A Comparative Study of Algorithm-Mediated Behavior and Toxic Discourse on Social Media Platforms.

Department of Computer Science

Doctor of Philosophy in Computer Science – XXXVI Cycle

Candidate

Gabriele Etta
ID number 1943331

Thesis Advisor

Prof. Walter Quattrociocchi

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science

2023

A Comparative Study of Algorithm-Mediated Behavior and Toxic Discourse on Social Media Platforms.
Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: etta@di.uniroma1.it

Contents

Vita and Publications **1**

1 Introduction **7**

1.1 Echo Chambers 9

1.2 Toxicity Dynamics 9

1.3 Misinformation and Society 9

1.4 Content Moderation 10

1.5 Advancements 11

2 Preliminaries and Definitions **15**

2.1 Data Collection 15

2.1.1 Facebook and Instagram 15

2.1.2 Reddit 16

2.1.3 Twitter 16

2.1.4 Voat 16

2.1.5 YouTube 16

2.1.6 Gab 16

2.1.7 Usenet 17

2.2 Data Transformation 17

2.2.1 URL Expansion 17

2.2.2 News Outlet Classification 17

2.3 Measuring user homophily in social network 18

2.3.1 User Leaning 18

2.3.2 Homophily in the Interaction Network 19

2.4 Networks 19

2.4.1 Definition 19

2.4.2 Bipartite Graph 19
2.4.3 Tree Graph 20
2.5 Tree Structural Metrics 20
2.6 Statistical Tools 22
2.6.1 Lifetime Estimation 22

3 Conspiracy Theories and social media platforms 25
3.1 Echo chambers and conspiracy in social media 26
3.1.1 Facebook 27
3.1.2 Twitter 28
3.1.3 YouTube 29
3.1.4 Reddit 30
3.1.5 Other social media platforms 30
3.2 Regulation and the risk of echo platforms 31
3.3 Conclusions and future works 32

4 Comparing the impact of social media regulations on news consumption 33
4.1 The COVID-19 infodemic on social media 34
4.2 Preliminaries and Definitions 36
4.2.1 Data Collection 36
4.2.2 Questionable and Reliable Sources 37
4.2.3 Comparison of power law distributions 38
4.3 Result and Discussion 39
4.3.1 Consumption Patterns 39
4.3.2 Characterizing Commenting Behavior for Questionable and Reliable posts 41
4.3.3 Quantifying Polarization 42
4.4 Conclusions 43

5 Entropy and complexity unveil the landscape of memes evolution 51
5.1 The advent of visual memes in the digital ecosystem 52
5.2 Methods 55
5.2.1 Data Breakdown 55

5.2.2 Clustering 55
5.2.3 Entropy and complexity 58
5.3 Results and Discussion 60
5.4 Conclusion 65

6 Characterizing Engagement Dynamics across Topics on Facebook 67
6.1 The attention economy in social media era 68
6.2 Materials and Methods 69
6.2.1 Overview of the data collection process 69
6.2.2 Metrics 72
6.3 Results and Discussion 75
6.3.1 Quantifying topic engagement evolution 75
6.3.2 Evaluating the relationship between topic engagement and controversy 76
6.3.3 Assessing the differences of engagement behaviors across topic categories 79
6.4 Conclusions 80

7 A Topology-Based Approach for Predicting Toxic Outcomes on Twitter and YouTube 83
7.1 Background 84
7.1.1 Defining toxicity on social media 84
7.2 The rising of toxicity in social media 85
7.2.1 Conversation cascades and toxicity dynamics 85
7.2.2 Machine Learning for toxicity identification 86
7.3 Preliminaries and definitions 86
7.3.1 Data Collection 86
7.3.2 Toxicity Labelling 88
7.3.3 Conversation Cascade Reconstruction 89
7.3.4 Cascade metrics 90
7.3.5 Permutation Test 91
7.3.6 Toxicity comment prediction in a conversation 91
7.4 Results 94

7.4.1 Toxicity Evolution 94

7.4.2 Structural analysis 96

7.4.3 Predicting the following toxic comment in a conversation 97

7.5 Conclusion 99

8 Toxicity in online conversations 105

8.1 Background 106

8.2 Materials and Methods 108

8.2.1 Data Collection 108

8.2.2 Logarithmic binning and conversation sizes 115

8.2.3 Toxicity detection and validation of Perspective API . 115

8.2.4 Burst analysis 119

8.2.5 Toxicity Detection on Usenet 120

8.2.6 Participation and toxicity in short conversations 121

8.2.7 Data availability 121

8.3 Results 122

8.3.1 Conversation size and toxicity 123

8.3.2 Conversation evolution and toxicity 126

8.3.3 Controversy and toxicity 128

8.4 Discussion 133

9 Final remarks and future direction 137

Bibliography 139

Appendix 177

Vita

2016

B.Sc. in Computer Science

104/110

University of Parma, Parma, Italy

2020

M.Sc. in Data Science

110/110

University of Padova, Padova, Italy

Publications

[1] C.M. Valensise, A. Serra, A. Galeazzi, **G. Etta**, M. Cinelli, W. Quattrociocchi, *Entropy and complexity unveil the landscape of memes evolution* Sci Rep 11, 20022 (2021).

[2] **G. Etta,** M. Cinelli, A. Galeazzi, C. M. Valensise, W. Quattrociocchi and M. Conti, *Comparing the Impact of Social Media Regulations on News Consumption*, in IEEE Transactions on Computational Social Systems, vol. 10, no. 3, pp. 1252-1262, June 2023, doi: 10.1109/TCSS.2022.3171391.

[3] **G. Etta**, A. Galeazzi, J.R. Hutchings , C.S. James Smith , M. Conti, W. Quattrociocchi, G.V. Dalla Riva (2022) *COVID-19 infodemic on Facebook and containment measures in Italy, United Kingdom and New Zealand.* PLOS ONE 17(5): e0267022. https://doi.org/10.1371/journal.pone.0267022

[4] M. Cinelli, **G. Etta**, M. Avalle, A. Quattrociocchi, N. Di Marco, C. Valensise, A. Galeazzi, W. Quattrociocchi, *Conspiracy theories and social media platforms*, Current Opinion in Psychology, Volume 47, 2022, 101407, ISSN 2352-250X, https://doi.org/10.1016/j.copsyc.2022.101407.

[5] A. Quattrociocchi, **G. Etta**, M. Avalle, M. Cinelli, W. Quattrociocchi, *Reliability of News and Toxicity in Twitter Conversations*, International Conference on Social Informatics, 2022, 245-256

[6] S. Alipour, N. Di Marco, M. Avalle, M. Cinelli, **G. Etta**,, W. Quattrociocchi, *The Drivers of Global News Spreading Patterns*, *To appear on Scientific Reports*, 2023

[7] **G. Etta**, N. Di Marco, M. Avalle, M. Cinelli, W. Quattrociocchi, *A Topology-Based Approach for Predicting Toxic Outcomes on Twitter and YouTube*, *Under Review on IEEE Big Data 2023*, 2023

Abstracts & Posters

[1] **G. Etta**, N. Di Marco, M. Avalle, M. Cinelli, W. Quattrociocchi, *Quantifying Topological Differences in Online Conversations*, IC2S2 2023

Summer Schools and Awards

[1] "Una tesi per la sicurezza nazionale" award, Presidenza del consiglio dei Ministri, Italy, 2021

[2] ComQuant Summer School, Koc University, Turkey, 2022

Abstract

In the evolving digital communication landscape, social media platforms are pivotal in shaping public opinion and societal narratives. These platforms, characterized by their democratized access to information and the facility for real-time engagement, have the potential to enrich public discourse significantly. However, they may also foster environments that restrict users' exposure to diversified content, contributing to the formation of echo chambers (i.e., groups of like-minded individuals where homogenous ideologies are reinforced), exacerbating user polarization, and promoting antisocial behaviors with severe implications for the broader democratic process. This dissertation explores the complex interplay between the dissemination of misinformation and its impact on online discourse, with a methodological innovation at its core. Recognizing that human behavior, as reflected in social media data, is inherently mediated by platform-specific algorithms, this research proposes a novel analytical framework. Adopting a comparative approach across various platforms seeks to unmask the underlying patterns of information propagation and user engagement, transcending the limitations imposed by algorithmic mediation. We first conduct a detailed examination of misinformation and conspiracy theory diffusion across social media landscapes, offering a quantitative assessment of the effectiveness of existing moderation policies. Then, we explore new expressive forms emerging within online dialogues, mainly through the lens of meme diffusion, to understand the relationship between viral content and the generation of controversial user reactions. Expanding on these insights, we conduct a comprehensive cross-platform analysis of toxic conversational dynamics, assessing how polarization contributes to the proliferation of hate speech online. Lastly, we discuss potential advancements in moderation tool designs, aiming to mitigate such digital hostility proactively. The empirical findings underscore the effectiveness of moderation tools in counteracting the spread of conspiracy theories. They reveal a tendency for viral topics to spark controversial and heated discussions, with toxicity levels intensifying progressively. Importantly, this research demonstrates the predictive ability of classifiers, trained on different stages of conversations, to identify the presence of toxic comments with high accuracy, even within a constrained feature set.

This dissertation underscores the necessity of comparative, cross-platform analysis to understand the digital communication ecosystem. It explores how platform-induced behaviors may influence public discourse, providing insights that bridge human behavioral studies and digital platform policies.

Chapter 1

Introduction

The growing adoption of digital technologies has profoundly transformed how people experience the world, accelerating the development of science, industries, and society. The newborn digital era, characterized by the increasing adoption of computers and storage systems, replaced the previous analogical solutions with unprecedented speed and pervasivity. The disruptive changes produced by these transitions led to data production with increasing speed, volumes, and heterogeneity. This change, combined with the interconnectivity due to the Internet, brought the digital era to a newer stage called the Big Data Era [1]. On one side, this era unveiled the necessity to find new ways of extracting information from datasets too large to be processed with the current methods. On the other, it provided numerous opportunities for creating platforms to connect users and produce content of various types. Such platforms, going under the name of Social Media Platforms, irreversibly shaped how users inform themselves, frame the world, and interact with other peers. The disintermediated information paradigm that these platforms introduced, where users can access pieces of content coming from a multitude of sources, represents a faster, more heterogeneous way to expand the knowledge of users in a real-time fashion with the events occurring around the world. At the same time, the constant exposure of users to an overabundant flow of information and news, whose trustworthiness may not always be guaranteed, contributed to the emergence of psychological and social dynamics of concern to maintain the well-being of individuals and society in general [2]. Indeed, social media users tend to seek and consume information that aligns with their pre-existing beliefs and avoid discordant information due to selective exposure from themselves to the content they consume. This phenomenon has led to the formation of so-called *echo chambers* [3], where individuals gather around shared narratives, isolating themselves from differing opinions. Group polarization theories [4] suggest that echo chambers can reinforce pre-existing beliefs and push the group toward increasingly extreme positions. When the opposing narratives refer to conspiracy theories, however, the effect of polarization and echo chambers generally becomes of particular concern, as it can have potentially harmful consequences for individuals and societies. Like misinformation, conspiracy theories appear on many social media platforms, involving a wide range of users, from believers to debunkers and passive observers. While non-conspiracy information from news outlets outpaces conspiracy-related content, the significance of user engagement around conspiracy theories should not be underestimated. A relevant aspect related to the spread of conspiracy theories online is that individuals who embrace them are often highly engaged and more likely to share false information. This dynamic has been particularly evident in recent years, when users were subjected to an unprecedented volume of content circulating online, often being unable to discern what information can be trusted. Such a scenario is referred to as *infodemic* and, combined with the proliferation of conspiracy theories, had dangerous effects on citizens' health and the stability of the democracies where these actions occurred.

In addition to the previous threats, the spread of misinformation, conspiracy theories, and, in general, the presence of polarized environments have contributed to the increase of what is referred to as *hate speech*, i.e., the presence of offensive and inappropriate language that promotes forms of intolerance, violence or hostility [5]. The deliberate circulation of these expression forms in online debates can have a remarkable impact on the physical and mental health of users, also promoting radicalization. Therefore, it is crucial to understand and characterize the interplay between misinformation and its effects on toxicity in online conversations. To achieve this goal, in this dissertation, we explore the interplay between the dissemination of misinformation and its impact on online discourse, providing advancements in identifying the elements attributable to user segregation and how moderating systems can improve to mitigate these phenomena.

