to illuminate how individual susceptibility to disinformation emerges in online environments. Strategically, it bridges the gap between analysis and intervention, offering a foundation for future tools and policies aimed at both understanding and countering online disinformation.

7.4 Future extension of TROPIC

The prototype TROPIC (Trustworthiness Rating of Online Publishers through online Interactions Calculation), for the analysis of online news publishers, operationalizes some of the main contributions of this thesis, providing expert annotators with a tool to (i) guide and prioritize annotation efforts, and (ii) support news publishers' reliability assessment. Future developments may further align the system with the framework introduced in Section 7.3, by integrating methods presented throughout the thesis:

- **Echo Chambers and Discursive Communities (Chapters 3 and 4):** Detecting echo chambers and discursive communities would enable the system to identify users' inclinations and the content shared within ideologically homogeneous clusters. Such information provides valuable indicators of bias, credibility, and content quality, supporting the prioritization of information flows and facilitating large-scale detection and mitigation of mis- and disinformation.
- **LLM-Based Evaluation of Journalistic Criteria (Section 5.2):** Incorporating LLM-based evaluations can complement expert judgment through synthetic annotators that assign reliability scores to articles or articles' publishers. Visual outputs such as score distributions would help experts identify statistically anomalous or contentious cases, while enabling semi- or fully automated assessments (see also Section 6.2).

These extensions would enhance the two core functionalities of TROPIC:

- *Guiding annotation efforts.* The current system highlights News Engagement Communities (NECs)—clusters of highly shared URLs, to identify publishers attracting disproportionate attention. Adding echo chamber and discursive community detection would refine this process by accounting for ideological homogeneity and community influence, while LLM-based evaluations could dynamically flag content along dimensions such as bias, factuality, or sensationalism.

- *Supporting reliability rating.* At present, source-level reliability assessment relies solely on expert annotation. Future versions could incorporate automated, criterion-based evaluations at both news and source levels (via aggregation). Presenting experts with score distributions per criterion would assist decision-making, while validated automatic pipelines could enable real-time monitoring or even replace manual ratings in high-volume, low-resource scenarios.

7.5 Future Research Directions

Three main directions are envisioned for future research:

1. Simulating Human Susceptibility to Disinformation with LLMs

Ongoing work investigates whether LLMs can be conditioned to reflect human cognitive diversity, e.g., by simulating individual differences in personality traits. Initial results from [154] suggest that models emulating traits like Conscientiousness and Openness replicate known patterns of resistance to disinformation, while other traits (e.g., Neuroticism) show limitations. This research aims to simulate heterogeneous users' judgment processes to identify vulnerabilities and design targeted interventions.

2. Subjective Evaluation of Information Quality

Moving beyond expert-based ground truths, a second line of work explores the subjective perception of quality, bias, and harm. LLMs are used to simulate virtual annotators representing diverse socio-cultural profiles, enabling modeling of:

- Perceived political bias;
- Sensationalistic or manipulative framing;

- Harmful narratives targeting vulnerable groups.

This will allow for the generation of *perceived quality distributions*, informing fairness-aware and context-sensitive credibility evaluation systems.

3. Applications and Societal Impact

The above-mentioned research directions support two primary applications:

- **Personalized Media Literacy Tools:** Simulated user profiles can power adaptive training interventions that respond to individual vulnerabilities, enhancing critical thinking and digital resilience.
- **Impact Forecasting and Risk Assessment:** Comparing objective and perceived quality evaluations enables forecasting of content impact across socio-cultural groups, supporting content moderation, platform regulation, and public communication strategies.

Appendix A

A

A.1 Detailed user NEC in Italian COVID-19 vaccination debate on Twitter/X

prendere da version iniziale di echo-chamber detection

A.2 Validated vs non-validated discursive communities

Let us summarize the procedure for inferring the presence of discursive communities (DiCo) in our dataset, as described in Section 2.3.2. Our approach focuses on the bipartite network of verified vs. unverified accounts, where a link represents the presence of at least one retweet from the unverified to the verified user. The network is then projected into the layer of verified users, resulting in a monopartite network in which the weights of the link represent the number of common (unverified) retweeters, i.e. the co-occurrences. Finally, the network is validated by comparing the empirical values with a maximum entropy null model (the BiCM [173]), including the information of the bipartite degree sequences.

At first glance, the validation procedure may seem like an unneces-

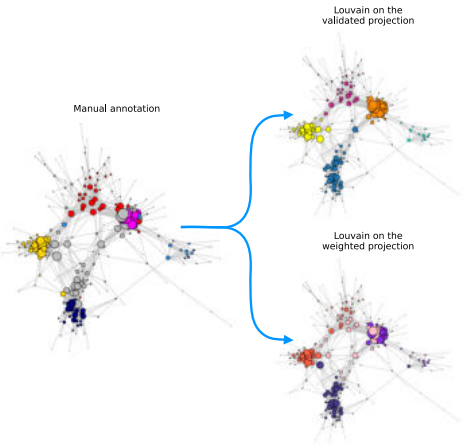


Figure 45: Comparison between the results of different community detections on the validated network of verified users. On the left, only politicians' accounts are colored according to their political affiliation (other verified accounts are gray). The first observation is that politicians with similar orientations cluster together in the validated projection. In this sense, a community detection run on this network returns partitions that are coherent with these political clusters (top right panel; nodes with the same color belong to the same community). The same is not quite true for a community detection algorithm run on the non-validated projection: in the latter case, the partitions only partially capture the political orientations present (lower right panel; again, nodes with the same color belong to the same community).

sary complication. The goal of the analysis is to extract similarities in the creation of new content based on common audiences, and it can be argued that even without extracting the significant structure of the network, the standard algorithms for community detection can find the relevant network structure.

Before directly comparing the results in the case of our dataset, let us first provide a methodological argument in favor of using the validated projection instead of the entire projection network. As mentioned above, the output of the procedure is a monopartite network in which

connections are present if the co-occurrences cannot be explained by the bipartite degree sequences. In this sense, the structure of the network is inferred by discounting the *original* bipartite information. If, instead, the projection network is not validated, the communities in the network are inferred using the information about the projected network, i.e., some kind of information *derived* from the original bipartite system. Note also that knowing the value of the co-occurrences does not allow going back to the bipartite structure of the system and causes a loss of information [96]. In this sense, the use of the original information available from the data should be preferred.

Nevertheless, the implications of such a choice could still be limited in our dataset and, therefore, we will examine the results of the different approaches. The first observation, already highlighted in many papers [**Mattei2022**, 9, 27, 33, 35, 107, 165, 166], is that, when the debate is political or societal (as in the case of our dataset), the accounts of politicians and political parties tend to cluster, according to their orientation, in the validated network of verified users. This is also the case for our dataset, as can be seen in the left panel of Fig. 45: the colored nodes represent the accounts of political parties and politicians, where the color is related to their political alliance[1]. The only exceptions are some Italia Viva accounts that are merged with some center-left politicians. Such behavior is justified by the fact that Italia Viva was created by politicians who left the PD because they were not satisfied with the current leadership. In this sense, it is not surprising to find links between former party members.

The Louvain algorithm, run on the validated projection, captures such groups, see the top right panel of Fig. 45 (nodes displaying the same colors belong to the same community).

[1]In dark yellow, the Movimento 5 Stelle; in dark blue, the right-wing parties Lega and Fratelli d'Italia; in sky blue, the center-right party Forza Italia; in magenta, the center-left party Italia Viva; in red, the democratic alliance, including PD (the Italian Democratic Party), +Europa, the Socialist Party, and the Green Party. In gray, other verified users whose political orientation is not given *a priori*, such as journalists, media, artists, NGOs, etc.

Even if running the (weighted) Louvain algorithm on the entire co-occurrence network yields, by definition, different results, they could still provide a coherent partition of the validated projection, since it represents the core of the co-occurrence network. Remarkably, discounting inferred information has a cost: the obtained partition is less coherent with the political orientations of the verified users than the former one, see the lower right panel of Fig. 45. For example, Movimento 5 Stelle and the center-left alliance are mixed. The situation is even worse for Italia Viva, which is split in 2, partly joining the center-left alliance accounts and partly mixed with Forza Italia. In this sense, we can say that the community detection on the validated projection gives cleaner partitions than those calculated on the non-validated network. Finally, comparing modularities computed on different types of networks is not particularly informative, but it can still give a rule-of-thumb idea about the organization of the network: in the case of the validated network, the modularity is $Q \simeq 0.66$, while in the case of the non-validated network, it is $Q \simeq 0.17$[2]. In this sense, the validated network has a more modular structure.

In summary, in the validated projection of verified users, politicians and political parties cluster according to their political affiliation, and therefore a community detection algorithm running on the validated projection will capture these groups. Instead, a community detection algorithm running on the entire co-occurrence network of verified users, where co-occurrences is the number of common unverified retweeters, adds some noise to the partition found, and the division between opposing groups is less clean.

[2]Note that the null models implemented by the two Louvain community detection algorithms are different. On the binary validated network, it is the standard binary configuration model, which considers the information of the bipartite degree sequences. On the total co-occurrence network, it is the weighted configuration model, thus including the information of the strength sequence.

Appendix B

B

B.1 Italian socio-political situation during the period of data collection

Here, we present the social context in which the analysis presented in 3.1 is set. This subsection is divided into two parts: the contagion evolution and the political situation. These two aspects are closely related.

B.1.1 Evolution of the Covid-19 pandemic in Italy

A first Covid-19 outbreak was detected in Codogno, Lodi, Lombardy region, on February 19th, 2020[1]. In the very next day, two cases were detected in Vò, Padua, Veneto region. On February 22nd, in order to contain the contagions, the government decided to put in quarantine 11 municipalities, 10 in the area around Lodi and Vò[2]. Nevertheless, the number of contagions raised to 79, hitting 5 different regions; one of the infected person in Vò died, representing the first registered Italian Covid-19 victim[3]. On February 23rd, there were already 229 confirmed cases

[1] Prima Lodi, ""Paziente 1", il merito della diagnosi va diviso... per due", 8th June 2020

[2] Italian Gazzetta Ufficiale, "DECRETO-LEGGE 23 Febbraio 2020, n. 6". The date is intended to be the very first day of validity of the decree.

[3] Il Fatto Quotidiano, "Coronavirus, è morto il 78enne ricoverato nel Padovano. 15 contagiati in Lombardia, un altro in Veneto", 22nd February 2020.

in Italy. The first lockdown should have lasted until the 6th of March, but due to the still increasing number of contagions in northern Italy, the Italian Prime Minister Giuseppe Conte intended to extend the quarantine zone to almost all the northern Italy on Sunday, March 8th[4]: traveling to and from the quarantine zone was limited to case of extreme urgency. A draft of the decree announcing the expansion of the quarantine area appeared on the website of the Italian newspaper *Corriere della Sera* on the late evening of Saturday 7th, causing some panic in the interested areas[5]: around 1000 people, living in Milan, but coming from southern regions, took trains and planes to reach their place of origin[6][7]. In any case, the new quarantine zone covered the entire Lombardy and partially other 4 regions.

Remarkably, close to Bergamo, Lombardy region, a new outbreak was discovered and the possibility of defining a new quarantine area on March 3rd was considered: this opportunity was later abandoned, due to the new northern Italy quarantine zone of the following days. This delay seems to have caused a strong increase in the number of contagions, making the Bergamo area the most affected one, in percentage, of the entire country[8]; at time of writing, there are investigations regarding the responsibility of this choice.

On March 9th, the lockdown was extended to the whole country, resulting in the first country in the world to decide for national quarantine[9]. Travels were restricted to emergency reasons or to work; all business activities that were not considered as essentials, as pharmacies and

[4]BBC News, "Coronavirus: Northern Italy quarantines 16 million people", 8th March 2020"

[5]The Guardian, "Leaked coronavirus plan to quarantine 16m sparks chaos in Italy", 8th March 2020

[6]il Messaggero, "Coronavirus, a Milano la fuga dalla "zona rossa": folla alla stazione di Porta Garibaldi", 8th March 2020"

[7]repubblica.it, "Coronavirus, l'illusione della grande fuga da Milano. Ecco i veri numeri degli spostamenti verso sud", 23rd April 2020

[8]sky.com, "Coronavirus: Italian army called in as crematorium struggles to cope with deaths", 19th March 2020.

[9]BBC News, "Coronavirus: Italy extends emergency measures nationwide", 10th March 2020

supermarkets, had to be closed. Until the 21st of March, lockdown measures became progressively stricter all over the country. Starting from the 14th of April, some retails activities, as children clothing shops, reopened. A first fall in the number of deaths was observed on the 20th of April[10]. A limited reopening started with the so-called "Fase 2" (*Phase 2*) on the 4th of May[11].

From the very first days of March, the limited capacity of the intensive care departments caused a re-organization of Italian hospitals, leading, e.g., to the opening of new intensive care departments[12]. Moreover, new communication forms with the patient relatives were proposed, new criteria for the intubated patients were developed, and, in the extreme crisis, in the most infected cases, the emergency management took to give priority to the hospitalisation to patients with a higher probability to recover[13].

Outbreaks were mainly present in hospitals [134]. Unfortunately, healthcare workers were contaminated by the virus[14]. This contagion resulted in a relative high number of fatalities: by the 22nd of April, 145 Covid deaths were registered among doctors. Due to the pressure on the intensive care capacity, even the healthcare personnel was subject to extreme stress, especially in the most affected areas[15].

B.1.2 Italian political situation during the pandemic

On August 8th 2019, the leader of Lega, the main Italian right wing party, announced to negate the support to the government of Giuseppe Conte,

[10] Al Jazeera, "Italy sees first fall of active coronavirus cases: Live updates", 20th April, 2020.

[11] Repubblica.it, "Coronavirus in Italia, verso primo ok spostamenti dal 4/5, non tra Regioni. Conte: "Non è un liberi tutti"", 22nd April 2020

[12] The New York Times, "Italy's Health Care System Groans Under Coronavirus — a Warning to the World", 12th March 2020.

[13] Il Corriere della Sera, "Coronavirus, il medico di Bergamo: "Negli ospedali siamo come in guerra. A tutti dico: state a casa"", 9th March 2020.

[14] Ansa.it, "Coronavirus: Ordini degli infermieri, 4 mila i contagiati", 29th March 2020.

[15] Internazionale, "Il dolore invisibile dei medici in corsia contro il coronavirus", 1st April 2020.

which was formed after a post-election coalition between the Movement 5 Stars -M5S- and the Lega. The Prime Minister Giuseppe Conte resigned on the 20th of August and opened to the political crisis. After few days of negotiation, M5S, the most represented party in the Italian parliament, agreed to form a new government with the Italian Democratic Party (*Partito Democratico*, PD). PD, on the other hand, agreed, upon the suggestion of the former secretary and Prime Minister Matteo Renzi. After the formation of the new government, again led by Giuseppe Conte, Matteo Renzi formed a new center-left party, *Italia Viva* (*Italy alive*, IV), due to some discord with PD; despite the split, Italia Viva continued to support the actual government, having some of its representatives among the ministers and undersecretaries, but often marking its distance with respect to both Pd and M5S.

Due to the great impact that Matteo Salvini and Giorgia Meloni - leader of Fratelli d'Italia, a right wing party- had on social media, they started a massive campaign against the government the day after its inauguration.

The regions of Lombardy, Veneto, Piedmont and Emilia-Romagna experienced the highest number of contagions during the pandemic; among those, the former 3 were administrated by the right and center-right wing parties, the fourth one by the PD. The disagreement in the management of the pandemic between regions and the central government was the occasion to exacerbate the political debate (in Italy, regions have a quite wide autonomy for healthcare). The regions administrated by the right wing parties criticised the centrality of the decisions regarding the lock down, while the national government criticised the health management (in Lombardy, the healthcare system has a peculiar organisation, in which the private sector is supported by public funding) and its ineffective measure to reduce the number of contagions. The debate was ridden even at a national level: the opposition criticized the financial origin of the support to the various economic sectors. Moreover, the role of the European Union in providing funding to recover Italian economics after the pandemic was debated.

B.2 Composition of the subcommunities in the validated network of verified Twitter users

Here, we detail the composition of the communities shown in Figure 1 of the main text. We remind the reader that, after applying the Louvain algorithm to the validated network of verified Twitter users, we could observe 4 main communities, that correspond to

1. Center right/Right wing parties and media (in steel blue);
2. Center-left wing (dark red);
3. 5 Stars Movement (*M5S*) (in dark orange);
4. Institutional accounts (in sky blue).

Starting from the center-left wing, we can find a slightly lighter red community, including various NGOs (the Italian chapter of UNICEF, Medecins Sans Frontieres, Action Aid, Emergency, Save the Children), various left-oriented journalists, VIPs and pundits[16]. Finally, we can find in this group political movements ('6000sardine') and politicians on the left of PD (as Giuseppe Civati, Pietro Grasso, Ignazio Marino) or on the left current of the PD (Laura Boldrini, Michele Emiliano, Stefano Bonaccini). A darker red sub-community turns out to be composed by the main politicians of the Italian Democratic Party (PD), as well as by representatives of the European Parliament (Italians and others) and some EU commissioners. The magenta group is mostly composed by the representatives of the newly founded Italia Viva, by the former Italian Prime Minister Matteo Renzi (December 2014 - February 2016) and former secretary of PD. In golden red, we can find the subcommunity of Catholic and Vatican groups. Finally, the dark violet red and light tomato subcommunities are composed mainly by journalists. Interestingly enough, the dark violet red contains also accounts related to the city of Milan (the

[16] As the cartoonists Makkox and Vauro, the singers Marracash, FrankieHiNRG, Ligabue and "il Volo" vocal band, and journalists from Repubblica (Ezio Mauro, Carlo Verdelli, Massimo Giannini), from La7 TV channel (Riccardo Formigli, Diego Bianchi).

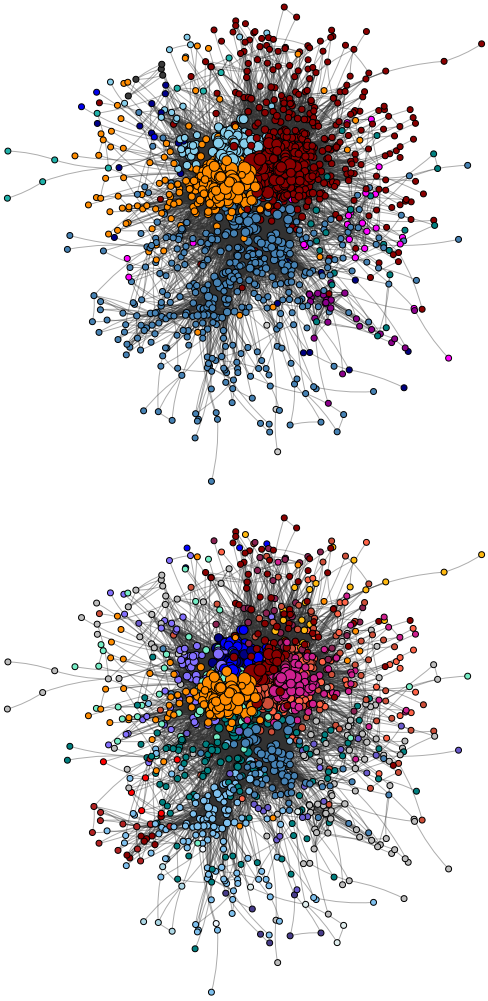


Figure 46: Validated projection of the bipartite network of verified/unverified accounts. In the top panel, the monopartite projection in which just communities are displayed. In the bottom panel, the subcommunities, obtained by rerunning the Louvain algorithm in each of the former 4 main communities.

major, the municipality, the public services account) and to the spoke person of the Chinese Minister of Foreign Affair.

In turn, also the orange (M5S) community shows a clear partition in substructures. In particular, the dark orange subcommunity contains the accounts of politicians, parliament representatives and ministers of M5S, as well as journalists close to the party, and the official account of *Il Fatto Quotidiano*, a newspaper explicitly supporting the M5S. Since one of the main leaders of the Movement, Luigi Di Maio, was also the Italian Minister of Foreign Affairs, we can find in this subcommunity also the accounts of several Italian embassies around the world, as well as the account of the Italian representatives at NATO, OCSE and OAS. In aquamarine, we can find the official accounts of some private and public, national and international, health institutes (as the Italian Istituto Superiore di Sanità, literally the *Italian National Health Institute*, the World Health Organization, and the Fondazione Veronesi), the Minister of Health Roberto Speranza, and some foreign embassies in Italy. Finally, in the Light Slate Blue subcommunity, we can find various Italian ministers as well as the Italian police and army forces.

Similar considerations apply to the steel blue community. In steel blue, the subcommunity of center-right/right wing parties (as Forza Italia, Lega and Fratelli d'Italia). The presidents of Lombardy, Veneto and Liguria, administrated by center-right/right wing parties, can be found here. (In the following this subcommunity is going to be called as FI-L-FdI, recalling the initials of the political parties contributing to this group.) The sky blue subcommunity includes the national federations of various sports, the official accounts of athletes and sport players (mostly soccer players) and their teams, as well as sport journals, newscasts and journalists. The teal subcommunity contains the main Italian news agencies, some of the main national and local newspapers, newscasts and their journalists. In this subcommunity, there are also accounts of many universities; finally, it includes also local public service newscasts. The firebrick subcommunity contains accounts related to the AS Roma football club; analogously, in dark red, official accounts of AC Milan and its players. The slate blue subcommunity is mainly composed by the official

accounts of radio and TV programs of Mediaset, the main private Italian broadcasting company, together with singers and musicians. Other smaller subcommunities include other sport federations and sports pundits.

Finally, the sky blue community is mainly composed by Italian embassies around the world. The navy subpartition contains also the official accounts of the President of the Republic, the Italian Minister of Defense and the one of the Commissioner for Economy at EU and former Prime Minister, Paolo Gentiloni.

B.3 Domain analysis for the validated network of verified users

Table 27 shows the percentage of the different types of domains for the 4 communities identified in the top panel of Fig. 46.

Community	#url	T	∼T	N	S	F	M	P	IS	ST	SE	UNC
only tweets												
steel blue	22029	74.5	0.9	2.7	3.3	0.1	0.0	0.7	0.0	0.0	0.0	17.8
dark red	9185	79.0	2.0	0.1	1.6	0.1	0.0	0.6	0.3	0.0	0.0	16.3
dark orange	3437	54.1	0.2	0.2	6.1	0.1	0.0	0.9	1.6	0.3	0.0	36.5
sky blue	1106	65.8	0.0	0.0	6.2	0.0	0.0	0.1	0.0	0.0	0.0	27.9
only retweets												
steel blue	2481	69.7	0.9	3.2	4.4	0.4	0.0	0.1	0.0	0.0	0.0	21.3
dark red	3563	71.4	1.9	0.1	3.7	0.4	0.0	0.6	0.6	0.0	0.0	21.3
dark orange	2202	41.0	0.5	0.9	8.7	0.4	0.0	0.6	1.4	0.7	0.0	45.8
sky blue	1051	38.3	1.5	0.1	12.7	0.3	0.0	0.1	0.8	0.0	0.1	46.1

Table 27: Annotation per communities – validated network of verified users. The colors are those of the greatest communities of the top panel of Fig. 46. Steel blue represent the discursive community of Media and center-right/right wing parties; in dark red, the center-left wing parties and their supporters; in dark orange, the supporters of Movimento 5 Stelle and, in sky blue, the official government accounts. The description of the various columns can be found in the Table 1 and Section 2.3.3. The presence of many more tweets than retweets may be surprising: actually, it is typical of verified users focusing their production in original messages, as already observed in [9, 34, 165].