1.1 Echo Chambers

Echo chambers, or filter bubbles, are defined as a group of like-minded individuals that promote and share content referring to the predominant ideology of the group itself. It has been shown [4] how echo chambers can nourish existing opinions in a group and promote ideological segregation of the users within. Quantitatively, results [6, 7, 8, 3] show how the interaction of users composing these bubbles reinforces the opinions of the members due to repeated contact with like-minded individuals. Such groups were observed to include members whose news diets were composed of sources following similar narratives, with no evidence of heterogeneity. At the same time, debunking acts [9] have been shown to produce negative sentiment by targeted users, reinforcing their position towards a topic instead of changing their minds.

1.2 Toxicity Dynamics

In online debates, the definition of what is referred to as toxic content has changed progressively through the years, influenced by cultural evolution and the many research areas that delve into this topic. Preliminary works defined as *hateful discourse* any form of expression where the author exhibits intense hatred towards an individual or group based on their identity [10]. From this definition, researchers focused on the various consequences that hateful discourses may bring, such as harassment [11, 12], the dissemination of false information [13] and trolling [14, 15]. By looking at the dynamics surrounding toxic discourses, it has been observed that users tend to focus their negative actions on a limited number of discussion threads [16], without clear indications of the presence of “pure haters" [17]. Nevertheless, it has been demonstrated that such exchanges do not significantly contribute to the proliferation of false information on social media [13].

1.3 Misinformation and Society

The exploration of misinformation is driven by its substantial impact on public society. Indeed, the influence of social media on various aspects, such as political elections and behavioral adoption, continues to be an unresolved issue, with concerns regarding its potentially detrimental effects on the democratic process. For instance, the involvement of fake news and bots in the 2016 U.S. presidential election on Twitter has raised apprehensions [18]. However, recent findings indicate that users' inclination to consume fake news may be contingent upon their political affiliations [19], and have identified that misinformation is mainly consumed by a specific subset of users with distinct characteristics [20]. Moreover, recent years have been characterized by global threats affecting the entire globe, such as the COVID-19 pandemic [21], climate change [22], and war conflicts [23]. These threats show the interconnected nature of our society and, on a digital landscape, how misinformation-induced segregation can translate into real-world events affecting people's lives [24, 25, 26]. This scenario, therefore, evidences how digital policies are essential in maintaining stable democracies and why studies about misinformation in society should drive them.

1.4 Content Moderation

The decentralized nature of online platforms has contributed to the emergence of toxic behaviours, misinformation, harassment and, in general, actions from users with the potential to harm the physical, psychological and social integrity of other users. Social media regulators, to contrast the presence of these behaviors, have been performing what is known as *content moderation*, defined as “the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [27]. The enforcement of moderation policies has evolved during this time, ranging from simple content removal or user banning, even with the help of user's reports, to more sophisticated solutions that involve the use of machine learning models that automatically detect content that are not allowed both on text or on multimedia contents [28, 29, 30]. Further, less invasive approaches like shadow banning or informed warnings were also enforced to temporarily restrict user's activity while educating themselves on a more responsible use of the digital space. Despite the impartial and more objective benefits that the designers of these systems promote to motivate the introduction of these decision tools, scholars and public opinion have expressed mainly their doubts about the fairness of decision systems in general [31, 32, 33, 34]. Indeed, the bias in the training data or the limitation imposed by the machine learning model has been observed to produce discrimination [35, 36] or lack of intervention [37], creating disparities through the users subjected to the decision of these algorithms.

1.5 Advancements

In this thesis, we investigate the critical factors that may be responsible for the evolution of toxicity in online conversations. To achieve this goal, we rely on data from mainstream social media platforms such as Facebook, Instagram, Twitter, Telegram, YouTube and Reddit or from unregulated ecosystems like Gab and Voat. To provide a comprehensive understanding of the mechanics behind news consumption, we extend social media data with the GDELT Event Database [38], gaining access to a worldwide database of events and news. Similarly, to quantify the evolution of toxicity dynamics over an extensive period, we include data from Usenet, gaining access to conversational data from the early stages of the Internet. The choice of using social media data is motivated by the digital interconnected world where society lives. Therefore, the answering of such questions can be performed by relying on the digital traces individuals leave through their day-to-day actions. This approach contributed to the emergence of *computational social science*, which aims to leverage the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviours [39]. The usage of digital data and the establishment of proper replication frameworks [40] can help to contrast the problem of scientific reproducibility, primarily known in the field of psychological and social science [41, 42]. The results of computational social science approaches have helped the emergence of insights related to several aspects of human interaction dynamics concerning news consumption and selective exposure [43, 41], polarization [19, 44, 45, 46, 47], echo chambers [4, 48, 49, 3] and toxicity dynamics on conversations [50, 51, 52, 13]. Moreover, in the case of worldwide events like the COVID-19 pandemic, digital data has helped to understand the effects of reopenings in communities [53], the tendency of users to consume misinformation rather than trustworthy news [24] or the economic and social consequences on specific countries [54]. Despite these promising opportunities, applying big data and computational approaches has raised a series of limitations that need to be considered. First, the increasing restrictions from big tech companies has remarkably reduced the access to data from researchers. Second, such data, even if accessible, may be incomplete, leading to partial or misleading results. Finally, results from studies on digital ecosystems, especially in the case of social media, may depend on specific settings of the environment itself. Any change conceiving these aspects, e.g. a different tweaking on recommendation/moderation algorithms or modifications in how users can interact and/or share content, may significantly impact the results of a study, making it irreproducible or pointing to different results.

Given the current research landscape, in this dissertation, we conduct an analysis to investigate the interplay between social media regulation policies and their effect in contrasting conspiracy theories and hate speech. In the first part, we provide an overview of the evolution of conspiracy theories on different social media. Then, we evaluate the interplay between conspiracy theories and the moderation enforced by these platforms by comparing Twitter and Gab during 2020. In the second part of the thesis, we investigate the concept of hate speech and toxicity in conversations. To do so, we quantitatively characterize engagement dynamics on news articles circulating on Facebook concerning several topics, assessing whether virality is more linked to the collection of adverse reactions from users. Then, we continue the investigation of controversial topics on Twitter and YouTube by comparing the topological and toxicity aspects of the Italian community on the 2022 Italian Political elections, representing a controversial topic, against the discussions around Italian Football, representing a topic close to the popular culture of the community. Finally, we exploit the analysis of toxicity dynamics by comprehensively analyzing conversation dynamics across eight platforms over 30 years—from the advent of Usenet to contemporary platforms, explicitly focusing on invariant patterns of toxic content across different platforms.

Our results show the presence of conspiracy theories nowadays affects all social media platforms, from mainstream to niche ones, involving all kinds of users, from genuine believers to debunkers and simple observers. When moderation policies are applied, we observe they have a concrete impact in reducing questionable contents, with the counter effect of establishing digital environments, referred to as *echo platforms*, with predominant narratives among all users. Despite the promising results in assessing the effects of moderation, the analysis of engagement dynamics on news pieces unveils how the virality of topics is associated with controversial reactions from users. From a conversational perspective, controversial topics, have been shown to produce more toxic conversations on average than their counterpart, involving more users with diverging opinions that gave birth to longer conversations compared to their counterparts. Furthermore, we showed how moderating tools that can identify toxic comments in a conversation might benefit from an ensemble approach with models trained on different stages of conversations. Finally, by extending the analysis of toxicity in online conversation to a broader perspective, we demonstrate how ideological polarization of opinions among users is one of the primary drivers of toxicity. Finally, we quantify how toxicity is associated with the length of conversations, how toxic language does not invariably lead people to leave a conversation, and that it does not necessarily escalate as discussions evolve.

The thesis is organized as follows. In Chapter 3, we provide an overview of conspiracy theories and their diffusion on social media platforms. In Chapter 4, we perform a first assessment of the effects of moderation policies in contrasting the circulation of questionable news during COVID-19. In Chapter 5 and 6, we investigate news from an engagement perspective, understanding the role of virality in the perception of users concerning specific topics. In Chapter 7, we assess the difference in the structure and the toxicity of conversations for the Italian community by comparing discussions about the 2022 Italian Elections against the Italian Football League. We then conclude our toxicity analysis in online conversations by expanding the insights of Chapter 8 with a longitudinal, multi-platform analysis. The conclusion and the future directions of this thesis are presented in Chapter 9.

Chapter 2

Preliminaries and Definitions

In the following chapter, we describe the methodologies and techniques employed in the thesis to extract, transform, and analyze data.

2.1 Data Collection

Data collection includes content from social media platforms, ranging from textual and multimedia elements to user activity, as well as news articles from all over the globe. In both cases, the gathering process was conducted through designated API services or open datasets following GDPR.

2.1.1 Facebook and Instagram

Meta, the company that owns Facebook and Instagram, allows researchers to collect posts from these platforms through CrowdTangle [55], a tool that tracks interactions on public content from Facebook pages, groups, and verified profiles. This tool provides content through different approaches, e.g., by conducting a keyword search or looking at all posts published by a set of accounts during a specific period. However, CrowdTangle does not allow the download of comments from the retrieved posts, and it does not include paid ads unless those ads began as organic, non-paid posts that were subsequently "boosted" using Facebook's advertising tools. Finally, CrowdTangle also does not include activity on private accounts or posts made visible only to specific groups of followers [56].

2.1.2 Reddit

The collection of posts, comments and user information from Reddit is available through their API. However, before the restrictions applied by Reddit concerning its data availability [57], data were mainly collected from the Pushshift service, which served as an archive of Reddit content to researchers interested in studying social media.

2.1.3 Twitter

Twitter, before the rebranding into X [58], provided privileged API access to the research community, known as Academic API Program [59]. Within this program, researchers could download posts, content and other information concerning user interactions, following and followers.

2.1.4 Voat

Voat was a news aggregator website that ceased to exist at the end of 2020. Prior research efforts led to the creation of a dataset publicly available to the community [60], which covers the entire lifetime of the platform from 8/1/2013 to 25/12/2020.

2.1.5 YouTube

YouTube offers the opportunity to collect data from its platform by accessing the YouTube Data API [61]. It allows the retrieval of different kind of information, like the subscribers of a channel, its videos, comments with related information, as well as the suggestions that the platform makes based on previous views.

2.1.6 Gab

Gab does not provide an official API service. However, the fact that the Gab structure relies on the one from Mastodon [62] allows researchers to gather data by performing HTTP requests to its API endpoints. With this approach, careful attention must be paid during data collection due to the GDPR.

2.1.7 Usenet

Usenet is a distributed discussion systems created in 1980 [63]. It is organized with a hierarchy of topics, where each of them is divided into different subjects. Since Usenet belongs to the early days of the Internet and therefore constitutes a cultural value of digital history, the collection of its data was performed by querying the Internet Archive [64]. Then, data is processed to reconstruct the conversation cascades of the different threads, resulting in a comparable structure with the current social media in circulation.

2.2 Data Transformation

2.2.1 URL Expansion

To infer the origin of news published on social media platforms, we classify the reliability of the domain referring to the news. These domains are extracted from the URL of the news article, which sometimes is shortened to match character limits. Therefore, a preliminary data transformation step in the analysis consists of resolving the URL back to its original form, with the ability to classify the publishing news outlet correctly.