Table 28 shows that the steel blue community (including both politicians and Media) is the most active one, even if it is not the most rep-

Community	#post	#url	#dist url	#domain	#user
steel blue	30877	24510	20718	648	417
dark red	17202	12748	10999	744	452
dark orange	8990	5639	4389	640	316
sky blue	3897	2157	1626	348	149
`only tweets`					
steel blue	26359	22029	19222	467	297
dark red	11275	9185	8435	430	329
dark orange	5240	3437	3042	351	245
sky blue	1738	1106	964	143	114
`only retweets`					
steel blue	4518	2481	2175	348	328
dark red	5927	3563	3050	483	399
dark orange	3750	2202	1633	423	264
sky blue	2159	1051	740	269	147

Table 28: Posts, urls, domains and users statistics per communities – validated network of verified users. The frequency of posts in the steel blue community is originated by the presence of Media in this group. Nevertheless, as we will see in Table 29, even the political subcommunity contained in the steel blue group is particular prolific.

resented: the number of users is lower than the one of the center-left community (the biggest one, in terms of numbers), but the number of posts containing a valid url is almost the double of that of the center-left group. The activity of steel blue verified users is more focused on content production (see the `only tweets` sub-table) than on sharing (see the `only retweets` sub-table). Retweets represent almost 14.6% of all posts from Media and right wing community, while in the case of the center-left community the value is 34.5%. This effect is observable even in the average `only tweets` post per verified user: a right-wing user and a Media user have an average of 88.75 original posts, against 34.27 for center-left users. These numbers are probably due to the presence, in the former community, of the Italian most accessed media, that spread their (original) pieces of news on Twitter. Table 29 shows the domain annotation per political sub-communities. The presence of urls from a

non reputable source in the steel blue community is more than 10 times higher than the second score in the same field for (`only tweets`). It is worth noting that, for the case of the dark orange and sky blue communities, which are smaller both in terms of users and number of posts, the presence of non classified sources is quite strong (it represents nearly 46% of the posts retweeted, for both the communities), as it is the frequency of posts linking to social network contents. Interestingly enough, verified users of both groups seem to focus slightly more on the same domains: there are, on average, 1.59 and 1.80 posts for each url domain, respectively for the dark orange and sky blue communities, and, on average, 1.26 and 1.34 posts for the steel blue and the dark red communities.

Subcommunity	#url	T	∼T	N	S	F	M	P	IS	ST	SE	UNC
FI-L-FdI	4759	56.4	2.3	12.8	14.5	0.1	0.0	3.1	0.0	0.0	0.0	10.8
Movimento 5 Stelle	2385	75.5	0.1	0.4	6.6	0.0	0.0	1.9	0.0	1.1	0.0	14.4
Italia Viva	857	25.3	26.6	0.1	10.0	0.7	0.1	8.4	0.5	0.0	0.0	28.3
Partito Democratico	643	64.4	0.6	0.3	9.2	0.8	0.0	0.0	3.6	0.0	0.2	20.9
`only tweets`												
FI-L-FdI	4177	59.0	2.1	13.0	14.6	0.0	0.0	3.5	0.0	0.0	0.0	7.8
Movimento 5 Stelle	1839	79.4	0.1	0.4	6.3	0.0	0.0	1.7	0.0	0.6	0.0	11.5
Italia Viva	458	19.2	39.1	0.2	9.0	0.2	0.2	11.4	0.0	0.0	0.0	20.7
Partito Democratico	370	71.9	0.5	0.5	5.4	1.1	0.0	0.0	3.8	0.0	0.0	16.8
`only retweets`												
FI-L-FdI	582	38.0	3.4	11.7	14.1	0.3	0.0	0.3	0.0	0.0	0.0	32.2
Movimento 5 Stelle	546	62.3	0.2	0.4	7.9	0.2	0.0	2.6	0.0	2.9	0.0	23.5
Italia Viva	399	32.3	12.3	0.0	11.3	1.3	0.0	5.0	1.0	0.0	0.0	36.8
Partito Democratico	273	54.2	0.7	0.0	14.3	0.4	0.0	0.0	3.3	0.0	0.4	26.7

Table 29: Domains annotation per political subcommunities - validated network of verified users. The incidence of reputable sources strongly reduces in the retweets for all the subcommunities, but Italia Viva. We argue that verified users are more cautious when writing their original messages, while they are more relaxed when sharing other messages. The references to Social Networks (S) are relatively strong in all the subcommunities. The description of the various columns can be found in the Table 1 and Section 2.3.3.

B.3.1 Hashtags by verified users

Figures 47 and 48 report statistics about the most diffused hashtags in the 4 political subcommunities. Actually, from the various hashtags, we can derive important information regarding the political discursive communities and their view about the pandemic and its management. First, M5S is the greatest user of hashtags: the two most used hashtags have

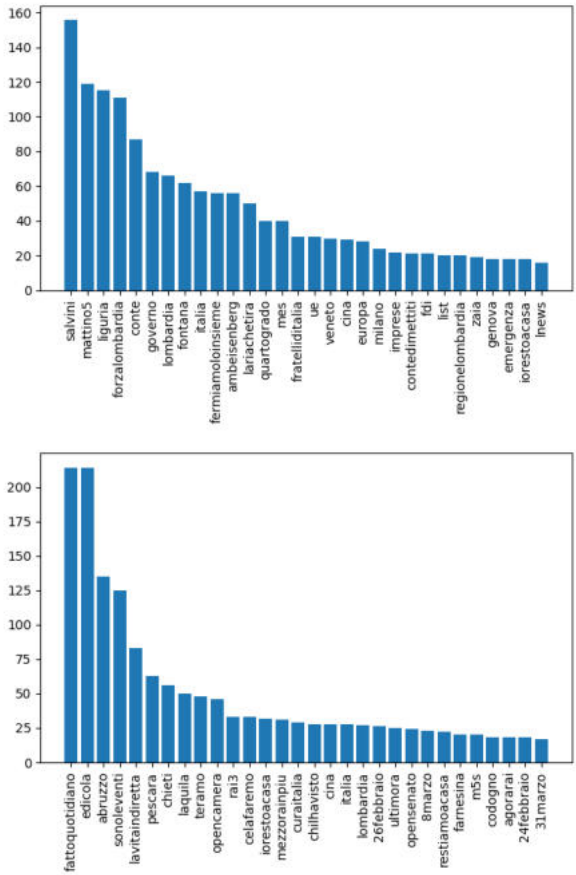


Figure 47: The 30 most diffused hashtags in the political subcommunities. Verified users. Top panel: right and center-right wing; bottom panel: 5 Stars Movement.

been used almost twice the most used hashtags by PD, for instance. This heavy usage is probably due to the presence in this community of journalists and of the official account of *Il Fatto Quotidiano*, a newspaper explicitly supporting M5S: indeed, the first two hashtags are "#ilfattoquotidiano" and "#edicola" (*kiosk*, in Italian).

There is a relevance of hashtags intended to encourage the population during the lockdown: it is the case of "#celafaremo" (*we will make it*), "#iorestoacasa" (*I am staying home*), "#fermiamoloinsieme" (*Let's stop it together*): "#iorestoacasa" is present in every community, but it ranks 13th in the M5S political community, 29th in the FI-L-FdI community, 2nd in the Italia Viva community and 10th in the PD one. Remarkably, "#celafaremo" is present only in the M5S group, as "#fermiamoloinsieme" can be found in the top 30 hashtags only in the center-right/right wing cluster. The PD, being present in various European institutions, mentions more European Union related hashtags ("#europeicontrocovid19", *Europeans against covid-19*), in order to ask for a common reaction of the EU. The center-right/right wing community has other hashtags as "#forzalombardia" (*Go, Lombardy!*; Lombardy region is administrated by a coalition of center-right and right wing parties.), ranking 2nd, and "#fermiamoloinsieme", ranking 10th. What is, nevertheless, astonishing, is the presence, among the most used hashtags in all communities, of the pair [politician/TV program] (as "#mattino5", "#lavitaindiretta", "#ctcf", "#dimartedì"). as if the main usage of hashtags is to promote the appearance of politicians in TV programs. Finally, hashtags by FI-L-FdI are mainly used to criticise the actions of the government, e.g., "#contedimettiti" (*Conte, resign!*).

B.4 Domain analysis for the directed validated network

Table 30 shows the number of tweets and retweets containing a url, and the tag assigned to the corresponding domain, for the directed validated network.

B.4.1 Hashtags by validated users

Figures 49 and 50 show the top 30 shared hashtags, for the various political subcommunities: the scales are different, due to the different activity of the various groups. Nevertheless, it is interesting to con-

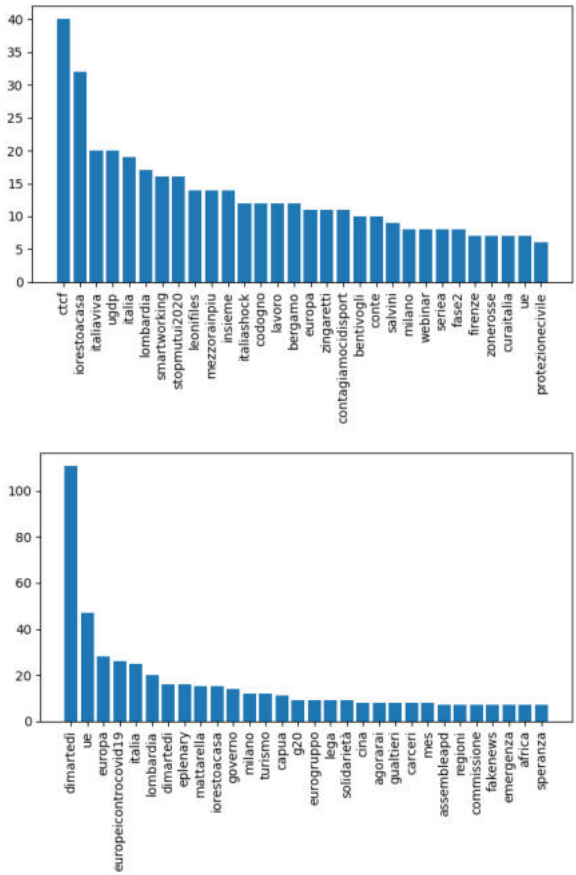


Figure 48: The 30 most diffused hashtags in the political subcommunities. Verified users. Top panel: Italia Viva (center-left); Bottom panel: Democratic Party (center-left).

sider the most used hashtags in the various subcommunities in order to have an idea of the standings of the different parties. The opposition, represented by FI-L-FdI, shows dissatisfaction by using hashtags like '#contedimettiti' (*Conte, resign!*), '#governodellavergogna' (*government of the disgrace*), '#governodelcontagio' (*government of the contagion*)

Sub-community	#url	T	∼T	N	S	F	M	P	IS	ST	SE	UNC
FI-L-FdI	457746	38.3	12.1	22.1	4.7	0.1	0.0	0.3	0.0	0.0	0.3	22.1
Italia Viva	155125	58.7	6.7	0.7	3.6	0.5	0.1	0.6	0.6	0.0	0.0	28.5
Movimento 5 Stelle	120244	63.8	1.4	3.1	4.7	0.2	0.0	2.8	0.1	0.1	0.0	23.8
Partito Democratico	6183	47.5	1.5	0.4	5.9	1.0	0.0	0.1	2.9	0.0	0.0	40.7
only tweets												
FI-L-FdI	95902	29.5	9.6	30.6	4.7	0.2	0.0	0.2	0.0	0.0	0.0	25.2
Italia Viva	33648	47.8	14.4	1.1	2.9	0.5	0.0	0.5	0.0	0.0	0.0	32.8
Movimento 5 Stelle	22940	56.3	1.4	2.7	3.9	0.5	0.0	1.4	0.0	0.1	0.0	33.7
Partito Democratico	1759	35.6	0.9	0.2	3.5	0.8	0.0	0.0	1.2	0.0	0.0	57.8
only retweets												
FI-L-FdI	361844	40.7	12.8	19.9	4.8	0.1	0.0	0.4	0.0	0.0	0.4	20.9
Italia Viva	121477	61.8	4.6	0.6	3.7	0.5	0.1	0.7	0.7	0.0	0.0	27.3
Movimento 5 Stelle	97304	65.5	1.4	3.2	4.9	0.2	0.0	3.2	0.1	0.1	0.0	21.4
Partito Democratico	4424	52.2	1.7	0.5	6.8	1.0	0.0	0.1	3.6	0.0	0.0	34.1

Table 30: Domains annotation per political sub-communities – directed validated network. The description of the various columns can be found in the Table 1 and Section 2.3.3. The impact of urls coming from Social Networks (S) is much lower than that in Table 29, when only verified users are considered. The consideration written in the caption of Table 29, about the high values of N domains when considering only retweets, is valid here for M5S and PD only.

and '#vogliamovotare' (*we want to vote*).

Actually, the political competition still shines through the hashtags even for the other communities: it is the case, for instance, of Italia Viva. In the top 30 hashtags, we can find '#salvini', '#lega', but also '#papeete'[17], '#salvinisciacallo' (*Salvini jackal*) and '#salvinimmmerda' (*Salvini asshole*). Italia Viva use hashtags supporting the population: '#iorestoacasa', '#restoacasa' (*I am staying home*), '#restiamoacasa' (*let's stay home*). Criticisms towards the management of Lombardy health system during the pandemic can be deduced from the hashtag '#commissariatelalombardia' (*put Lombardy under receivership*) and '#fontana' (the Lega administrator of the Lombardy region).

Movimento 5 Stelle has the name of the main leader of the opposition, '#salvini', as first hashtag and it supports criticism to the Lombardy Administration with the hashtags '#fontanadimettiti' (*Fontana, resign!*)

[17]Matteo Salvini, while Minister of Internal Affairs, prepared the political crisis in 2019 from the *Papeete Beach* resort in Milano Marittima, Italy (Il Sole 24Ore, *Salvini, dal Papeete all'opposizione: l'agosto terribile del "capitano"*, 1st September 2019). His staying was advertised by a huge TV and social media covering and marked as a lack of respect towards the Republican institutions by his opponents. Instead, his supporters admired his closeness to the population.

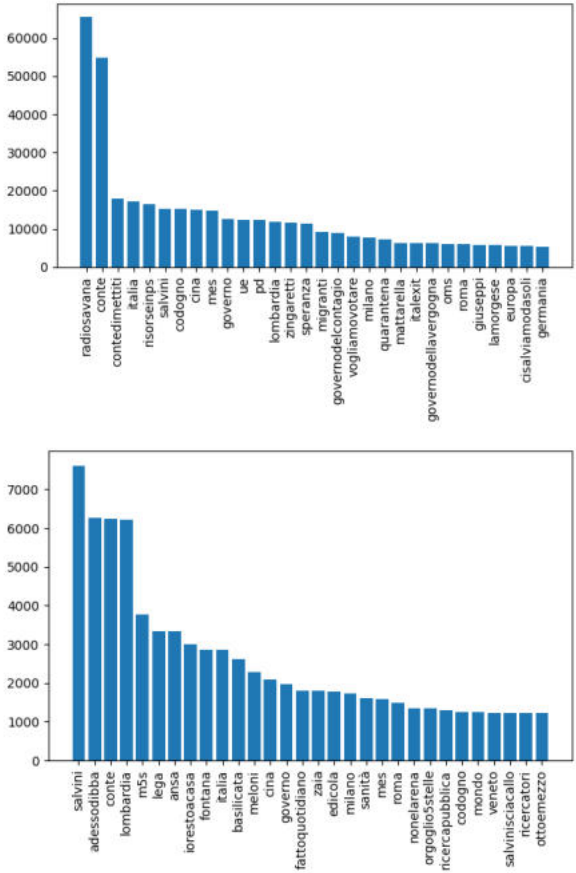


Figure 49: The 30 most diffused hashtags in the political sub-communities, directed validated network. Top panel: right and center-right wing discursive community. Bottom panel: 5 Stars Movement.

and '#gallera', the Health and Welfare Minister of the Lombardy Region, considered the main responsible for the bad management of the pandemic. Nevertheless, we can highlight even some hashtags encouraging the population during the lockdown, as the above mentioned '#iorestoacasa', '#restoacasa' and '#restiamoacasa'. It is worth mentioning that the

government measures, and the corresponding M5S campaigns, are accompanied by specific hashtags: '#curaitalia' is the name of one of the decree of the prime Minister to inject liquidity in the Italian economy, '#acquistaitaliano' (*buy Italian products!*), instead, advertises Italian products to support the national economy.

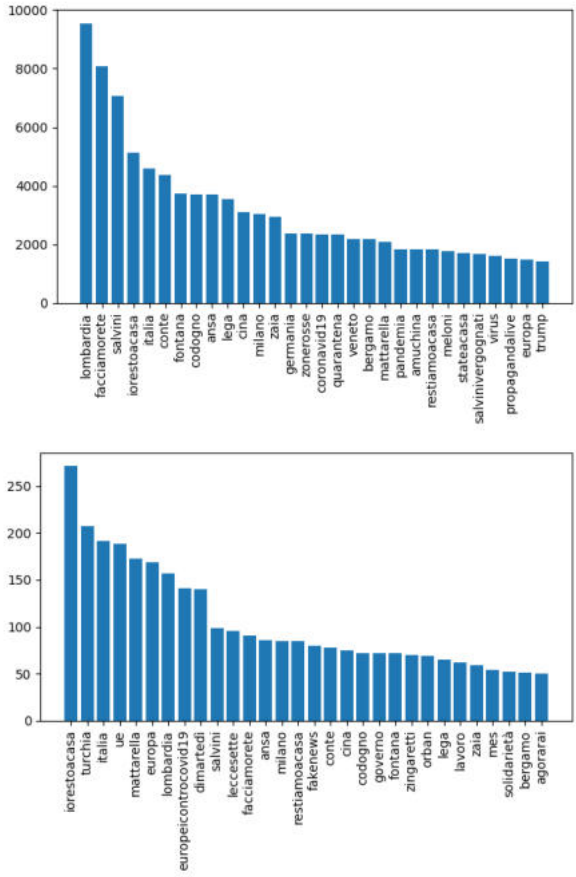


Figure 50: The 30 most diffused hashtags in the political subcommunities, directed validated network. Top panel: Italia Viva, Bottom panel: Democratic Party.

B.5 Label propagation comparison

In the main text, we solved the problem of assigning the orientation to all relevant users in the validated retweet network via a label propagation. The approach is similar, but different to the one proposed in [34], the differences being in the starting labels, in the label propagation algorithm and in the network used. In this section we will revise the method employed in the present article, as compared it to the one in [34] and evaluate the deviations from other approaches.

First step of our methodology is to extract the polarisation of verified users from the bipartite network, as described in Section 5.1 of the main text, in order to use it as seed labels in the label propagation.

In reference [34], a measure of the "adherence" of the *unverified* users towards the various communities of *verified* users was used in order to infer their orientation, following the approach in [9], in turn based on the polarisation index defined in [15]. This approach was extremely performing when practically all unverified users interact at least once with verified one, as in [9]. While still having good performances in a different dataset as the one studied in [34], we observed isolated deviations: it was the case of users with frequent interactions with other unverified accounts of the same (political) orientation, randomly retweeting a different discursive community verified user. In this case, focusing just on the interaction with verified accounts, those nodes were assigned a wrong group. The labels for the polarisation of the unverified users defined [34] were subsequently used as seed labels in the label propagation. Due to the possibility described above of wrongly assigning labels to unverified accounts, in the present paper, we consider only the tags of verified users, since they pass a strict validation procedure and are more stable.
There is another difference in the label propagation used here against the one in [34]: in the present paper we used the label propagation of [167], while the one in [34] was quite home-made. As in reference [167], the seed labels of [34] are fixed, i.e. are not allowed to change[18]. The main

[18] Actually, in [167] seed labels may be allowed to vary. Due to our application, we consider here the version in which they remain fixed, since the validation procedure is quite

difference is that, in case of a draw, among the labels of the first neighbours, in [167] a tie is removed randomly, while in the algorithm of [34] the label is not assigned and goes into a new run, with the newly assigned labels. Moreover, the updated of labels in [167] is asynchronous, while it is synchronous in [34]. We opted for the one in [167] for being actually a standard in the label propagation algorithms, being stable, more studied, and faster[19].
Finally, differently from the procedure in [34], we applied the label propagation not to the entire (undirected version of the) retweet network, but on the (undirected version of the) validated one. (The intent of choosing the *undirected version* is that in both case in which a generic account is significantly retweeting or being retweeted by another one, they do probably share some vision of the phenomena under analysis, thus we are not interested in the direction of the links, in this situation.) The rationale in using the validated network is to reduce the calculation time (due to the dimensions of the dataset), while obtaining an accurate result. While the previous differences from the procedure of [34] are dictated by conservativeness (the choice of the seed labels) or by the adherence to a standard (the choice of [167]), this last one may be debatable: why choosing the validated network should return "better" results than the ones calculated on the entire retweet network? We consider the case of a single day (in order to reduce the calculation time) and studied 6 different approaches:

1. a Louvain community detection [19] on the undirected version of the validated network of retweets;

2. a Louvain community detection on the undirected version of the unweighted retweet network;

3. a Louvain community detection on the undirected version of the weighted retweet network, in which the weights are the number of retweets from user to user;

strict.

[19] In the present paper we used the implementation of the label propagation in [167] that can be found in the python-igraph python module.

4. a label propagation a la Raghavan et al. [167] on the directed validated network of retweets;

5. a label propagation a la Raghavan et al. on the unweighted retweet network;

6. a label propagation a la Raghavan et al. on the weighted retweet network, the weights being the number of retweets from user to user.

Actually, due to the order dependence of Louvain [79], we run several times the Louvain algorithm after reshuffling the order of the nodes, taking the partition in communities that maximise the modularity. Similarly, the label propagation of [167] has a certain level of randomness: we run it several times and choose the most frequent label assignment for every node.

In order to compare the results obtained with the various approaches, we calculated the Variation of Information (VI, [131]). VI considers exactly the different in information contents captured by two different partition, as consider by the Shannon entropy. Results are reported in the matrix in Figure 51 for the 23th of February (results are similar for other days). Even when using the weighted retweet network as "exact" result, the partition found by the label propagation of our approach has a little loss of information, comparable with the one of using an unweighted approach. Indeed, the results found by the various community detection algorithms show little agreement with the label propagation ones. Nevertheless, we still prefer the label propagation procedure, since the validated projection on the layer of verified users is theoretically sound and has a non trivial interpretation.

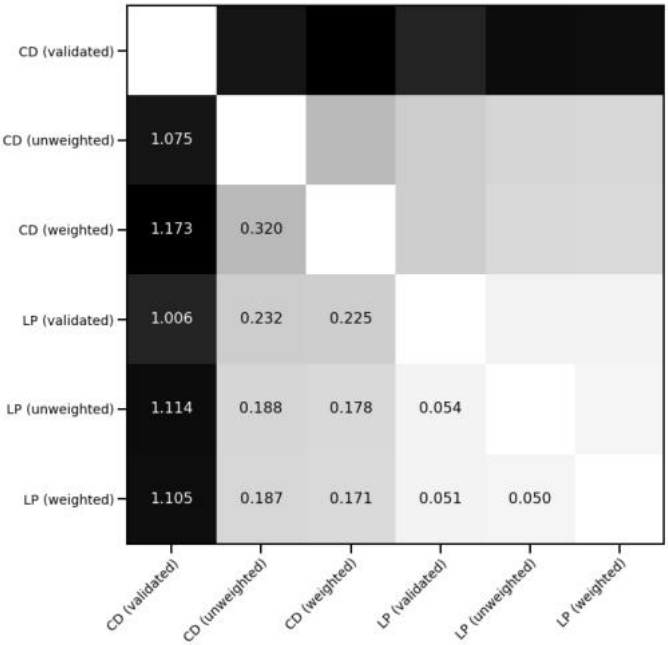


Figure 51: The Variation of Information table for the 23rd February 2020.(The date was chosen randomly.) The community detection algorithms do not agree so much even among themselves. Instead, the label propagation approaches results are quite similar. Due to this behaviour, we focus on the lightest one, i.e. the one calculated on the validated retweet network.

Appendix C

C

C.1 Relevant Italian online publishers

Table 31: The dataset analyzed in this study comprises Italian online publishers, as sourced from [148].