2.2.2 News Outlet Classification

To evaluate the reliability of information circulating on both social media, we employed a source-based approach. We built a dataset of news outlets' domains from our dataset where each domain is labeled either as *Questionable* or *Reliable*. The classification relied on two fact-checking organizations called MediaBias/FactCheck (MBFC, https://mediabiasfactcheck.com) and NewsGuard (NG, https://www.newsguardtech.com/). On MBFC, each news outlet is associated with a label that refers to its political bias, namely: *Right, Right-Center, Least-Biased, Left-Center, and Left*. Similarly, the website also provides a second label that expresses its reliability, categorizing outlets as *Conspiracy-Pseudoscience, Pro-Science* or *Questionable*. Noticeably, the *Questionable* set includes a wide range of political biases, from *Extreme Left* to *Extreme Right*. For instance, the *Right* label is associated with Fox News, the *Questionable* label to Breitbart (a famous right extremist outlet), and the *Pro-Science* label to *Science*. MBFC also provides a classification based on a *ranking bias score* that depends on four categories: *Biased Wording/Headlines, Factual/Sourcing, Story Choices,* and *Political Affiliation.* Each category is rated on a $0-10$ scale, with 0 indicating the absence of bias and 10 indicating the presence of maximum bias. The *bias outlet score* is computed as the average of the four score categories. Likewise, NG classifies news outlets into four categories based on nine journalistic criteria, each of them having a specific score whose sum ranges between 0 and 100. Outlets with a score of at least 60 points are considered compliant with the basic standards of credibility and transparency. Otherwise, they are recognized as outlets that lack of credibility. A different characterization is provided for humor and platforms websites, not accounting for the categorization process.

2.3 Measuring user homophily in social network

2.3.1 User Leaning

To measure the extent to which a user is associated with the consumption of questionable or reliable contents, we introduce the *user leaning* q. We define it in the range $q \in [0,1]$, where 0 means that a user posts contents exclusively associated with reliable sources, and 1 means that a user puts into circulation only questionable posts.

Formally, the user leaning can be defined as follows: let $\mathcal{P}$ be the set of all posts with a URL matching a domain in our dataset and $\mathcal{U}$ the set containing all the users with at least a categorized post. At each element $p_j \in \mathcal{P}$ is associated a binary value $l_j \in \{0,1\}$ based on the domain of the link contained: if the URL refers to a domain classified as questionable then $l_j = 1$, otherwise $l_j = 0$. Considering a user u_i in a bipartite network between users and posts, then the user leaning q_i of a user u_i can be defined as:

$$q_i = \frac{1}{k_i}\sum_{j=1}^{k_i} l_j \ , \tag{2.1}$$

where l_j is the leaning score of the j-th neighbor of the user u_i, and k_i is the number of categorized contents that the user posted.

2.3.2 Homophily in the Interaction Network

Given a network of users, homophily can be defined as the nodes' tendency to interact with others with similar characteristics. In network terms, this translates into a node i with a given leaning x_i more likely to be connected with nodes with a leaning close to x_i [65]. In this thesis, this concept can be assessed by defining, for each user i, the average leaning of their neighbourhood as

$$x_i^N \equiv \frac{1}{k_i^{\rightarrow}} \sum_j A_{ij} x_j, \tag{2.2}$$

where A_{ij} is the adjacency matrix of the interaction network, $A_{ij} = 1$ if there is a link from node i to node j, $A_{ij} = 0$ otherwise, and $k_i^{\rightarrow} = \sum_j A_{ij}$ is the out-degree of node i. The presence of homophily is assessed by studying the relationship $x_i \sim x_i^N$.

2.4 Networks

2.4.1 Definition

The basis for the conceptualization of a network is a graph $G = (V, E)$, being V the set of n nodes and E the set of m edges. The nodes are denoted as $i, j \in V$ or, similarly, $i, j = 1, \ldots, n$, and the edge that formalizes the connection between i and j is denoted as $(i, j) \in E$. We denote as Å the adjacency matrix of a graph G, which is a n-squared binary matrix taking values 0 or 1, where the element $A_{ij} = 1$ if nodes i and j are connected and $A_{ij} = 0$ otherwise.

2.4.2 Bipartite Graph

The bipartite graph is a graph in which the vertex set V is the union of two disjoint independent sets called the partitions of G. The equivalent of an adjacency matrix for a bipartite graph is a $h \times p$ rectangular matrix called incidence matrix B that takes values 0 or 1, where the element $B_{ij} = 1$ if nodes i and j are connected.

A bipartite graph can be easily projected onto one of its partitions by performing an operation called one-mode projection that can be formalized in terms of the product $P = B^T B$, in the case we are projecting onto the partition of size p, and $P = BB^T$ if we are projecting onto the partition of size h. P is a symmetric matrix whose elements P_{ij} are nonnegative numbers that represent, in the case of off-diagonal elements, the number of shared links of the nodes i and j to the partition of size h or p. The diagonal elements of the matrix P are also nonnegative numbers that represent the degree of the node in the bipartite graph. Since the elements on the diagonal of the matrix P have a different meaning concerning the elements away from the diagonal, it is common practice to set the diagonal elements $P_{ii} = 0$.After such treatment, the matrix P can also be called the co-occurrence matrix, where two elements are interconnected if they share at least one partition with an external node. Also, the number of co-connections between i and j is represented by the link weight, i.e., by the element P_{ij} of the matrix P.

2.4.3 Tree Graph

A tree graph can be defined as a pair $T = (V, E)$, where $V = \{1, \ldots, n\}$ represents the set of nodes and $E = \{1, \ldots, m\}$ the set of links. We consider directed trees, with n nodes and $m = n - 1$ links.

2.5 Tree Structural Metrics

Size

The tree size is the number of nodes in the tree, denoted as $n = |V|$, where $|\cdot|$ is the cardinality of the set V. In this dissertation, we assume the size is the total number of replies in the first mentioned conversation tree, assuming that a user can post multiple replies and interact with different users within the conversation.

Depth

The tree depth $D(T)$ is the distance d of the deepest node in the conversation, which also coincides with the tree's diameter, i.e., the longest shortest path between the root node and any other node in the graph. The depth can be expressed as follows: $D(T) = max\ (d_{rj})\ \forall j\ ,\ j \neq r$ where r is the root node.

Wiener Index

The Wiener index measures the structural complexity of the tree and its potential virality [66] and is defined as the average shortest path between each pair of nodes i, j. In the case of a directed tree, the Wiener index can be defined as:

$$W(T) = \frac{2}{n(n-1)} \sum_{i} \sum_{j>i} d_{ij} \tag{2.3}$$

where $\frac{2}{n(n-1)}$ is a normalization factor to account for all paths among couples of nodes. The Wiener index ranges between $[1, \infty)$ and, in general, it is minimized for broadcast structures and maximized for low branching structures [66].

Toxicity Ratio

The toxicity ratio is the average number of toxic nodes in the conversation tree T, considering the number of toxic replies out of the total number in the conversation. The toxicity ratio can be defined as

$$TR(T) = \frac{card\{V \mid s > t\}}{card\{V\}}, \tag{2.4}$$

where s is the toxicity score of the comment $v \in V$ and t is the toxicity threshold value, such that a comment that satisfies the equation $s > t$ is considered toxic.

Average Toxicity Distance

The average toxicity distance is the average normalized distance of toxic comments from the root, defined as

$$TD(T) = \frac{1}{card\{V \mid s > t\}} \sum_{j} \frac{d_{rj}}{D(T)}. \tag{2.5}$$

TD(T) is bounded in $(0, 1]$, and low values of this quantity imply that toxic comments are, on average, located close to the root.

Assortativity

The assortativity coefficient r measures the extent to which similar nodes tend to be connected with each other [67]. Being the analogue of Pearson's correlation coefficient, it varies in the range $[-1, 1]$ with negative values indicating disassortativity (i.e., nodes with different features tend to be interconnected more than expected at random) and positive values indicating assortativity (i.e., nodes with similar features tend to be interconnected more than expected at random). Assortativity values close to zero are related to the distribution of node features close by chance. We consider as node feature their toxicity score, and to compute the assortativity coefficient, we ignore the direction of the edges, obtaining the following equation:

$$r(T) = \frac{\sum_{ij}(a_{ij} - \frac{k_i k_j}{2m})x_i x_j}{\sum_{ij}(a_{ij}x_i^2 - \frac{k_i k_j}{2m}x_i x_j)}, \tag{2.6}$$

where a_{ij} is the element the adjacency matrix $A = (a_{ij})_{i,j \in V}$ in which $a_{ij} = 1$ ($a_{ij} = 0$) indicates the presence(absence) of an edge between nodes i and j, $k_i = \sum_{j=1}^{n} a_{ij}$ is the node degree,and x_i is the feature assigned to node i.

2.6 Statistical Tools

2.6.1 Lifetime Estimation

To estimate users and pages based on the different interactions performed or received during their existence, we employ the Kaplan-Meier estimator. It is a non-parametric statistic for quantifying a survival function defined on discrete interval times. Let $S(t)$ be a function representing the probability of having a lifetime greater than the time t, such that

$$S(t) = P(\rho > t), \tag{2.7}$$

where $t = 0, 1, \ldots$. However, in real-life cases, the true survival function $S(t)$ is never known. Therefore, we define an estimator, which is the fraction of observations that survived for a specific amount of time t_i, where $i = 1, \ldots, T$. This results in the following definition

$$\hat{S}(t) = \prod_{i:\ t_i \leq t} (1 - \frac{d_i}{n_i}), \tag{2.8}$$

where t_i is the time when at least one event happened, d_i is the number of events (e.g., deaths) that happened at time t_i, and n_i represents the number of observations at risk, i.e., the individuals known to have survived up to time t_i, which means that they did not die or they have been censored instead. To summarize, this estimator computes, at each time, t_i, the product of the survival until that time.

Chapter 3

Conspiracy Theories and social media platforms

Social media platforms have remarkably shaped how users inform themselves, interact with other peers and perceive the world. Such changes have unveiled psychological mechanisms on what news users consume and how they decide their peers to be surrounded with online. It is, therefore, crucial to understand the digital landscape where information circulates, addressing the potential threats and causes that affect its consumption.

In this chapter, we start the thesis by introducing an overview of news consumption patterns related to conspiracy content on mainstream (Facebook, Twitter, YouTube, and Reddit) and niche social media platforms like Gab. In such a context, opinion polarization and echo chambers are pivotal communication elements around conspiracy theories. A relevant role may also be played by the content moderation policies enforced by each social media platform. Indeed, banning content or users from social media could lead to a level of user segregation that goes beyond echo chambers and reaches the entire social media space, up to the formation of "echo platforms". The insurgence of echo platforms is a new online phenomenon that needs to be investigated, as it potentially fosters many dangerous phenomena that we observe online, including the spreading of conspiracy theories.

3.1 Echo chambers and conspiracy in social media

The advent of social media had a profound impact on how people access information and interact online. Users tend to acquire information they like, filter out information they do not, and join groups of like-minded peers around a shared narrative called echo chambers [45, 68, 3]. According to group polarization theory, an echo chamber can act as a mechanism to reinforce existing opinions moving the entire group toward more extreme positions. In many instances, conspiracy theories are the pivot around which echo chambers develop and grow [6, 3, 69]. Considering that the spreading of conspiracy theories can have potentially harmful consequences for individuals and societies [70, 71], understanding the proliferation of such theories in online environments, especially in the context of an infodemic [72, 73], becomes of fundamental importance.

A relevant aspect concerning the spreading of conspiracy theories online is empirical evidence for the fact that individuals endorsing conspiracy content were highly engaged and more responsive to endorse deliberately false or other questionable information [45]. Indeed, in the context of internet and social media, conspiracy theories seems to be strongly related to misinformation, with which they share many aspects ranging from the presence of questionable elements in their narrative to the reasons why they appeal to potential believers [74]. Also, the way in which conspiracy theories propagate in online communities seem to present structural features that are remarkably similar [75] to those of (mis)information cascades happening on Facebook [68] and Twitter [76]. However, results concerning structural differences in information cascades should be taken with caution [77] given the intrinsic limitation of a false/true or science/conspiracy dichotomy and the inherent unbalance of datasets referring to online content due to the moderation policies affecting them. [78].

Like misinformation, conspiracy theories appear on most social media platforms, from mainstream to niche ones, involving all kind of users, from genuine believers to debunkers and simple observers [79]. While non-conspiracy information by news outlets outpaces conspiracy-related content, the relevance of the users' engagement around conspiracy is not negligible [68, 76].

In the following sections we provide an overview of the typical information consumption patterns on Facebook, Twitter, YouTube, and Reddit in relationship to conspiracy content. In the last section, we briefly discuss results about niche social media platforms.