Publisher	Website
Avvenire	www.avvenire.it
Il Corriere Del Giorno	www.ilcorrieredelgiorno.it
Il Piccolo	www.ilpiccolo.gelocal.it
Corriere Della Sera	www.corriere.it
Il Fatto Quotidiano	www.ilfattoquotidiano.it
Il Post	www.ilpost.it
Domani	www.editorialedomani.it
Il Foglio	www.ilfoglio.it
Il Primato Nazionale	www.ilprimatonazionale.it
Il Gazzettino	www.ilgazzettino.it
Il Giornale	www.ilgiornale.it
Il Quotidiano Del Molise	www.quotidianomolise.com
Il Giornale Di Sicilia	www.gds.it
Il Giorno	www.ilgiorno.it
Il Resto Del Carlino	www.ilrestodelcarlino.it
Il Manifesto	www.ilmanifesto.it
Il Secolo D'Italia	www.secoloditalia.it
Il Mattino	www.ilmattino.it
Il Messaggero	www.ilmessaggero.it
Il Sole 24 Ore	www.ilsole24ore.com
Il Tirreno	www.iltirreno.gelocal.it
La Gazzetta Del Mezzogiorno	www.lagazzettadelmezzogiorno.it
La Nuova Ferrara	www.lanuovaferrara.gelocal.it
La Nazione	www.lanazione.it
La Nuova Padania	www.lanuovapadania.it
La Nuova Sardegna	www.lanuovasardegna.it
La Repubblica	www.repubblica.it
La Stampa	www.lastampa.it
La Verità	www.laverita.info
Libero	www.liberoquotidiano.it
Libertà	www.liberta.it
L'Unione Sarda	www.unionesarda.it
Open	www.open.online
Stopcensura	www.stopcensura.online

C.2 NewsGuard criteria

C.3 GDI criteria

Criteria name	Short name	Definition	points
Does not repeatedly publish false content	RepFalseCont	The site does not repeatedly produce stories that have been found - either by journalists at NewsGuard or elsewhere - to be clearly and significantly false, and which have not been quickly and prominently corrected.	22
Gathers and presents information responsibly	InfoResp	Content providers are generally fair and accurate in reporting and presenting information. They reference multiple sources, preferably those that present direct, firsthand information on a subject or event or from credible second hand news sources, and they do not egregiously distort or misrepresent information to make an argument or report on a subject.	18
Regularly corrects or clarifies errors	ErrCorr	The site makes clear how to report an error or complaint, has effective practices for publishing clarifications and corrections, and notes corrections in a transparent way.	12.5
Handles the difference between news and opinion responsibly	NewsOpDiff	Content providers who convey the impression that they report news or a mix of news and opinion distinguish opinion from news reporting, and when reporting news, do not egregiously cherry pick facts or stories to advance opinions. Content providers who advance a particular point of view disclose that point of view.	12.5
Avoids deceptive headlines	AvDecHeadlines	The site generally does not publish headlines that include false information, significantly sensationalize, or otherwise do not reflect what is actually in the story	10
Website discloses ownership and financing	DiscOwnFin	The site discloses its ownership and/or financing, as well as any notable ideological or political positions held by those with a significant financial interest in the site, in a user-friendly manner.	7.5
Clearly labels advertising	LabAds	The site makes clear which content is paid for and which is not.	7.5
Reveals who's in charge, including possible conflicts of interest	RevConfOfInt	Information about those in charge of the content is made accessible on the site	5
The site provides the names of content creators, along with either contact or biographical information	ContCreators	Information about those producing the content is made accessible on the site	5

Table 32: NewsGuard criteria specifications.

Criteria Name	Short Name	Sub-indicator	Definition
Headline accuracy	HeadAcc	-	Measures how well the headline reflects the article content. Indicates clickbait.
Byline information	ByInfo	-	Measures availability of author info, promoting accountability.
Lede present	LedePres	-	Presence of a fact-based lede. Indicates journalistic standards.
Common coverage	ComCov	-	Whether the event is covered by another credible local source.
Recent coverage	RecCov	-	Checks if the event happened within 30 days of article publication.
Negative targeting	NegTarg	-	Whether a group or individual is negatively targeted. Indicates hate or bias.
Article bias	ArtBias	-	Evaluates neutrality or bias of the article.
Sensational language	SensLang	-	Measures degree of sensationalism in language.
Visual presentation	VisPres	-	Measures sensationalism in imagery.
Attribution	Attr	-	Evaluates policies on attribution of facts, sources, and media.
Comment policies	CommPol	Policies	Site policies on moderating user-generated content.
		Moderation	Mechanisms to enforce those policies.
Editorial principles and practices	EdPrincPract	Editorial independence	Measures policies that ensure editorial independence.
		Adherence to narrative	Detects ideological alignment in editorial content.
		Content guidelines	Checks for factual accuracy and bias mitigation policies.
		News vs. analysis	Distinction between news and opinion content.
Ensuring accuracy	EnsAcc	Pre-pub. fact-checking	Existence of fact-checking before publication.
		Post-pub. corrections	Transparency and clarity of corrections process.
Funding	Fund	Diversified incentives	Number of funding sources. Avoids reliance on few entities.
		Accountability to readers	Whether reader support is a revenue source.
		Transparent funding	Clarity of financial backing and sponsors.
Ownership	Own	Owner-operator division	Distinct editorial and financial governance.
		Transparent ownership	Disclosure of ownership structure and affiliations.

Table 33: GDI criteria specifications.

Bibliography

[1] *Ad Fontes Media*. URL: https://adfontesmedia.com/ (visited on 10/23/2024).

[2] Lada A. Adamic and Natalie S. Glance. "The political blogosphere and the 2004 U.S. election: divided they blog". In: *3rd International Workshop on Link discovery, LinkKDD 2005, Chicago, Illinois, USA, August 21-25, 2005*. 2005, pp. 36–43.

[3] Ahmet Aker, Kevin Vincentius, and Kalina Bontcheva. "Credibility and Transparency of News Sources: Data Collection and Feature Analysis." In: *NewsIR@ SIGIR*. 2019, pp. 15–20.

[4] *Angular*. URL: https://angular.dev/ (visited on 10/23/2024).

[5] Oriol Artime et al. "Effectiveness of dismantling strategies on moderated vs. unmoderated online social platforms". In: *Sci. Rep.* 10.1 (Dec. 2020), p. 14392. ISSN: 20452322. DOI: 10.1038/s41598-020-71231-3. URL: https://doi.org/10.1038/s41598-020-71231-3.

[6] Ron Artstein and Massimo Poesio. "Inter-coder agreement for computational linguistics". In: *Computational linguistics* 34.4 (2008), pp. 555–596.

[7] John W. Ayers et al. "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum". In: *JAMA Internal Medicine* 183.6 (June 2023), pp. 589–596. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2023.1838. URL: https://doi.org/10.1001/jamainternmed.2023.1838.

[8] Parisa Bazmi, Masoud Asadpour, and Azadeh Shakery. "Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility". In: *Information Processing & Management* 60.1 (2023), p. 103146.

[9] Carolina Becatti et al. "Extracting significant signal of news consumption from social networks: the case of Twitter in Italian political elections". In: *Palgrave Communications* 5 (1 Dec. 2019), pp. 1–16. ISSN: 20551045. DOI: 10.1057/s41599-019-0300-3.

[10] L. Belcastro et al. "Analyzing voter behavior on social media during the 2020 US presidential election campaign". In: *Soc Netw Anal Min.* 12.1 (2022). DOI: 10.1007/s13278-022-00913-9.

[11] Daniele Bellutta and Kathleen M Carley. "Investigating coordinated account creation using burst detection and network analysis". In: *Journal of big Data* 10.1 (2023), pp. 1–17.

[12] Y Benjamini and Y Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J. R. Stat. Soc. B* 57.1 (1995), pp. 289–300. ISSN: 00359246.

[13] Nicolas Berlinski et al. "The Effects of Unsubstantiated Claims of Voter Fraud on Confidence in Elections". In: *Journal of Experimental Political Science* 10.1 (2023), pp. 34–49. DOI: 10.1017/XPS.2021.18.

[14] D. Beskow et al. "Introducing Bothunter: A tiered approach to detection and characterizing automated activity on Twitter". In: *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2018.

[15] Alessandro Bessi et al. "Users polarization on Facebook and Youtube". In: *PLoS One* 11.8 (2016). ISSN: 19326203. DOI: 10.1371/journal.pone.0159641.

[16] Md Momen Bhuiyan et al. "Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–26.

[17] John Bianchi et al. "Evaluating Trustworthiness of Online News Publishers via Article Classification". In: *arXiv preprint arXiv:2401.01781* (2024).

[18] *BiCM*. URL: https://github.com/mat701/BiCM (visited on 10/23/2024).

[19] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[20] Matyáš Boháček et al. "Fine-grained Czech News Article Dataset: An Interdisciplinary Approach to Trustworthiness Analysis". In: *arXiv preprint arXiv:2212.08550* (2022).

[21] Alexandre Bovet and Hernán A. Makse. "Influence of fake news in Twitter during the 2016 US presidential election". In: *Nature Communications* 10.7 (2019).

[22] S. Bowman and C Willis. *We Media: How Audiences are Shaping the Future of News and Information*. The Media Center at the American Press Institute. 2003.

[23] Samantha Bradshaw and Philip N. Howard. "How does junk news spread so quickly across social media? Algorithms, advertising and exposure in public life". In: *Oxford Internet Institute - White Paper* (2018).

[24] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[25] J. Bright et al. "Does Campaigning on Social Media Make a Difference? Evidence From Candidate Use of Twitter During the 2015 and 2017 U.K. Elections". In: *Communication Research* (2020). DOI: https://doi.org/10.1177/0093650219872394.

[26] Jonathan Bright. "Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism". In: *Journal of Computer-Mediated Communication* 23.1 (Jan. 2018), pp. 17–33. ISSN: 1083-6101. DOI: 10.1093/jcmc/zmx002. eprint: https://academic.oup.com/jcmc/article-pdf/23/1/17/23822774/zmx002.pdf. URL: https://doi.org/10.1093/jcmc/zmx002.

[27] Matteo Bruno, Renaud Lambiotte, and Fabio Saracco. "Brexit and bots: characterizing the behaviour of automated accounts on Twitter during the UK election". In: *EPJ Data Science 2022 11:1* 11 (1 Mar. 2022), pp. 1–24. ISSN: 2193-1127. DOI: 10.1140/EPJDS/S13688-022-00330-0. URL: https://epjdatascience.spr

ingeropen.com/articles/10.1140/epjds/s13688-022-00330-0.

[28] Ceren Budak. "What happened? the spread of fake news publisher content during the 2016 us presidential election". In: *The World Wide Web Conference*. 2019, pp. 139–150.

[29] Michael Butter and Peter Knight. *Routledge Handbook of Conspiracy Theories*. Routledge, 2021.

[30] G. Caldarelli et al. "The role of bot squads in the political propaganda on Twitter". In: *Communications Physics* 3 (2019), pp. 1–15.

[31] Guido Caldarelli. *Scale-Free Networks Complex Webs in Nature and Technology*. Oxford University Press, 2010, pp. 1–328.

[32] Guido Caldarelli et al. "Analysis of online misinformation during the peak of the COVID-19 pandemics in Italy". In: *CoRR* abs/2010.01913 (2020). arXiv: 2010.01913. URL: https://arxiv.org/abs/2010.01913.

[33] Guido Caldarelli et al. "Flow of online misinformation during the peak of the COVID-19 pandemic in Italy". In: *EPJ data science* 10.1 (2021), p. 34.

[34] Guido Caldarelli et al. "The role of bot squads in the political propaganda on Twitter". In: *Communications Physics* 3.1 (2020), pp. 1–15.

[35] Guido Caldarelli et al. "The role of bot squads in the political propaganda on Twitter". In: *Commun. Phys.* 3.1 (Dec. 2020), pp. 1–15. ISSN: 23993650. DOI: 10.1038/s42005-020-0340-4. arXiv: 1905.12687.

[36] Qiang Cao et al. "Uncovering Large Groups of Active Malicious Accounts in Online Social Networks". In: *ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2014, pp. 477–488.

[37] Dallas Card et al. "The Media Frames Corpus: Annotations of Frames Across Issues". In: *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, July 2015, pp. 438–444. DOI: 10.3115/v1/P15-2072.

[38] Tatiana Celadin et al. “Displaying news source trustworthiness ratings reduces sharing intentions for false news posts”. In: *Journal of Online Trust and Safety* 1.5 (2023).

[39] Alessandro Celestini et al. “Information disorders on Italian Facebook during COVID-19 infodemic”. In: *arXiv* (July 2020). arXiv: 2007.11302. URL: http://arxiv.org/abs/2007.11302.

[40] Andrea Ceron. “Internet, news, and political trust: The difference between social media and online media outlets”. In: *Journal of computer-mediated communication* 20.5 (2015), pp. 487–503.

[41] N. Chavoshi, H. Hamooni, and A. Mueen. “DeBot: Twitter Bot Detection via Warped Correlation”. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016, pp. 817–822. DOI: 10.1109/ICDM.2016.0096.

[42] Emily Chen, Kristina Lerman, and Emilio Ferrara. “Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set”. In: *JMIR Public Health Surveill* 6.2 (2020). DOI: 10.2196/19273.

[43] Giulio Cimini et al. “The statistical physics of real-world networks”. In: *Nature Reviews Physics* 1.1 (Jan. 2019), pp. 58–71. DOI: 10.1038/s42254-018-0002-6.

[44] Matteo Cinelli et al. “The COVID-19 social media infodemic”. In: *Sci. Rep.* 10.1 (Dec. 2020), p. 16598. ISSN: 20452322. DOI: 10.1038/s41598-020-73510-5. arXiv: 2003.05004. URL: www.nature.com/scientificreports.

[45] Bart De Clerck, Luis E. C. Rocha, and Filip Van Utterbeeck. “Maximum entropy networks for large scale social network node analysis”. In: *Applied Network Science 2022 7:1* 7 (1 Sept. 2022), pp. 1–22. ISSN: 2364-8228. DOI: 10.1007/S41109-022-00506-7. URL: https://appliednetsci.springeropen.com/articles/10.1007/s41109-022-00506-7.

[46] Bart De Clerck et al. “Maximum Entropy Networks Applied on Twitter Disinformation Datasets”. In: *Studies in Computational Intelligence* 1016 (2022), pp. 132–143. ISSN: 18609503. DOI: 10.1007/978-3-030-93413-2_12/COVER. URL: https://link.springer.com/chapter/10.1007/978-3-030-93413-2_12.

[47] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. URL: https://doi.org/10.1177/001316446002000104.

[48] M Conover, J Ratkiewicz, and M Francisco. "Political polarization on twitter." In: *Icwsm* (2011). ISSN: 15205126. DOI: 10.1021/ja202932e.

[49] Michael D. Conover et al. "Partisan asymmetries in online political activity". In: *EPJ Data Science* (2012). ISSN: 21931127. DOI: 10.1140/epjds6.

[50] Michael D. Conover et al. "Predicting the political alignment of twitter users". In: Jan. 2011. ISBN: 9780769545783. DOI: 10.1109/PASSAT/SocialCom.2011.34.

[51] Stefano Cresci. "A decade of social bot detection". In: *Commun. ACM* 63.10 (2020), pp. 72–83.

[52] Stefano Cresci et al. "Fame for sale: efficient detection of fake Twitter followers". In: *Decision Support Systems* 80 (2015), pp. 56–71.

[53] Stefano Cresci et al. "Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling". In: *IEEE Transactions on Dependable and Secure Computing* 15.4 (2018), pp. 561–576.

[54] Stefano Cresci et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race". In: *Proceedings of the 26th International Conference on World Wide Web Companion (WWW'17)*. ACM. 2017, pp. 963–972.

[55] Zeyu Dai, Himanshu Taneja, and Ruihong Huang. "Fine-grained Structure-based News Genre Categorization". In: *Events and Stories in the News*. Association for Computational Linguistics, Aug. 2018, pp. 61–67. URL: https://aclanthology.org/W18-4308.

[56] William T. Daniel and Lukas Obholzer. "Reaching out to the voter? Campaigning on Twitter during the 2019 European elections". In: *Research & Politics* 7.2 (2020), p. 2053168020917256. DOI: 10.1177/2053168020917256.

[57] Rocco De Nicola, Marinella Petrocchi, and Manuel Pratelli. "On the efficacy of old features for the detection of new bots". In: *Information Processing & Management* 58.6 (2021). copyrights © of the publisher 2025 Elsevier B.V., p. 102685.

[58] Michela Del Vicario et al. "Echo Chambers: Emotional Contagion and Group Polarization on Facebook". In: *Sci. Rep.* (2016). ISSN: 20452322. DOI: 10.1038/srep37825.

[59] Michela Del Vicario et al. "Mapping social dynamics on Facebook: The Brexit debate". In: *Soc. Networks* 50 (2017), pp. 6–16. ISSN: 03788733. DOI: 10.1016/j.socnet.2017.02.002.

[60] Michela Del Vicario et al. "The spreading of misinformation online". In: *Proceedings of the National Academy of Sciences* 113.3 (2016), pp. 554–559. DOI: 10.1073/pnas.1517441113.

[61] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[62] Reinhard Diestel. *Graph Theory (Graduate Texts in Mathematics)*. 2006, p. 415. ISBN: 3540261834. DOI: 10.1109/IEMBS.2010.5626521. arXiv: arXiv:1102.1087v6. URL: http://www.amazon.com/Graph-Theory-Graduate-Texts-Mathematics/dp/3540261834.

[63] *Disinformation as Adversarial Narrative Conflict*. https://www.disinformationindex.org/blog/2022-06-22-disinformation-as-adversarial-narrative-conflict/. 2022.

[64] *Docker*. URL: https://www.docker.com/ (visited on 10/23/2024).

[65] Yingtong Dou et al. "User preference-aware fake news detection". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2051–2055.

[66] Karen M Douglas et al. "Understanding conspiracy theories". In: *Political psychology* 40 (2019), pp. 3–35.

[67] Robert M. Entman. "Framing: Toward Clarification of a Fractured Paradigm". In: *Journal of Communication* 43.4 (1993), pp. 51–58. DOI: https://doi.org/10.1111/j.1460-2466.1993.tb01304.x.

[68] *European Digital Media Observatory (EDMO) Project*. https://edmo.eu/resources/training-programme/. 2024.

[69] *Facebook's Tips to Spot False News*. https://www.facebook.com/formedia/blog/third-party-fact-checking-tips-to-spot-false-news. 2024.

[70] Robert Faris et al. "Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election". In: *Berkman Klein Center Research Publication* 6 (2017).

[71] *FastAPI*. URL: https://fastapi.tiangolo.com (visited on 10/23/2024).

[72] Emilio Ferrara. "Disinformation and social bot operations in the run up to the 2017 French presidential election". In: *First Monday* 22.8 (2017). URL: https://firstmonday.org/ojs/index.php/fm/article/view/8005.

[73] Emilio Ferrara. "GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models". In: *Journal of Computational Social Science* 7.1 (2024), pp. 549–569.

[74] Emilio Ferrara et al. "Characterizing social media manipulation in the 2020 US presidential election". In: *First Monday* (2020).

[75] Emilio Ferrara et al. "The Rise of Social Bots". In: *Commun. ACM* 59.7 (June 2016), pp. 96–104. ISSN: 0001-0782. DOI: 10.1145/2818717. URL: http://doi.acm.org/10.1145/2818717.

[76] Seth Flaxman, Sharad Goel, and Justin M. Rao. "Filter Bubbles, Echo Chambers, and Online News Consumption". In: *Public Opinion Quarterly* 80.S1 (Mar. 2016), pp. 298–320. ISSN: 0033-362X. DOI: 10.1093/poq/nfw006.

[77] Forbes Business Council. *Self-Publishing Versus Traditional Publishing: Pros And Cons For Leaders To Consider*. https://www.forbes.com/sites/forbesbusinesscouncil/2022/08/15/self-publishing-versus-traditional-publishing-pros-and-cons-for-leaders-to-consider/. Accessed: 2023-12-22. 2022.

[78] Santo Fortunato. "Community detection in graphs". In: *Phys. Rep.* 486.3-5 (2010), pp. 75–174. ISSN: 03701573. DOI: 10.1016/j.physrep.2009.11.002.

[79] Santo Fortunato and Marc Barthélemy. "Resolution limit in community detection." In: *Pnas* 104.1 (Jan. 2007), pp. 36–41. ISSN: 0027-8424. DOI: 10.1073/pnas.0605965104. arXiv: 0607100v2 [arXiv:physics]. URL: http://www.ncbi.nlm.nih.gov/pubmed/17190818%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1765466.

[80] M. Freeze, M. Baumgartner, P. Bruno, et al. "Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect". In: *Polit Behav* (2021), pp. 1433–1465. DOI: https://doi.org/10.1007/s11109-020-09597-3.

[81] Thomas M.J. Fruchterman and Edward M. Reingold. "Graph drawing by force-directed placement". In: *Softw. Pract. Exp.* 21.11 (Nov. 1991), pp. 1129–1164. ISSN: 1097024X. DOI: 10.1002/spe.4380211102.

[82] Riccardo Gallotti et al. "Assessing the risks of 'infodemics' in response to COVID-19 epidemics". In: *Nat. Hum. Behav.* 4.12 (Dec. 2020), pp. 1285–1293. ISSN: 23973374. DOI: 10.1038/s41562-020-00994-6. arXiv: 2004.03997. URL: https://doi.org/10.1038/s41562-020-00994-6.

[83] C. Gangware and W. Nemr. *Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age*. Park Advisors, 2019.

[84] Diego Garlaschelli and Maria I. Loffredo. "Maximum likelihood: Extracting unbiased information from complex networks". In: *Physical Review E* 78 (1 2008), pp. 1–5. ISSN: 15393755. DOI: 10.1103/PhysRevE.78.015101.

[85] R. Kelly Garrett. "Echo chambers online?: Politically motivated selective exposure among Internet news users". In: *Journal of Computer-Mediated Communication* 14 (2 Jan. 2009), pp. 265–285. ISSN: 10836101. DOI: 10.1111/J.1083-6101.2009.01440.X.

[86] Christina Georgacopoulos and Grayce Mores. "How Fake News Affected the 2016 Presidential Election". In: *faculty.lsu.edu – White Paper* (2020).

[87] Maria Giatsoglou et al. "ND-Sync: Detecting Synchronized Fraud Activities". In: *PAKDD*. Springer, 2015.

[88] Yoav Goldberg and Graeme Hirst. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers, 2017. ISBN: 1627052984.

[89] Sandra González-Bailón, Javier Borge-Holthoefer, and Yamir Moreno. "Broadcasters and Hidden Influentials in Online Protest Diffusion". In: *American Behavioral Scientist* (2013). ISSN: 00027642. DOI: 10.1177/0002764213479371.

[90] Pietro Gravino et al. "The supply and demand of news during COVID-19 and assessment of questionable sources production". In: *Nature human behaviour* 6.8 (2022), pp. 1069–1078.

[91] Nir Grinberg et al. "Political science: Fake news on Twitter during the 2016 U.S. presidential election". In: *Science* 363.6425 (Jan. 2019), pp. 374–378. ISSN: 10959203.

[92] Stefano Guarino et al. "Information disorders during the COVID-19 infodemic: The case of Italian Facebook". In: *Online Soc. Networks Media* 22 (2021), p. 100124. URL: https://doi.org/10.1016/j.osnem.2021.100124.

[93] Stefano Guarino et al. "Information disorders during the COVID-19 infodemic: The case of Italian Facebook". In: *Online Social Networks and Media* 22 (Mar. 2021), p. 100124. ISSN: 2468-6964. DOI: 10.1016/J.OSNEM.2021.100124.

[94] Andrew M. Guess et al. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" In: *Science* 381.6656 (2023), pp. 398–404. DOI: 10.1126/science.abp9364.

[95] Andrew M. Guess et al. "Reshares on social media amplify political news but do not detectably affect beliefs or opinions". In: *Science* 381.6656 (2023), pp. 404–408. DOI: 10.1126/science.add8424.

[96] Jean-Loup Guillaume and Matthieu Latapy. "Bipartite structure of all complex networks". In: *Information Processing Letters* 90.5 (2004), pp. 215–221. ISSN: 0020-0190. DOI: https://doi.org/10.1016/j.ipl.2004.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0020019004000754.

[97] Jörg Haßler et al. *Campaigning on Facebook in the 2019 European Parliament Election*. Palgrave Macmillan Cham, 2021. DOI: 10.1007/978-3-030-73851-8.

[98] Paul Hawken. *The Next Economy*. New York, NY: Nenry Holt & co., 1983.

[99] *Hello GPT-4o*. https://openai.com/index/hello-gpt-4o/. 2024.

[100] Bahareh Heravi. "Storytelling Structures in Data Journalism: Introducing the Water Tower structure". In: *Computation+ Journalism 2022*. 2022.