3.1.1 Facebook

In the past years Facebook has been accused of being a vehicle for conspiracy theories and misinformation spreading arguably more than other social media platforms, facing criticism on a number of themes including vaccine hesitancy[1], the Russian interference in the 2016 U.S. elections [2] climate change denial, [3] the so-called "infowars" case,[4] and more recently the role played in fomenting the 2021 U.S. Capitol Hill attack[5] and in the debate around COVID-19[6]. Despite that, quantitative investigation on misinformation/conspiracy consumption on Facebook has been somewhat limited - e.g. compared to Twitter - possibly due to increasing restrictions in data accessibility following the Cambridge Analytica case in 2018, when a private firm used Facebook users' data without consent to profile individuals and send them personalized political advertisements.[7]

However, the presence of strong polarization and echo-chambers on Facebook, arguably facilitated by the News Feed (now just Feed) algorithm [41], is well-documented and it is likely to be among the primary drivers in the information diffusion dynamics on the platform [45, 6, 80, 3, 81]. The characterization of conspiracy news consumption and spreading has often been based on a comparison with that of scientific content. Within this frame-

[1] www.theguardian.com/technology/2019/nov/13/majority-antivaxx-vaccine-ads-facebook-funded-by-two-organizations-study

[2] www.theguardian.com/technology/2021/apr/12/facebook-fake-engagement-whistleblower-sophie-zhang

[3] www.scientificamerican.com/article/climate-denial-spreads-on-facebook-as-scientists-face-restrictions/

[4] www.vox.com/2018/7/16/17577426/media-left-right-facebook-define-journalism

[5] www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/

[6] www.thetimes.co.uk/article/facebook-page-for-covid-conspiracy-theorists-has-hundreds-of-thousands-of-followers-7c285b05f

[7] https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

work, the existence of conspiracy theory related community structures with echo-chamber-like behavior has been reported and it was shown that polarized consumers of conspiracy content are highly focused on posts pertaining only to their own community and very active in diffusing such contents [45, 80, 82]. These users tend to be little to not affected by exposure to debunking posts, and their community shows a dominant negative attitude in response to them [9], suggesting the inefficacy of the debunking approach within echo chambers. Importantly, the (automatic) identification of conspiracy theories on Facebook proves difficult. In [77] the authors devised a classifier relying on the structural features of the content propagation cascades to find that conspiracy and science content reverberates in a way which is hard to distinguish from one another.

Recently, Facebook has intensified its efforts against the proliferation of conspiracy theories and misinformation[8] through enhancing content moderation activities on the platform (see e.g. the case of Qanon[9]). The effectiveness of those actions, however, is currently still debated [83, 84].

3.1.2 Twitter

In the last decade, Twitter has been widely used as a workbench for studying social phenomena including the spreading of conspiracy theories and misinformation. Some examples include Brexit [85], the Catalan referendum [86], the US presidential elections [18, 19], and the COVID-19 vaccines debate [24]. One of the Twitter peculiarities is the presence of automated accounts often used to amplify the diffusion of controversial content about different topics [18, 19], including conspiracy theories. The role of automation [87] and, more in general, coordination [88], combined with the recommendation algorithm and biases in user choices fostered the emergence of polarization and echo chambers on the platform. Echo chambers has been detected around strongly debated topics such as abortion, gun control and climate change [3]. Moreover, some communities tend to be consistent across topics: for example, climate change deniers are usually closer to conservative political position, while activists tends to sympathize for liberals [89].

[8] https://about.fb.com/news/tag/misinformation/

[9] https://www.bbc.com/news/world-us-canada-54443878

Twitter is actively engaged in combating the spreading of problematic content by enforcing several strategies, from prohibiting political advertisement, to moderation and accounts ban. The results of these actions seem to impact the amount of misinformation and conspiracy theories circulating, such as in the case of the 2019 European elections [90] or 2020 US presidential elections [91]. Yet, conspiracy theories may still be popular among some groups of users [92].

3.1.3 YouTube

Research involving the role of conspiracy theories (and misinformation) on YouTube is quite recent. A motivation for this new line of research involving YouTube data is perhaps due to the media coverage received by the platform for what concerns its role in exacerbating users' opinions by means of its video recommendation algorithm[10]. According to some studies [93, 94], the algorithm seems to be responsible for the creation of the filter bubble and eventually of rabbit holes, i.e., loops of questionable and conspiracy contents suggested by the algorithm, creating a vicious circle of problematic recommendations. Despite a continuous effort to moderate inappropriate and problematic videos and comments, YouTube, just like other platforms, hosts a number of conspiracy related contents. Furthermore, a recent research involving three different conspiracy theories has shown that videos related to conspiracy display a higher popularity in terms of number of views, than videos aiming at debunking them [95]. This result is in line with previous studies on the platform [96] according to which conspiracy users tend to interact more with like-minded peers. Evidence for the presence of echo chambers was also found during the COVID-19 debate [97], after categorising channels along two dimensions corresponding to their reliability and political bias. As expected, a wide share of low-reliability and high-bias channels were also responsible for sharing conspiracy videos. Nevertheless, evidence about echo chambers in partisan discussion is still debated [98, 99].

[10] https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

3.1.4 Reddit

Given its growing popularity, Reddit has recently captured the attention of researchers. Reddit posts are gathered within communities of interest called "subreddits", covering specific topics and actively moderated by their members. On this social media, the recommendation algorithm apparently does not nourish the echo chamber effect [3, 100], with people interacting with news from a relatively wide set of sources. The lack of polarization shown by Reddit users found support in the study of misinformation and opinion dynamics [100]. Despite Reddit's low users polarization, researchers put a remarkable effort into investigating how conspiracy theories evolve on the platform and how certain users may still radicalize around specific narratives. Analyses of the linguistic dimension [101] was employed to show that similar narrative motifs are shared among different, even unrelated, conspiracy theories, or to identify early warnings of users departure from conspiratorial communities [102]. Data from Reddit have been used also to investigate the social dynamics associated to joining conspiratorial communities [103], highlighting marginalization from other communities as the main driver for such a phenomenon. Eventually in [75], researchers found structural differences in discussion cascades between conspiracy and science related subreddits.

3.1.5 Other social media platforms

Less popular social media platforms are active matter of investigation, as they are often populated by users hit by moderation policies (e.g. bans) on mainstream platforms. Gab [62, 81], Voat [60] and Parler [104] are among the most studied alternative platforms ensuring “free-speech” to their users, in which conspiracy theories and related discussion proliferate [62, 105, 104, 81]. Together with social media platforms even messaging platforms such as Telegram [106] are getting researchers' attention for their user base. Interestingly, before and during the Capitol Hill riot, Gab and Parler registered a higher average activity, with a contextual increase in the usage of anti-social language [107].

3.2 Regulation and the risk of echo platforms

Social media platforms present rather different features with respect to diffusion and hosting of conspiracy theories that may be associated with differences in the implemented feed algorithm and enforced moderation policies [43, 3, 108].

A primary problem is that banning problematic users from one platform may induce their migration to other ones where no moderation is enforced; this was the case of Infowars, a far-right conspiracy website whose owner was banned by Facebook, Twitter and YouTube, but gained large support (more than 200k followers) on Gab. One one side, his ban reduced the overall number of users exposed either accidentally or algorithmically to borderline content, on the other the event may have contributed to exacerbate the radicalisation of his original audience and induced a migration in an even-more polarised environment.

In this regard, it was observed how the presence of regulation policies plays an important role in shaping people exposure to questionable content [81]. Indeed, by comparing a regulated and an unregulated social media, namely Twitter and Gab, it was demonstrated how the presence of moderation pursued by Twitter significantly reduced questionable content, with a consequent affiliation towards reliable sources in terms of engagement and comments. Conversely, the lack of precise regulation on Gab resulted in the tendency of users to engage with both types of content, showing a preference for the questionable one.

As we described in Section 3.1.5, the cases of Gab, Parler and Voat could not be singletons, since similar situations could be re-created on new social media as, for instance, The Truth created by Donald Trump. More generally, content moderation may have unexpected effect at a global level. Previous work observed the effect of YouTube algorithmic moderation (i.e., reducing the visibility of certain videos) of conspiracy contents was effective on the platform [94] but, at the same time, it inflated the spread of the same videos (shared as YouTube links) on other social media platforms such as Twitter and Reddit. Similarly, the act of banning users and whole communities of conspiracy content on Reddit positively correlated with migrations of users on

Voat that was hosting akin groups [60].

The new platforms reached by migrating users possess the features of “echo platforms”, that can be defined as a social media platform colonized by users belonging to a specific echo chamber. This new online phenomenon calls for further investigations, as it could potentially add a multiplicative factor to many potentially problematic phenomena that we observe online, including the spreading of conspiracy theories.

3.3 Conclusions and future works

In this chapter, we reviewed several contributions related to the circulation and communication of conspiracy theories on a comprehensive set of social media platforms, from mainstream to niche ones. Overall, polarization and echo chambers seem to be two main aspects that characterize social dynamics around conspiracy theories. These two phenomena may be somewhat reinforced by the role of recommendation algorithms and moderation policies differing from platform to platform.

In particular, the effects of the platforms’ moderation policies that is banning or penalizing controversial content’s visibility remain unclear. If, on the one hand, moderation may reduce the visibility and spread of conspiracy theories and misinformation, on the other hand, it may trigger users’ migration toward less regulated platforms. This phenomenon can shift the structure of the online environment from echo chambers to echo platforms, where users may join different social media based on their narrative instead of segregating into communities with opposite opinions, definitely reducing their exposure to a diverse set of contents.

Several aspects of conspiracy theories remain not clear. First, little is known about the dynamics of the conspiracy theory’s popularity, especially at the early stage. Why one theory becomes popular online is not yet well understood, and it is crucial to implement countermeasures to mitigate their spreading. Second, the advantages and disadvantages of actions such as the user or content banning and the dismantling of communities debating around conspiracy theories. Future works should focus and clarify both aspects.

Chapter 4

Comparing the impact of social media regulations on news consumption

In Chapter 3, we provided an overview of the impact of social media platforms on information consumption patterns, peer interactions, and the potential consequences for society and democracy. We delved into the world of conspiracy content on mainstream social media platforms, particularly Facebook, Twitter, YouTube, and Reddit, focusing on niche platforms. In doing so, we explored the role of opinion polarization and echo chambers as central elements in disseminating conspiracy theories. We also underlined the importance of content moderation policies enforced by these platforms, highlighting the potential consequences of banning content or users, which could lead to user segregation on a broader scale, ultimately giving rise to what we've termed "echo platforms".

Given the current digital landscape, in this chapter, we quantitatively compare two social media that enforced opposite moderation methods, Twitter and Gab, to assess the interplay between news consumption and content regulation concerning COVID-19. We analyze the two platforms on about three million pieces of content, analyzing user interaction concerning news articles. We first describe users' consumption patterns on the two platforms, focusing on the political leaning of news outlets. Finally, we characterize the echo chamber effect by modelling the dynamics of users' interaction networks. Our results show that the presence of moderation pursued by Twitter produces a significant reduction of questionable content, with a consequent affiliation towards reliable sources in terms of engagement and comments. Conversely, the lack of precise regulation on Gab results in the tendency of the user to engage with both types of content, showing a slight preference for the questionable ones, which may account for dissing/endorsement behaviour. Twitter users show segregation towards reliable content with a uniform narrative. Gab, instead, offers a more heterogeneous structure where users, independently of their leaning, follow people who are slightly polarized towards questionable news.

4.1 The COVID-19 infodemic on social media

The COVID-19 outbreak [109], which was declared as a pandemic by the World Health Organization (WHO) on 11 March 2020 [21], changed several aspects of our everyday life both in the online and offline sphere. For instance, the news diet of users was remarkably modified in its structure by introducing a considerable amount of information referring to a new topic. This phenomenon was accelerated by social media platforms, which are known for shaping discussions on a wide range of issues, including politics, climate change, economics, migration, and health [110, 111, 19, 46], unveiled how users online tend to consume information adhering to their system of beliefs and to ignore dissenting information. This selective exposure of users to specific pieces of content represents an important phenomenon to be taken into consideration, especially when users are exposed to a considerable amount of information referring to a new topic, like the COVID-19 pandemic [21], which generated an overabundant flow of information and news, whose trustworthiness may not always be guaranteed, especially online. This phenomenon, referred as infodemic [112, 2] reportedly affects people's behavior [113] in a harmful way. This aspect calls for urgent investigations of the turbulent dynamics of the online infosphere, complementary to the monitoring of the spreading of infections [114, 115, 108]. Indeed, the current infodemic may foster the tendency of users a) to acquire information adhering to their system of beliefs

[45], b) to ignore dissenting information [9], c) to form polarized groups around a shared narrative [6]. Two common factors to such behaviors carried on by users are opinion polarization [116], one of the dominating traits of online social dynamics, and echo chambers [3]. Divided into echo chambers, users account for the coherence with their preferred narrative rather than the actual value of the information [117, 118, 77]. Such evidence for polarization and online echo chambers seems to be related to a feedback loop between individual choices and algorithm recommendations towards like-minded contents [41, 43, 3]. However, other presumably harmless factors like the enforcement of content regulation may play a role in increasing online polarization. Indeed, it was recently observed that moderation policies and removal actions/bans of users produce adverse effects in terms of online polarization [119, 120, 121]. Users who got banned often consider this action as a badge of honor, rejoining the same social media under new identities or migrating to more tolerant platforms. The result could be either a reinforcement of their (extreme) opinion or reduced exposure to opposing voices. Therefore, raising awareness about the collateral costs of content policy and other interventions is crucial for making social media a less toxic environment.

In this study, we perform a comparative analysis between two social media platforms to study the differences of content circulation given the enforcement of different moderation policies. We select Twitter as a representative of content-regulated social media and Gab, a social network known for its willingness to ensure free speech by using little to no content moderation [62], like its counterpart. Despite their differences in how content policy is applied, both platforms are characterized by a similar platform design. Users are allowed to post and interact with content, together with their ability to create connections with other users. We perform our analysis on a timespan between 1/1/2020 and 30/09/2020, covering the first global wave of COVID-19. The dataset includes about three million posts and comments related to the COVID-19 topic expressed from more than one million users. We investigate consumption patterns from a user and post perspective on the two social media, assessing differences in terms of engagement. We extend this analysis by taking into account the trustworthiness of the contents published, classifying news sources accordingly to a categorization based on Media Bias/Fact Check [122] and NewsGuard [123]. An akin type of classification was exploited in several papers [19, 108, 3, 9] bringing essential insights on the circulation of misinformation online. Therefore, we employ this dichotomy by classifying posts as *Questionable* or *Reliable* depending on their credibility. The same labeling was used to model the persistence of users repeatedly commenting under a post of the same outlet category. Finally, we investigate the presence of homophily, i.e., the tendency of users to aggregate around common interests, by measuring the relationship between users and their tendency to post questionable content. We find that Twitter is characterized by the existence of two echo chambers of radically different sizes, in which the biggest one contains users more inclined to consume reliable content. In summary, the bulk of users on Twitter seems to share and interact with verified content.

Oppositely, users on Gab show a lack of a clear preference between the two types of outlets. Questionable posts are preferred in terms of commenting persistence. However, reliable posts are more likely to be commented on as time passes. Coherently, the existence of echo chambers on Gab is not as evident as observed in the case of Twitter due to the presence of users with a relatively heterogeneous leaning. We conclude that a valid content regulation policy produces tangible results in contrasting misinformation spreading.

4.2 Preliminaries and Definitions

In this section, we present the methodology applied in this study. We start by introducing the data collection process of posts from Twitter and Gab together with its categorization. Then, we describe the theoretical tools behind the analysis of engagement patterns, homophily and survival lifetime.

4.2.1 Data Collection

The collection of all posts related to COVID-19 was designed to capture the corresponding debate on social media by gathering posts and comments from both platforms on a period that ranges from 1/1/2020 to 30/09/2020. We first analyzed the most searched terms worldwide related to the pandemic on Google Trends. We selected four terms based on their interest and significance over time, namely: *coronavirus, corona, covid, covid19.* These terms served as a proxy to retrieve posts on the two social media whose hashtags matched exactly at least one of the four terms from Google Trends.

For Gab, we queried their API to obtain posts that exactly matched at least one of the search hashtags. Due to some modifications made by the platform during the study, the API stopped providing results in chronological order in June 2020. Therefore, we started collecting all posts from the general stream until the end of the analysis period, filtering by hashtag as we previously described. Such a shift in the collecting process did not affect the dataset. Indeed, searching by hashtag on Gab produces the same result as gathering all posts from a specific period and filtering with the same rationale. Then, to provide an equal comparison with Twitter, we considered only those posts from 28/01/2020. In the end, the collecting process produced an initial dataset of $\sim$ 204K posts, $\sim$ 130K of them containing a search hashtag and a link.

The collection of Twitter posts related to the COVID-19 pandemic relied on a public dataset [124] covering this specific topic. The dataset consists of a collection of tweet IDs, starting from 28/01/2020, which contains keywords and accounts that were trending at the time [124]. As the authors stated, due to the evolving nature of the pandemic and of online conversations [124], the list of accounts [125] and representative keywords [126] were constantly updated throughout the time. Due to rate limitations imposed by Twitter API, we retrieved up to $10K$ posts per hour each day, for a total of $\sim$ 2.6M posts in the entire analysis period.
In the end, we filtered these posts by retaining only those whose hashtags exactly matched at least one of the four search terms employed for the study and with a link. Such filtering reduced the dataset to a total of $\sim$ 1.1M posts.

4.2.2 Questionable and Reliable Sources

To categorize the trustworthiness of news outlets, we extracted the links from posts and obtained their reliability by associating the news outlet to the learning provided by MBFC and NG (see Section 2.2.2 for further details). On MBFC, all the outlets already classified as *Questionable* or belonging to the category *Conspiracy-Pseudoscience* were labelled as *Questionable*. The

Platform	Downloaded	Containing search hashtag and link	Categorized	Questionable	Reliable
Gab	205 458	130 864	83 784	49 772	34 012
Twitter	2 668 286	1 110 030	244 430	25 121	219 309
Total	2 873 744	1 240 894	328 214	74 893	253 321

Table 4.1. Data breakdown of posts for Gab and Twitter.

remaining categories were labelled as *Reliable*. Several other works have employed the use of MBFC to categorize posts for their trustworthiness [3, 108, 19, 127, 128], providing evidence of its reliability.

Coherently, outlets on NG were classified based on their score, maintaining the dichotomy provided by the website. We choose a score of 60 as a threshold to consider an outlet as *Reliable* (score > 60); otherwise, it is referred to as *Questionable* (score ≤ 60).

Considering a total of 2738 news outlets provided by the two organizations, 2701 belonging to MBFC and 37 to NG, we end up with 814 outlets classified as Questionable and 1924 outlets classified as Reliable. This labeling was employed to categorize the Gab and Twitter datasets obtained in Section 4.2.1. As a result of this, for Gab we obtained a dataset of $\sim$ 83K posts, $\sim$ 49K of them labeled as Questionable and the remaining $\sim$ 34K as Reliable, while on Twitter, we obtained a dataset of $\sim$ 244K posts, $\sim$ 25K Questionable and the remaining $\sim$ 219K as Reliable. In the end, the categorization process produced a total of $\sim$ 320K posts, $\sim$ 74K classified as Questionable and the remaining $\sim$ 250K as Reliable.

The total quantities obtained through the data collection and classification process are shown in Table 4.1.

4.2.3 Comparison of power law distributions

Most quantities related to the activity of users on social media show a heavy tailed distribution of discrete variables. Given the discrete nature of such distributions, we could not rely on Kolmogorov-Smirnov test [129] to assess whether two distributions present significant differences between each other. Indeed, such a test assumes that distributions must be continuous, and the presence of a large number of ties in the long-tailed distributions that we want to compare may lead to the computation of biased p-values. To overcome this issue, we employed a methodology proposed in *Zollo et al.*[9] which makes use of a Wald Test [130] to assess significant differences between the scaling parameters of two long-tailed distributions.

4.3 Result and Discussion

This study aims at performing a comparative analysis of two social media, namely Twitter and Gab, in order to understand how news consumption and social dynamics change in presence of two radically different types of content regulation policies (more stringent in the case of Twitter, almost absent in the case of Gab). The following results provide insights to explain this behavior from different perspectives. At first, we analyze the engagement of users with posts, which we consider as separated into two categories named questionable and reliable. Then, we quantify the commenting behavior of users and posts. Lastly, we provide a network analysis to measure the tendency of users to aggregate with like-minded peers, describing how the presence of content regulation may be correlated with the polarization towards specific narratives.

4.3.1 Consumption Patterns

We investigate how the engagement on the two social media differs in relationship with the COVID-19 topic. Figure 4.1 compares the engagement distribution for posts and users. Despite the differences in terms of scale that are attributable to the size of the platforms' user base, we observe that both frequency distributions are long-tailed. This feature provides a first evidence in the consumption of news, showing that interaction patterns are similar regardless of the content moderation imposed.

Next, we extend the analysis of consumption patterns by categorizing posts, based on their outlet leaning, into Questionable or Reliable. The resulting distributions from the application of this dichotomy are represented in Figures 4.2 and 4.3. Figure 4.2 displays the distribution of the number of likes and shares (reblogs or retweets) obtained by posts in our dataset, together with the corresponding cumulative. Similarly, Figure 4.3 describes the frequency distribution of the same kind of interactions from a user perspective. In general, we observe how Twitter users show higher levels of engagement with reliable posts, establishing a clear gap from questionable ones that increases during the analysis period. This difference can be attributed to the commitment of Twitter to limit the spreading of unverified contents [131]. The opposite scenario happens on Gab, in which the consumption patterns do not show a clear sign of polarization towards a specific kind of narrative. This provides some evidence of how users belonging to segregated environments like Gab are not interested in the origin of the content itself. Instead, they tend to self-segregate within environments in which they can consume and spread questionable content. Therefore, the lack of regulation on this platform may allow them to perform information operations [132], i.e., a category of actions taken by organized actors (governments or non-state actors) to distort domestic or foreign political sentiment, against other users who do not share the mainstream system of beliefs of the community.

In order to assess the similarity between the distributions deriving from the consumption patterns of questionable and reliable posts, we fit power-law distributions to such data and perform a statistical evaluation of their scaling parameters using the Wald test. For Gab, all the obtained p-values were significantly higher than 0.05, describing how questionable and reliable distributions are characterized by similar distribution patterns despite the difference in size. The same behavior is found on Twitter, except for the likes distribution whose p-value is less than 0.001, describing a significant difference in the way users engage with questionable and reliable content.

We can conclude that content moderation is a remarkable difference between the two platforms. Indeed, Twitter displays a higher presence of reliable than questionable content. In line with this result, reliable posts receive higher levels of engagement. Conversely, Gab appears to be associated with a more heterogeneous leaning of the users. This heterogeneity may provide a warning about possible misinformation operations conducted by users in the platform. In the end, we observed how moderation may play a role in the contrast of misinformation and how it may be responsible for the emergence of segregation among users whose news diet is mainly based on contents targeted by the regulation policies.

4.3.2 Characterizing Commenting Behavior for Questionable and Reliable posts

To quantify the persistence of comments concerning users and posts, we employed Kaplan-Meier estimates of two survival functions. The first accounts for the period between the first and last comment received from posts. The second instead considers the period between the first and last comment made by a user. To characterize any significant difference in the two survival functions, we perform the Peto & Peto test [133]s. The upper panel of Figure 4.4 shows the Kaplan-Estimates computed on Gab, grouped by outlet category. The test performed on its post and user lifetimes produces a p-value of 0.026 and 0.001, respectively. Therefore, we can conclude that the commenting persistence on Gab may be subjected to the outlet category of the post commented. Indeed, post lifetime on questionable posts reports a lower probability of being commented as time increases despite its longer persistence, reaching a maximum 340 days. Results from user lifetime estimation, instead, describe how users are more likely to comment on questionable posts for the first 240 days after post creation. After that time, the survival probability becomes higher on reliable posts.
In the end, we can conclude that the commenting behaviors on Gab reflect the general leaning of its community. Users are more likely to comment on questionable posts since their contents adhere to a common system of beliefs oriented to conspiracy theories. Coherently, the significant commenting persistence reported on reliable posts may describe the desire of users to express their dissent against the narratives introduced from such posts.