[101] Hendrik Heuer and Elena L. Glassman. "Reliability Criteria for News Websites". In: *ACM Trans. Comput.-Hum. Interact.* 31.2 (Jan. 2024). ISSN: 1073-0516. DOI: 10.1145/3635147. URL: https://doi.org/10.1145/3635147.

[102] Yili Hong. "On computing the distribution function for the Poisson binomial distribution". In: *Comput. Stat. Data Anal.* 59.1 (2013), pp. 41–51. ISSN: 01679473.

[103] Tamanna Hossain et al. "COVIDLies: Detecting COVID-19 Misinformation on Social Media". In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics, Dec. 2020. DOI: 10.18653/v1/2020.nlpcovid19-2.11. URL: https://www.aclweb.org/anthology/2020.nlpcovid19-2.11.

[104] Philip N Howard et al. *Social media, news and political information during the US election: Was polarizing content concentrated in swing states?* Tech. rep. Data Memo 2017.8. Oxford, UK: Project on Computational Propaganda, 2017.

[105] Pik-Mai Hui et al. "BotSlayer: DIY Real-Time Influence Campaign Detection". In: *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM*. AAAI Press, 2020, pp. 980–982. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/7370.

[106] *Iffy Index*. URL: https://iffy.news/index/ (visited on 10/23/2024).

[107] "Italian Twitter semantic network during the Covid-19 epidemic". In: *EPJ Data Science 2021 10:1* 10 (1 Sept. 2021), pp. 1–27. ISSN: 2193-1127. DOI: 10.1140/EPJDS/S13688-021-00301-X. URL: https://link.springer.com/articles/10.1140/epjds/s13688-021-00301-x%20https://link.springer.com/article/10.1140/epjds/s13688-021-00301-x.

[108] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. "Human heuristics for AI-generated language are flawed". In: *Proceedings of the National Academy of Sciences* 120.11 (2023), e2208839120. DOI: 10.1073/pnas.2208839120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2208839120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2208839120.

[109] Kathleen Hall Jamieson and oseph N. Cappella. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press, 2008.

[110] E.T. Jaynes. *Information Theory and Statistical Mechanics*. 1957. DOI: 10.1103/PhysRev.106.620. arXiv: arXiv:1011.1669v3.

[111] Dariusz Jemielniak and Aleksandra Przegalinska. *Collaborative Society*. MIT Press, 2020.

[112] M. Jiang, P. Cui, and C. Faloutsos. "Suspicious Behavior Detection: Current Trends and Future Directions". In: *IEEE Intelligent Systems* 31.1 (2016), pp. 31–39.

[113] Antino Kim, Patricia L. Moravec, and Alan R. Dennis. "Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings". In: *Journal of Management Information Systems* 36.3 (2019), pp. 931–968. DOI: 10.1080/07421222.2019.1628921.

[114] Joshua Klayman. "Varieties of Confirmation Bias". In: *Psychology of Learning and Motivation* 32 (1995). Ed. by Jerome Busemeyer, Reid Hastie, and Douglas L. Medin, pp. 385–418. ISSN: 0079-7421. DOI: https://doi.org/10.1016/S0079-7421(08)60315-1.

[115] Jon M. Kleinberg. "Authoritative sources in a hyperlinked environment". In: *J. ACM* (1999). ISSN: 00045411. DOI: 10.1145/324133.324140.

[116] Anastasia Kozyreva et al. “Toolbox of individual-level interventions against online misinformation”. In: *Nature Human Behaviour* (2024), pp. 1–9.

[117] David M. J. Lazer et al. “The science of fake news”. In: *Science* 359.6380 (2018), pp. 1094–1096. DOI: 10.1126/science.aao2998. eprint: https://www.science.org/doi/pdf/10.1126/science.aao2998. URL: https://www.science.org/doi/abs/10.1126/science.aao2998.

[118] Patrick Lewis et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[119] Jeroen van Lidth de Jeude et al. “Reconstructing Mesoscale Network Structures”. In: *Complexity* 2019 (2019), 5120581:1–5120581:13.

[120] Hause Lin et al. “High level of correspondence across different news domain quality rating sets”. In: *PNAS Nexus* 2.9 (2023).

[121] Hause Lin et al. “High level of correspondence across different news domain quality rating sets”. In: *PNAS Nexus* 2.9 (Sept. 2023), pgad286. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgad286.

[122] Renan S Linhares et al. “Uncovering coordinated communities on twitter during the 2020 us election”. In: *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. 2022, pp. 80–87.

[123] Siyi Liu et al. “Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence”. In: *Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Nov. 2019, pp. 504–514. DOI: 10.18653/v1/K19-1047.

[124] Luca Luceri et al. “Evolution of bot and human behavior during elections”. In: *First Monday* (2019).

[125] H. B. Mann and D. R. Whitney. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60.

[126] Shawn Martin, W. M. Brown, and Brian N. Wylie. *Dr.L: Distributed Recursive (Graph) Layout*. [Computer Software] https://doi.org/10.11578/dc.20210416.20. Nov. 2007.

[127] Mattia Mattei et al. “Bow-tie structures of twitter discursive communities”. In: *Scientific Reports* 12.1 (2022), p. 12944.

[128] Nour El-Mawass, Paul Honeine, and Laurent Vercouter. “SimilCatch: Enhanced social spammers detection on Twitter using Markov Random Fields”. In: *Information Processing & Management* 57.6 (2020), p. 102317. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2020.102317. URL: http://www.sciencedirect.com/science/article/pii/S0306457320308128.

[129] Mary L McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia medica* 22.3 (2012), pp. 276–282.

[130] *MediaBias Fact Check*. URL: https://mediabiasfactcheck.com/ (visited on 10/23/2024).

[131] Marina Meila. “Comparing clusterings by the variation of information”. In: *Learn. theory Kernel Mach. 16th Annu. Conf. Learn. Theory 7th Kernel Work. COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003 Proc.* (2003), p. 173. ISSN: 03029743. DOI: 10.1007/978-3-540-45167-9_14.

[132] Meta. *Meta Journalism Project*. https://www.facebook.com/journalismproject/programs/third-party-fact-checking. [Online; accessed 08-April-2022]. 2022.

[133] Eni Mustafaraj and P. Takis Metaxas. “From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search”. In: *Web Science: Extending the Frontiers of Society On-Line*. 2010.

[134] Mirco Nacoti et al. “At the Epicenter of the Covid-19 Pandemic and Humanitarian Crises in Italy: Changing Perspectives on Preparation and Mitigation”. In: *Catal. non-issue content* (2020). DOI: 10.1056/CAT.20.0080.

[135] Preslav Nakov et al. “Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, 2022, pp. 495–520. ISBN: 978-3-031-13643-6.

[136] *NewsGuard*. URL: https://www.newsguardtech.com (visited on 10/23/2024).

[137] *NewsGuard's rating process criteria*. https://www.newsguardtech.com/ratings process-criteria/. 2024.

[138] *GDI product and FAQ*. https://www.disinformationindex.org/product/.

[139] *NewsGuard's AI Safety Suite*. https://www.newsguardtech.com/solutions/ai-safety-suite/#. 2024.

[140] Raymond S Nickerson. "Confirmation bias: A ubiquitous phenomenon in many guises". In: *Review of general psychology* 2.2 (1998), pp. 175–220.

[141] Lukas Obholzer and William T Daniel. "An online electoral connection? How electoral systems condition representatives' social media use". In: *European Union Politics* 17.3 (2016), pp. 387–407. DOI: 10.1177/1465116516630149.

[142] *OpenAI GPT-4o*. https://platform.openai.com/docs/models/gpt-4o. 2024.

[143] Juyong Park and Mark E J Newman. "Statistical mechanics of networks". In: *Phys. Rev. E* 70.6 (Dec. 2004), p. 66117. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.70.066117.

[144] K. Pearson. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *Philosophical Magazine Series* 5.50 (1900), pp. 157–175.

[145] Gordon Pennycook and David G. Rand. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning". In: *Cognition* 188 (2019). The Cognitive Science of Political Thought, pp. 39–50. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2018.06.011. URL: https://www.sciencedirect.com/science/article/pii/S001002771830163X.

[146] Gordon Pennycook et al. "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings". In: *Management Science* 66.11 (2020), pp. 4944–4957. DOI: 10.1287/mnsc.2019.3478.

[147] M Petrocchi and A Spognardi. *The Online News Market in Italy*. Online: https://www.disinformationindex.org/country-studies/2022-1-31-the-online-news-market-in-italy/. 2022.

[148] Marinella Petrocchi and Angelo Spognardi. *Disinformation Risk Assessment: The Online News Market in Italy*. https://www.disinformationindex.org/country-studies/2022-1-31-the-online-news-market-in-italy/xiii. accessed on 04/04/2024. 2022.

[149] Francesco Pierri et al. *VaccinItaly: monitoring Italian conversations around vaccines on Twitter*. 2021. arXiv: 2101.03757 [cs.SI].

[150] Jakub Piskorski et al. "SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup". In: *Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, July 2023, pp. 2343–2361. DOI: 10.18653/v1/2023.semeval-1.317.

[151] Stiene Praet, David Martens, and Peter Van Aelst. "Patterns of democracy? Social network analysis of parliamentary Twitter networks in 12 countries". In: *Online Social Networks and Media* 24 (2021). DOI: 10.1016/j.osnem.2021.100154.

[152] Manuel Pratelli and Marinella Petrocchi. "A Structured Analysis of Journalistic Evaluations for News Source Reliability". In: *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media, ICWSM 2022 Workshops, Atlanta, Georgia, USA [hybrid], June 6, 2022*. Ed. by Pedro O. S. Vaz de Melo, Wei Jeng, and Cody Buntain. 2022. DOI: 10.36190/2022.51. URL: https://doi.org/10.36190/2022.51.

[153] Manuel Pratelli and Marinella Petrocchi. "A Structured Analysis of Journalistic Evaluations for News Source Reliability". In: *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*. 2022. URL: https://doi.org/10.36190/2022.51.

[154] Manuel Pratelli and Marinella Petrocchi. "Evaluating the Simulation of Human Personality-Driven Susceptibility to Misinformation with LLMs". In: *arXiv preprint arXiv:2506.23610* (2025).

[155] Manuel Pratelli, Fabio Saracco, and Marinella Petrocchi. "Entropy-based detection of Twitter echo chambers". In: *PNAS Nexus* 3.5 (Apr. 2024), pgae177. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgae177. eprint: https://academic.oup.com/pnasnexus/article-pdf/3/5/pgae177/58004712/pgae177.pdf. URL: https://doi.org/10.1093/pnasnexus/pgae177.

[156] Manuel Pratelli, Fabio Saracco, and Marinella Petrocchi. "TROPIC – Trustworthiness Rating of Online Publishers Through Online Interactions Calculation". In: *Advances in Information Retrieval*. Ed. by Claudia Hauff et al. Cham: Springer Nature Switzerland, 2025, pp. 407–412. ISBN: 978-3-031-88717-8.

[157] Manuel Pratelli, Fabio Saracco, and Marinella Petrocchi. "Unveiling News Publishers Trustworthiness Through Social Interactions". In: *Proceedings of the 16th ACM Web Science Conference*. WEBSCI '24. Stuttgart, Germany: Association for Computing Machinery, 2024, pp. 139–148. ISBN: 9798400703348. DOI: 10.1145/3614419.3644015. URL: https://doi.org/10.1145/3614419.3644015.

[158] Manuel Pratelli et al. "Evaluation of Reliability Criteria for News Publishers with Large Language Models". In: *Proceedings of the 17th ACM Web Science Conference 2025*. Websci '25. Association for Computing Machinery, 2025, pp. 179–188. ISBN: 9798400714832. DOI: 10.1145/3717867.3717924. URL: https://doi.org/10.1145/3717867.3717924.

[159] Manuel Pratelli et al. "Online disinformation in the 2020 U.S. election: swing vs. safe states". In: *EPJ Data Science* 13.1 (2024), p. 25. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-024-00461-6. URL: https://doi.org/10.1140/epjds/s13688-024-00461-6.

[160] Manuel Pratelli et al. "Swinging in the States: Does disinformation on Twitter mirror the US presidential election system?" In: *arXiv preprint arXiv:2303.12474* (2023).