Next, we examine the commenting persistence on Twitter. Results from Peto & Peto test on the post and user lifetimes report a p-value equal to 0.011 and 0.0055 respectively, stating how the survival functions on both lifetimes differentiate with respect to the outlet category of the posts commented. Indeed, such estimations on Twitter describe a uniformity in the commenting behavior for the reliable category. This fact also provides further evidence about how the presence of content moderation can discourage users from expressing their views under posts whose authority is not verified.

In summary, Gab demonstrates how the lack of content policy helps the emergence of the narratives that characterize this environment, resulting in a discrepancy between the outlet categories with the most commenting persistence on the two lifetimes. However, when the content policy is applied, like on Twitter, such discrepancy dissolves, resulting in a commenting behavior that favors reliable content.

4.3.3 Quantifying Polarization

The presence of content moderation may affect how users develop homophily, i.e., the tendency to surround themselves with other peers who share the same narratives or system of beliefs. To quantify this phenomenon, we build a network in which the nodes represent the users i with their corresponding leaning x_i, while the edges represent the *following* relationship with other users that occurs on the social media. This representation allows us to measure the neighborhood leaning x_i^N, i.e., a measure of the characteristic leaning of the network surrounding user i. Figure 4.5 displays the joint distribution between the individual leaning of a user x_i and its corresponding neighborhood leaning x_i^N, on Twitter and Gab. In addition to this, the marginal probability distributions $P(x)$ and $P^N(x)$, referring to the individual and average neighborhood leaning, are represented on their corresponding axis. Lastly, the density of users at point (x, x^N) is represented as a contour map: the brighter the color in that point, the higher the user density. Results described in Figure 4.5(a) show the presence of homophily on Twitter, characterized by a strong correlation of leanings in correspondence of low values. The existence of a second echo chamber of incomparable size made of users with high individual leaning, and therefore not represented in the main figure but only visible in the marginal distributions, signals strong segregation between two communities. This finding also indicates how content regulations may affect the shape of the news diet of users in the context of the COVID-19 pandemic. Indeed, the concentration around small values for both leanings provides evidence about the effectiveness of the moderation imposed by the platform against posts and users that promote questionable content. On the other side, Gab shows a more heterogeneous behavior, as represented in Figure 4.5(b). Indeed, the joint distribution spreads over different values of the individual leaning domain, with the highest mode represented in correspondence of the point $(0.6, 0.6)$. We observe that on average users, regardless of their leaning, are surrounded by a neighborhood skewed towards questionable contents. Only very few users have a reliable-based leaning, who are also likely to be those with a weaker activity since they could be on Gab just for curiosity or dissing. Furthermore, the outlet category of the news that users post is not relevant anymore since the user's peers share a leaning with a high value. Finally, these findings may suggest that questionable news is employed to support the narrative of the environment, whilst reliable ones are only used to perform information operations by changing the original meaning of the posts through a comment.

4.4 Conclusions

In this chapter, we compared two social media, Twitter and Gab, to investigate the interplay between content regulation policies and news consumption. We provide quantitative measures of such differences by evaluating the engagement of users and posts. These measures are then extended by providing a categorization of news outlets. Next, we measure the commenting persistence of users and posts to describe their ability to express themselves under posts belonging to a specific outlet category. In the end, we characterize the presence of homophily, investigating how users with a specific leaning are more likely to surround themselves with users who share the same narratives.

Our results show how the application of content regulation, performed by Twitter, may be association with the contrast of fake news and conspiracy theories, shaping the news consumption and the polarization of users towards reliable content. The avoidance of these countermeasures, carried on by Gab, provides results that underline the presence of patterns possibly related to information operations. Indeed, users tend to engage with questionable and reliable content comparably. However, their commenting behavior and the assessment of the homophily in this environment describe a systematic affiliation towards questionable contents.

We conclude that content policies cover an important role against the circulation of harmful content, especially in the context of the COVID-19 pandemic. Our work provides meaningful evidence in this direction, indicating how a lack of content policy is associated with the emergence of harmful narratives that promote questionable content and mistrust everything that goes against them.

Our study presents some limitations. First, we only cover two social media platforms, the former representing a regulated ecosystem and the latter an unregulated one. Therefore, a generalization of our results to other social media platforms should be taken with caution. Moreover, the Twitter dataset was obtained through a sampling from the original seed made available by Chen et al. [124] as explained in Section 4.2.1, thus it only presents a partial view of the whole debate on the social media platform. Finally, Twitter and Gab may present significant differences in the demography of their users. However, data collection from social media, in general, involves pitfalls and biases on this perspective [134], without having the control to precisely define a set of participants with the same characteristics.

Future implementations of this study may then focus on extending the plethora of social media platforms involved, tracking a broader spectrum of the debate. A further focus on the different kinds of interactions on the platforms is needed, concerning the leaning of the post. Indeed, results from Glenski et al. [135] identified how Twitter users tend to interact with disinformation sources more often and faster than with trusted sources. Such implementation may also represent a good asset to account for the dissing/endorsement behavior promoted by users in segregated environments like Gab, analyzing those mechanisms from a textual perspective. Furthermore, a topological analysis of users who perform information operations in such environments may be relevant to understand their inner dynamics and promote specific countermeasures. Finally, the observed signs of an association between content regulation policies and misinformation reduction may be extended by including further moderated/unmoderated platforms and/or by observing the evolution of consumption dynamics within specific communities in the same social media.

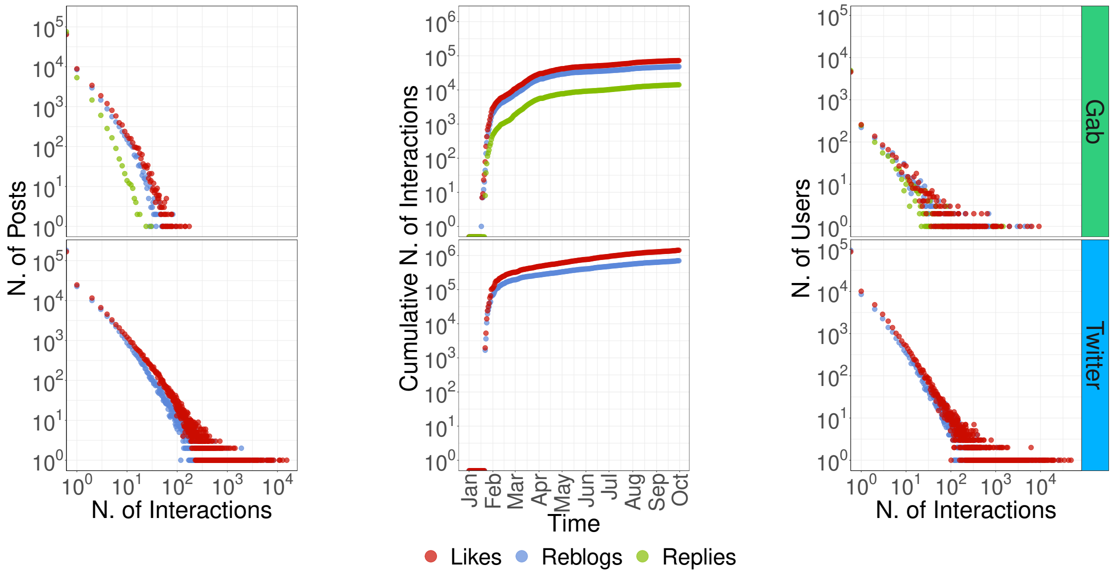


Figure 4.1. Representation of the engagement collected on Gab (upper panel) and Twitter (bottom panel). *Left column*: frequency distribution of the interactions for posts, defined as *Likes*, *Reblogs* (or *Retweets)* and *Replies.* A like is generally considered positive feedback on a news item. A reblog indicates a desire to spread a news item to friends. A reply can have multiple features and meanings and can generate collective debate. Both social media shows a heavy-tailed distribution that allows room for large deviations, i.e., some posts go viral. *Middle column*: evolution of the cumulative number of interactions over time. The general trend shows a rapid increase during February 2020, in parallel with the spreading of the COVID-19 outbreak. The absence of replies on Twitter is due to the limitations provided by their API. *Right column:* frequency distribution of interactions received by users. Similarly to posts, the distribution is heavy-tailed, describing how users tend to collect similar values of different interactions as their number increases.

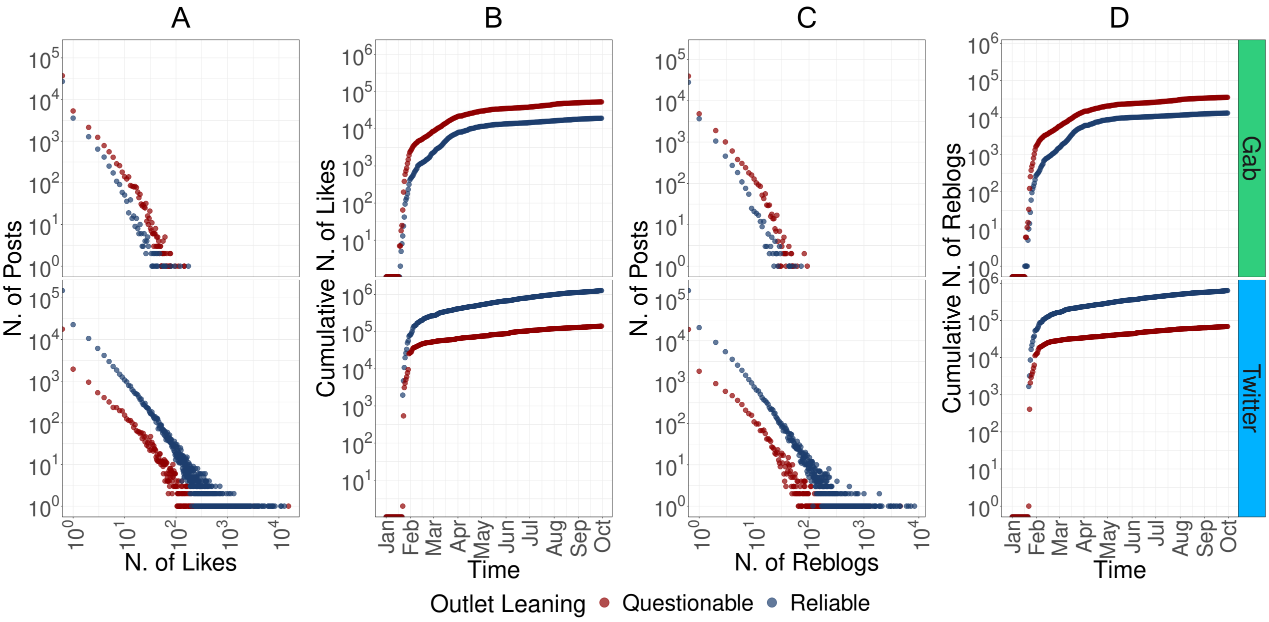


Figure 4.2. *Column A-B*: categorized distribution of the number of posts against the number of likes they received with its cumulative evolution. The distributions show some evidence about content preference on both platforms. Users on Gab show higher levels of engagement with questionable posts, supported by the lack of clear content regulation. Twitter, oppositely, shows strong evidence about the engagement with reliable content, with a remarkable gap between the two categories. From a cumulative perspective, the regulation imposed by Twitter is associated with an increasing divergence between questionable and reliable posts, showing how the latter category produces the highest engagement in the platform. The same does not apply to Gab, whose divergence seems not to increase during the analysis period. *Column C-D:* categorized distribution of the number of posts against the number of reblogs or retweets they received with its cumulative evolution. The previous considerations also apply to this kind of interaction, describing the willingness of users to inject the contents they support into the news feed of their followers.

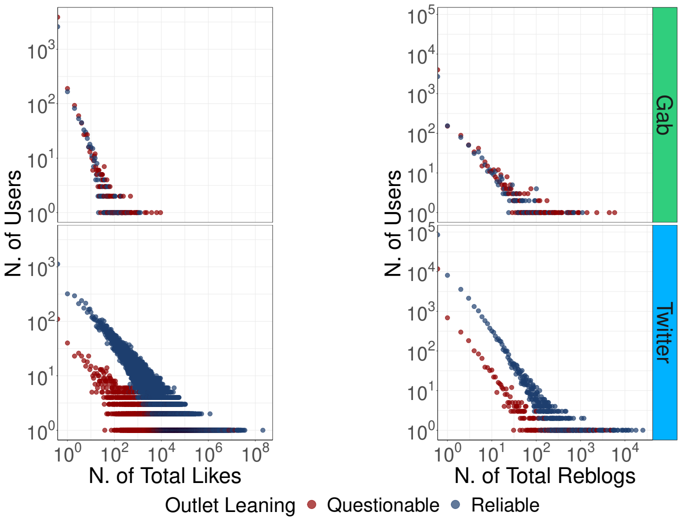


Figure 4.3. Distribution of likes (left column) and reblogs (right column) received by users posting Questionable or Reliable contents on Gab (upper panel) and Twitter (bottom panel). The figure shows how the presence of content regulations, performed by Twitter, results in a greater engagement with users who post reliable content. Gab, instead, shows a mixed endorsement pattern in which the engagement with users does not depend on the category of the content they post.

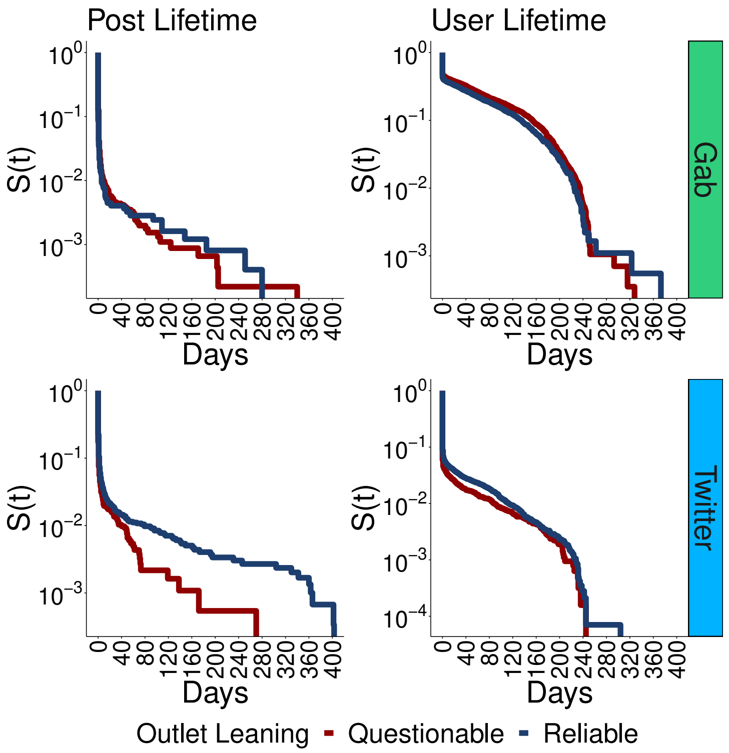
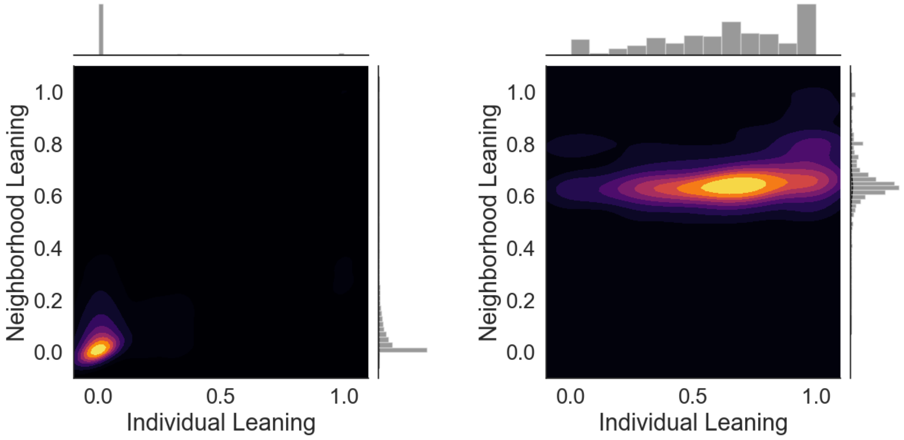


Figure 4.4. Kaplan-Meier estimates for Gab (upper panel) and Twitter (lower panel), grouped by outlet category.

Left column: estimates obtained through the computation of post lifetime, i.e., the period between the first and last comment a post received. *Right column*: estimates obtained through the computation of post lifetime, i.e., the period between the user's first and last comment.

Gab shows how the lack of content regulation is associated with a commenting behavior that underlines a preference towards questionable content. This behavior is characterized by a discrepancy between the outlet category with the highest commenting persistence both on user and post lifetimes. By contrast, the introduction of content policies from Twitter makes reliable content those with the highest commenting persistence, which does not depend on the lifetime perspective.



(a) Twitter

(b) Gab

Figure 4.5. Joint distribution between individual and average neighborhood leaning of all users posting classifiable contents at least three times on Twitter (left) and Gab (right). The figure shows further evidence about the regulation imposed by Twitter which is associated with the existence of a unique echo chamber of users with strong posting habits towards reliable content. Oppositely, Gab shows the presence of an echo chamber in which both individual and neighborhood leanings are concentrated around high values of the intervals, with a greater dispersion due to the mixed posting habits of users.

Chapter 5

Entropy and complexity unveil the landscape of memes evolution

In Chapter 4, we examined how content regulation and user interaction on specific social media platforms, Twitter and Gab, influenced the spread of information, especially concerning COVID-19. Indeed, we observed how the lack of moderation can affect the news consumption of the entire platform and, consequently, the language and narratives portrayed by users. Among the different elements users employ to express themselves, visual memes are an emerging aspect of the internet system of signification, and their structure evolves by adapting to a heterogeneous context. A fundamental question is whether they present culturally and temporally transcendent characteristics in their organizing principles. In this chapter, we study the evolution of 2 million visual memes from Reddit over ten years, from 2011 to 2020, regarding their statistical complexity and entropy. We find support for the hypothesis that memes are part of an emerging form of internet metalanguage: on one side, we observe exponential growth with a doubling time of approximately six months; on the other side, the complexity of meme contents increases, allowing and adapting to represent social trends and attitudes.

5.1 The advent of visual memes in the digital ecosystem

Social media radically changed the way we consume information and interact online [68, 136, 137]. Online interactions, indeed, influence social dynamics by favoring the formation of homophilic groups around shared narratives and attitudes and thus bursting group polarization [6, 3, 138]. In this scenario, multimedia content such as videos, photos, and pictures represents an essential portion of online communication, especially within social media platforms. Online communication can be read through the lenses of Dawkins' *cultural memes* [139], whose definition applies to almost all online information vehicles. Cultural memes represent a unit of cultural information transmitted and replicated; writing posts, sharing personal videos, expressing "likes" are examples of this concept. While Dawkins' model of cultural evolution is nowadays considered insufficient to comprehend the complex cultural phenomena of information transmission [140, 141, 142, 143], its evolutionary pattern still represents a valid basis for describing fundamental features of memes diffusion. In this chapter, we investigate the role and evolution of a particular kind of cultural meme, namely template images that undergo modifications or get some text overlapped, conventionally referred to just as *memes.* In the following, we adopt this convention. According to Dawkins' hypothesis, cultural memes [144] are characterized by the three essential elements of evolutionary theory: replication, variation, and selection. In the case of visual memes, the replication mechanism is self-evident. It consists of modifying an image, e.g., with some text, to represent a given situation. Moreover, replication of memes is facilitated by their consistency with other cultural memes present in the online environment, such as short videos, pictures, or short texts. Variation is an intrinsic feature of visual memes. Indeed, new memes are continuously created to target funny situations or jokes about political or societal events and compete for users' attention flowing across online communities. Finally, selection occurs when a meme cannot attract human attention nor adapt to transmit new contents and disappears adapting to the fast online environment. Among the online cultural memes that underwent relatively strong selection, we find, for example, blogs and discussion forums that have been replaced mainly by online social media; similarly, the emoji's introduction strongly reduced the use of *ascii art* symbols.

So far, a large body of research quantitatively investigated the features of different online cultural memes, not limited to images. Textual memes were analyzed by Leskovec et al. [145] as a proxy for the cycle of online news consumption. Ienco et al. [146] studied the problem of ranking memes, i.e., selecting those memes to be displayed to users to maximize the network activity on the platform. Romero et al. [147] studied online memes propagation in the form of Twitter hashtags. Bauckhage [148] investigated the epidemic dynamics of 150 famous memes, applying models from mathematical epidemiology to account for the growth and decline of visual memes. Ratkiewicz and coworkers [149] developed a framework for analyzing the diffusion of politics-related tweets. Weng et al. [150] studied meme virality through an agent-based approach, accounting for the limited attention each user can spend in online environments. In [151] the popularity of memes is correlated with the underlying network community structure. In [152] clustering techniques are applied to identify text-based memes, leveraging the content, metadata, and network structure of social data. Coscia [153] studied the popularity of memes leveraging measures of similarity between memes. In [154] tri-grams are used to cluster posts from Reddit. In [155] textual memes popularity is investigated looking at linguistic features. Adamic et al. [156] explored a large corpus of textual data from Facebook modeling the propagation of information as a Yule process. Dubey et al. [157] employed a Deep Learning architecture to process memes, extract the underlying template and explore its variations. An extensive analysis of visual memes is performed by Zanettou and coworkers [158] exploiting perceptual hashing to cluster visual memes together and explore the connections between the meme content and the communities in which it circulates. In [159] a deep-learning classifier for memes is proposed to explore the role of memes instead of non-meme images during elections. These investigations developed relevant insights and tools for handling and researching the world of internet memes. Nevertheless, little attention is given to the fundamental aspects of the evolution of memes in terms of visual features and conveyed information. Eventually, no evidence is reported for the hypothesis of internet memes constituting a metalanguage of the internet [142, 143].

In this chapter, we investigate the general evolution pattern of memes as an online communication artifact. To this aim, we leverage the evolutionist approach to define and measure the evolution rate, i.e., the number of new templates that appear online per time unit, the variation rate, i.e., the number of new instances of the same template that are produced in time; this quantity is particularly relevant concerning memes' popularity. As artistic expressions have been effectively investigated exploiting network science and physics concepts [160], we compute the trajectory of memes in the entropy-complexity plane. Specifically, these measures, grounded on the physics of complex systems, have been employed to investigate painting arts [161], revealing a temporal pattern towards higher complexity.

The basis of this investigation is a massive dataset of 2 million Reddit memes over ten years. Each image has been classified and ascribed to a template through a Machine Learning pipeline composed by an unsupervised Deep-learning based classifier followed by a density-based clustering algorithm (see 5.2). Our investigation shows that the memes ecosystem size is exponentially increasing, with a doubling time of approximately six months, indicating that replication is currently the leading process. Concerning selection, we observe that memes' persistence is dominated by rapid early adoption. The variation pattern is captured by the trajectory in the entropy–complexity plane. Similarly to what happens in painting arts, we observe a tendency towards structures with increasing visual complexity; early memes were made up of simple foreground images (e.g., animals or explicit human expressions) on plain backgrounds, while later ones involve more articulated scenes (e.g., modified movie frames).

As cultural signs, memes are strictly connected to the broader cultural system in which they are embedded. While their ultimate theoretical definition is still elusive and debated in terms of methodological frames, our results indicate that memes appear as one of the most productive and adaptable areas of digital communication, functioning as a metalanguage of cultural dynamics and evolving in progressive forms of textual complexity.