[161] Toby Prike, Lucy H Butler, and Ullrich KH Ecker. "Source-credibility information and social norms improve truth discernment and reduce engagement with misinformation online". In: *Scientific Reports* 14.1 (2024), p. 6900.

[162] Piotr Przybyla. "Capturing the Style of Fake News". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 490–497. URL: https://doi.org/10.1609/aaai.v34i01.5386.

[163] Lotte Pummerer et al. “Societal effects of COVID-19 conspiracy theories”. In: *Social Psychological and Personality Science* (2021).

[164] Walter Quattrociocchi, Guido Caldarelli, and Antonio Scala. “Opinion dynamics on interacting networks: Media competition and social influence”. In: *Sci. Rep.* 4 (2014). ISSN: 20452322. DOI: 10.1038/srep04938.

[165] Tommaso Radicioni et al. “Analysing Twitter semantic networks: the case of 2018 Italian elections”. In: *Scientific Reports 2021 11:1* 11 (1 June 2021), pp. 1–22. ISSN: 2045-2322. DOI: 10.1038/s41598-021-92337-2. URL: https://www.nature.com/articles/s41598-021-92337-2.

[166] Tommaso Radicioni et al. “Networked partisanship and framing: A socio-semantic network analysis of the Italian debate on migration”. In: *PLOS ONE* 16 (8 Mar. 2021), e0256705. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0256705. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0256705.

[167] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. “Near linear time algorithm to detect community structures in large-scale networks”. In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* (2007). ISSN: 15393755. DOI: 10.1103/PhysRevE.76.036106.

[168] Kevin Roitero et al. “Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19”. In: *CoRR* abs/2107.11755 (2021).

[169] Martin Rosvall and Carl T Bergstrom. “Maps of random walks on complex networks reveal community structure”. In: *Proceedings of the national academy of sciences* 105.4 (2008), pp. 1118–1123.

[170] Alessandro Rovetta and Akshaya Srikanth Bhagavathula. “COVID-19-related web search behaviors and infodemic attitudes in Italy: Infodemiological study”. In: *J. Med. Internet Res.* (2020). ISSN: 14388871. DOI: 10.2196/19374.

[171] Fabio Saracco et al. “Detecting early signs of the 2007–2008 crisis in the world trade”. In: *Sci. Rep.* 6 (July 2016), p. 30286. ISSN: 2045-2322. DOI: 10.1038/srep30286. arXiv: 1508.03533. URL: http://www.nature.com/articles/srep30286%20http://arxiv.org/abs/1508.03533%20http://dx.doi.org/10.1038/srep30286.

[172] Fabio Saracco et al. “Inferring monopartite projections of bipartite networks: An entropy-based approach”. In: *New J. Phys.* 19.5 (July 2017), p. 16. ISSN: 13672630. DOI: 10.1088/1367-2630/aa6b38. arXiv: 1607.02481. URL: http://arxiv.org/abs/1607.02481.

[173] Fabio Saracco et al. “Randomizing bipartite networks: the case of the World Trade Web”. In: *Scientific Reports* 5 (1 Sept. 2015), p. 10595. ISSN: 2045-2322. DOI: 10.1038/srep10595. URL: http://www.nature.com/articles/srep10595.

[174] Mohsen Sayyadiharikandeh et al. “Detection of Novel Social Bots by Ensembles of Specialized Classifiers”. In: *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management*. 2020, pp. 2725–2732.

[175] M. Scharkow. “Thematic content analysis using supervised machine learning: An empirical evaluation using German online news”. In: *Quality and Quantity* 47 (2013), pp. 761–773. DOI: 10.1007/s11135-011-9545-7.

[176] Ross J. Schuchard and Andrew T. Crooks. “Insights into elections: An ensemble bot detection coverage framework applied to the 2018 U.S. midterm elections”. In: *PLOS ONE* 16.1 (Jan. 2021), pp. 1–19. DOI: 10.1371/journal.pone.0244309. URL: https://doi.org/10.1371/journal.pone.0244309.

[177] *Scikit-learn cohen's kappa score implementation*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html. 2024.

[178] *Selenium*. https://www.selenium.dev/. 2024.

[179] Chengcheng Shao et al. “Anatomy of an online misinformation network”. In: *Plos one* 13.4 (2018), e0196087.

[180] Chengcheng Shao et al. “The spread of low-credibility content by social bots”. In: *Nature communications* 9.1 (2018), pp. 1–9.

[181] Karishma Sharma, Emilio Ferrara, and Yan Liu. *Identifying Coordinated Accounts in Disinformation Campaigns*. 2020. arXiv: 2008.11308 [cs.SI].

[182] Anu Shrestha et al. “Joint Credibility Estimation of News, User, and Publisher via Role-Relational Graph Convolutional Networks”. In: *ACM Transactions on the Web* 18.1 (2023), pp. 1–24.

[183] Kai Shu, Suhang Wang, and Huan Liu. "Beyond news contents: The role of social context for fake news detection". In: *Proceedings of the twelfth ACM international conference on web search and data mining*. 2019, pp. 312–320.

[184] N.V. Smirnov. "Estimate of deviation between empirical distribution functions in two independent samples". In: *Bull. Moscow Univ.* 2.2 (1939), pp. 3–16.

[185] Sonic Research Group at Nortwestern Univ. *Episode 5: The Bits and Bots of the Web with Fil Menczer*. online: https://tinyurl.com/y49gmchk. 2020.

[186] Michael Soprano et al. "The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale". In: *Inf. Process. Manag.* 58.6 (2021), p. 102710.

[187] Tiziano Squartini and Diego Garlaschelli. "Analytical maximum-likelihood method to detect patterns in real networks". In: *New Journal of Physics* 13 (2011), p. 083001. ISSN: 13672630. DOI: 10.1088/1367-2630/13/8/083001.

[188] Tiziano Squartini and Diego Garlaschelli. *Maximum-entropy networks. Pattern detection, network reconstruction and graph combinatorics*. Springer International Publishing, 2017, p. 116. ISBN: 9783319694368.

[189] Ivan Srba et al. "A Survey on Automatic Credibility Assessment Using Textual Credibility Signals in the Era of Large Language Models". In: *ACM Trans. Intell. Syst. Technol.* (Sept. 2025). ISSN: 2157-6904. DOI: 10.1145/3770077. URL: https://doi.org/10.1145/3770077.

[190] Galen Stockinh et al. *The Role of Alternative Social Media in the News and Information Environment*. Pew Research Center. 2022. URL: https://www.pewresearch.org/journalism/2022/10/06/the-role-of-alternative-social-media-in-the-news-and-information-environment/.

[191] Xing Su et al. "Hy-DeFake: Hypergraph Neural Networks for Detecting Fake News in Online Social Networks". In: *arXiv preprint arXiv:2309.02692* (2023).

[192] *The Global Disinformation Index*. URL: https://www.disinformationindex.org/ (visited on 10/23/2024).

[193] *The Trust Project - trust indicators*. https://thetrustproject.org/trust-indicators/. 2024.

[194] Yannis Theocharis et al. "A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates". In: *Journal of Communication* 66.6 (Oct. 2016), pp. 1007–1031. ISSN: 0021-9916. DOI: 10.1111/jcom.12259.

[195] Jeff Tollefson. "How Trump turned conspiracy theory research upside down". In: *Nature* 590 (2021), pp. 192–193.

[196] Twitter. *Empowering people on Twitter to create a better-informed world*. https://twitter.github.io/birdwatch/. [Online; accessed 08-April-2022]. 2022.

[197] Aleksandra Urman. "Context matters: political polarization on Twitter from a comparative perspective". In: *Media, Culture & Society* 42.6 (2020), pp. 857–879. DOI: 10.1177/0163443719876541. eprint: https://doi.org/10.1177/0163443719876541. URL: https://doi.org/10.1177/0163443719876541.

[198] Patti M Valkenburg and Jochen Peter. "Comm Research—Views from Europe— Five Challenges for the Future of Media-Effects Research". In: *International Journal of Communication* 7 (2013), p. 19.

[199] Livia Van Vliet, Petter Törnberg, and Justus Uitermark. "Political Systems and Political Networks: The Structure of Parliamentarians' Retweet Networks in 19 Countries". In: *International Journal of Communication* 15 (2021), pp. 2156–2176.

[200] Onur Varol et al. "Online Human-Bot Interactions: Detection, Estimation, and Characterization". In: *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*. 2017, pp. 280–289. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587.

[201] Onur Varol et al. "Online human-bot interactions: Detection, estimation, and characterization". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 11. 1. 2017.

[202] Bimal Viswanath et al. "Strength in Numbers: Robust Tamper Detection in Crowd Computations". In: *Proceedings of the 2015 ACM on Conference on Online Social Networks*. ACM. 2015, pp. 113–124.

[203] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (2018), pp. 1146–1151.

[204] Iain S. Weaver et al. “Communities of online news exposure during the UK General Election 2015”. In: *Online Social Networks and Media* 10-11 (2019), pp. 18–30. ISSN: 2468-6964. DOI: https://doi.org/10.1016/j.osnem.2019.05.001. URL: https://www.sciencedirect.com/science/article/pii/S2468696418301290.

[205] *wget*. https://www.gnu.org/software/wget/. 2024.

[206] *Writing differences feature, news and investigative articles*. https://gspieler.medium.com/whats-the-difference-writing-hard-news-feature-and-investigative-journalism-40811e0092c. 2024.

[207] Kai-Cheng Yang, Pik-Mai Hui, and Filippo Menczer. “How Twitter data sampling biases US voter behavior characterizations”. In: *PeerJ Computer Science* 8 (2022), e1025.

[208] Kai-Cheng Yang and Filippo Menczer. “Large language models can rate news outlet credibility”. In: *CoRR* abs/2304.00228 (2023). DOI: 10.48550/arXiv.2304.00228. URL: https://doi.org/10.48550/arXiv.2304.00228.

[209] Kai-Cheng Yang et al. “Arming the public with AI to counter social bots”. In: *CoRR* abs/1901.00912 (2019). URL: http://arxiv.org/abs/1901.00912.

[210] Kai-Cheng Yang et al. “Scalable and Generalizable Social Bot Detection through Data Selection”. In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 1096–1103.

[211] Kai-Cheng Yang et al. “Scalable and generalizable social bot detection through data selection”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 01. 2020, pp. 1096–1103.

[212] Kai-Cheng Yang et al. *The COVID-19 Infodemic: Twitter versus Facebook*. 2020. arXiv: 2012.09353 [cs.SI].

[213] Sarita Yardi et al. “Detecting Spam in a Twitter Network”. In: *First Monday* 15.1 (2010).

[214] Jinyi Ye et al. “Susceptibility to Unreliable Information Sources: Swift Adoption with Minimal Exposure”. In: *Proceedings of the ACM on Web Conference 2024*. WWW ’24. , Singapore, Singapore, Association for Computing Machinery, 2024, pp. 4674–4685. ISBN: 9798400701719. DOI: 10.1145/3589334.3648154. URL: https://doi.org/10.1145/3589334.3648154.

[215] Junjie Ye et al. *A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models*. 2023. arXiv: 2303.10420 [cs.CL].

[216] Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. "Automated fact-checking: A survey". In: *Lang. Linguistics Compass* 15.10 (2021). URL: https://doi.org/10.1111/lnc3.12438.

[217] Amy X Zhang et al. "A structured response to misinformation: Defining and annotating credibility indicators in news articles". In: *Companion Proceedings of the The Web Conference 2018*. 2018, pp. 603–612.

[218] Xinyi Zhou et al. "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Oct. 2020). DOI: 10.1145/3340531.3412880. URL: http://dx.doi.org/10.1145/3340531.3412880.

[219] Fabiana Zollo et al. "Debunking in a world of tribes". In: *PLoS One* 12.7 (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0181821.

[220] Fabiana Zollo et al. "Emotional dynamics in the age of misinformation". In: *PLoS One* 10.9 (2015). ISSN: 19326203. DOI: 10.1371/journal.pone.0138740.

[221] Arkaitz Zubiaga and Aiqi Jiang. "Early Detection of Social Media Hoaxes at Scale". In: *ACM Trans. Web* 14.4 (2020), 18:1–18:23.