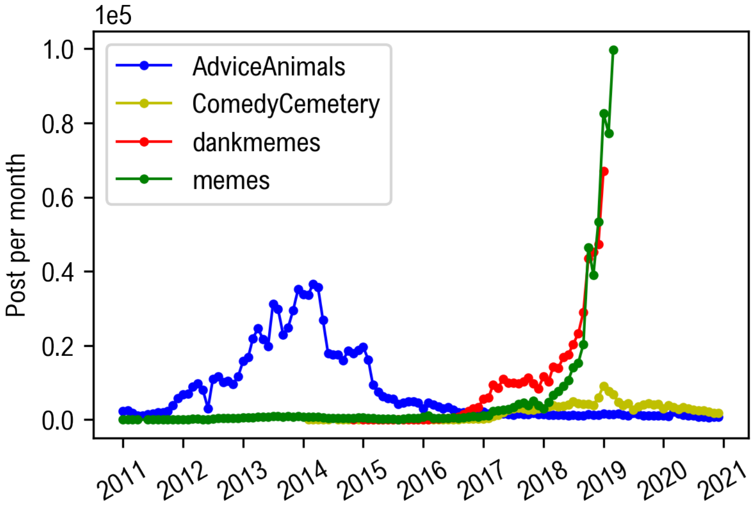


Figure 5.1. Dataset used in this study. The total amount of downloaded memes is about 2 million.

5.2 Methods

5.2.1 Data Breakdown

Reddit is an online social media platform that aggregates users in communities of interest. In the last years, it has been widely employed to perform academic research on online communities, and the number of active users on this platform is constantly increasing [162].

The visual memes used as dataset for this study were downloaded through the Pushshift Reddit Dataset, selecting four communities (subreddits) explicitly devoted to share and discuss about memes, namely: *r/AdviceAnimals*, *r/memes*, *r/CemeteryComedy* and *r/dankmemes*. Data were collected considering a ten years window, from 2011 to 2020. In Figure 5.1 the number of downloaded posts per each community is reported, as function of time. Not all the communities started their activity simultaneously.

5.2.2 Clustering

One of the main features of visual memes is their recurrent nature: starting from an initial template, memes are produced through text or image modifications resulting each time in a new instance that stems from the original template. To measure the evolution and variation rate of visual memes, it is crucial to cluster them according to the underlying template. As the collected number of memes is about two million, such a large amount of images calls for automatic classification methods.

Our unsupervised clustering procedure is divided into two steps: first, we apply, to our knowledge, the state of the art Deep Learning implementation for unsupervised image clustering, that is called SCAN [163] (Semantic Clustering by Adopting Nearest neighbors), followed by a further clustering procedure through the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm [164].

Specifically, the SCAN algorithm works in two steps. In the first one, self-supervised learning is used to train a neural network with parameters ϑ that maps images ($x_i, i = 1 \dots N$) into feature representations $\phi_\vartheta(x_i) \in \mathbb{R}^d$. Therefore, each image is represented by a vector of dimension d, with d being the dimension of the embedding space carrying semantically meaningful information about its content. The parameters ϑ are determined by minimizing the loss function given by the distance δ between the representation of the image and the representation of their augmentations:

$$\min_{\vartheta} \delta(\phi_\vartheta(x_i), \phi_\vartheta(T[x_i]))\,.$$

The augmentation of an image may be a rotation, an affine or perspective transformation, etc.

In the second step, a neural network with parameters η is used to classify an image x_i and its nearest neighbors sampled exploiting its corresponding representation $\phi_\vartheta(x_i)$. As the second task is a classification, the output is a probability distribution over the considered classes for the given image: $\phi_\eta(x) \in [0,1]^C$, where C is the number of clusters considered.

The loss function, in this case, is made of two terms:

where $\mathcal{D}$ is the dataset comprising all images, $\mathcal{N}_x$ is the set of nearest neighbors of image x, $\mathcal{C}$ is the set of clusters, and $\langle \cdot \rangle$ is the dot product. The first term aims to maximize the probability that the image and its nearest neighbors are classified in the same class. The second term avoids the formation of a single cluster containing all the images and forces the spread of the predictions uniformly across the clusters. Eventually, for each processed image, we get a representation vector of size $d = 2048$ and a set to which it belongs.

The SCAN algorithm provides us with an informative and compact representation for each image in our dataset together with a first, high-level clustering. SCAN divides the corpus into four clusters that group the visual memes into three broad and general categories: animals (two sets), humans, and others. The two clusters with animal images have been merged together. In other words, memes containing humans represent the 50% of the corpus, followed by 25% of animals and 25% of other kind of contents.

To obtain a template-based clustering, we exploited HDBSCAN [164]. For each high-level cluster obtained from SCAN, the entire corpus of memes can be represented by a matrix whose rows are the representations $\phi_\vartheta(x_i)$. To make the problem computationally more tractable, Principal Component Analysis was used to reduce the features' dimension from 2048 to 20. This allowed us to better fit our computational resources and did not cause any reduction in the quality of the clustering. By applying HDBSCAN to such a matrix, we were able to get a label for each image of our corpus and to separate the memes by their template. Notably, HDBSCAN can separate clusters from noisy points. Part of the corpus does not belong to any template, and therefore is marked as noise and grouped in a large "cluster of noise". Despite this suitable property of the algorithm, some clusters may result made up of images whose template is not the same for all. A purity measurement is therefore required to exclude from the analysis too heterogeneous clusters, i.e. clusters below a given purity threshold.

In a completely unsupervised framework, the quality of clustering is in general not easy to evaluate [165]. A quantity that is usually employed as an objective function for clustering is the average-pairwise distance $\overline{S}_k$, which evaluates the intra-cluster homogeneity [166], i.e. how much each element of the cluster is, on average, similar to all the others. Its definition is given by

$$\overline{S}_k = \frac{1}{N_k^2} \sum_{x_i, x_j \in \mathcal{C}_k}^{N_k} ||\phi_\vartheta(x_i) - \phi_\vartheta(x_j)||^2 \,.$$

In our case, we employed the latter quantity to measure the purity of the clusters identified by HDBSCAN, together with the cluster size. In Figure 5.2(a) for each cluster is reported $\overline{S}_k$ and the cluster size. The red dots are the clusters

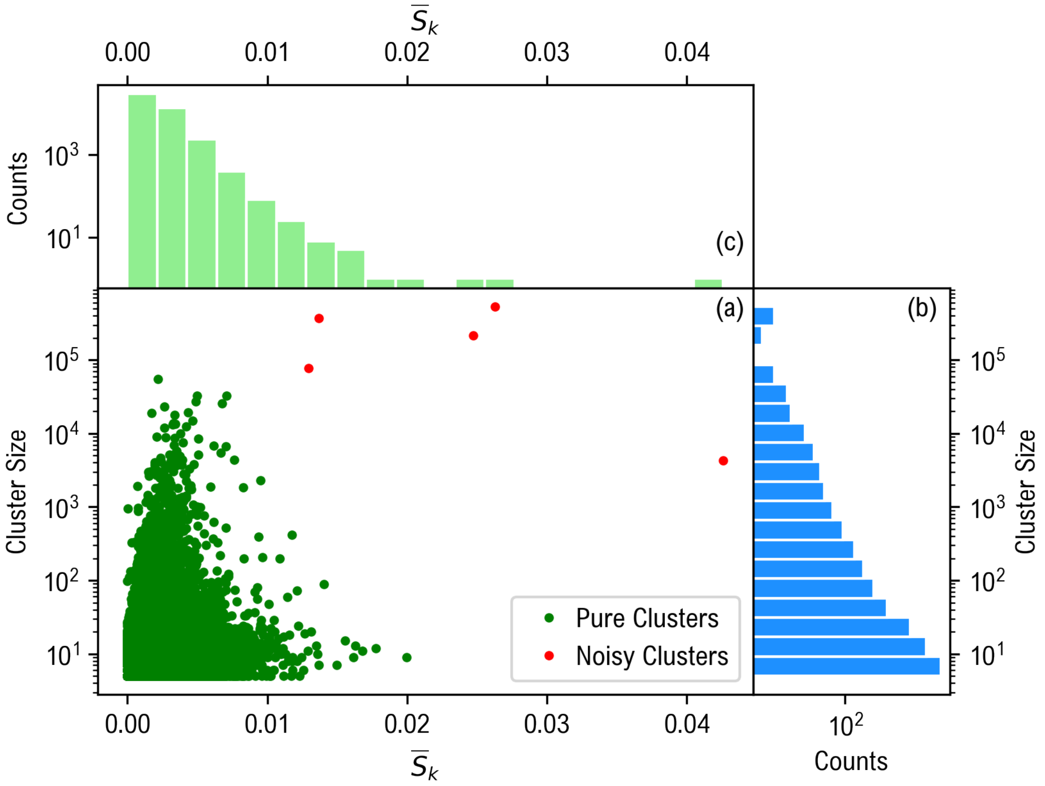


Figure 5.2. Panel (a): joint distribution of average-pairwise distance $\overline{S}$ and cluster size. Red dots represent "noisy clusters" identified by HDBSCAN and outliers with respect to the cluster size and $\overline{S}_k$ distributions, shown respectively in panels (b) and (c).

that are not considered for the following analysis. Four of them correspond to the "noisy clusters" retrieved by HDBSCAN. All of them result in outliers with respect to the joint size and $\overline{S}_k$ distribution, whose marginals distributions are reported in panels (b) and (c) respectively.

In Figure 5.3 some memes from a pure cluster and a noisy one are reported.

5.2.3 Entropy and complexity

Permutation entropy H and statistical complexity C are two quantities that can be used to synthesize general properties of images, based on the value and relative disposition of their pixels. In the following, we give a minimal description of both quantities, and we refer the reader to original articles for

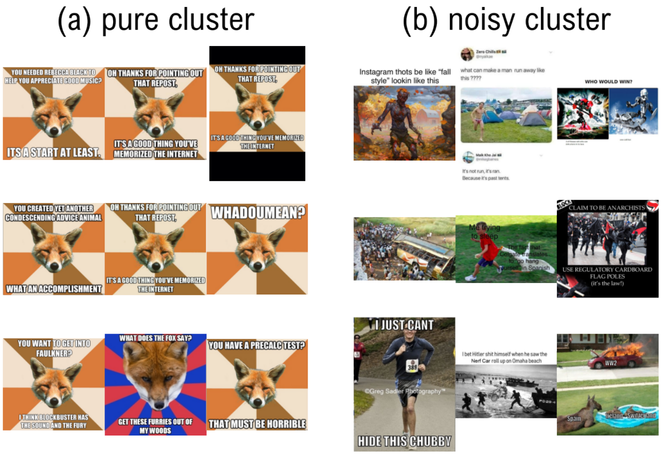


Figure 5.3. Examples of memes clusters. In a pure cluster, shown in panel (a), memes are different instances of the same template; insead, in a noisy cluster, shown in panel (b), it is not possible to detect a common template for all instances.

more formal details [167, 168, 169, 170, 171]. Permutational entropy measures the degree of disorder in the pixel arrangement. High values indicate high pixel randomness, while low values correspond to more regular patterns. Statistical complexity instead measures the amount of "structural" complexity. Non-trivial spatial patterns give rise to positive values, while extremely ordered or disordered patterns correspond to low values.

To compute H and C all colored images were converted to grayscale. Thus, each image consists of two-dimensional matrix. Next, following [161], for all the 2×2 submatrices comprised in the image the relative ordering of the pixel is computed. For a collection of four elements, a number of $4! = 24$ possible permutations can be obtained. By counting the relative occurrence of each permutation, a probability mass function

$$P = \{p_1 \dots p_N\} \qquad \text{with} \qquad \sum_{i=1}^{N} p_i = 1\,,$$

can be built. Shannon's entropy is then computed over this probability distribution to obtain the permutation entropy

$$H(P) = \frac{S(P)}{\log(N)} = \frac{1}{\log(N)} \sum_{i=1}^{N} p_i \log\left(\frac{1}{p_i}\right)\,. \tag{5.1}$$

Given the probability mass function P, its discrepancy with respect to a uniform distribution U

$$U = \{u_1 \dots u_N\} \qquad \text{with} \qquad \sum_{i=1}^{N} u_i = 1\,,$$

is obtained by computing the Jensen-Shannon divergence

$$D(P,U) = S\left(\frac{P+U}{2}\right) - \frac{S(P)}{2} - \frac{S(U)}{2}\,.$$

Combining this quantity with $H(P)$, the statistical complexity can be computed as

$$C(P) = \frac{D(P,U)\,H(P)}{D^*}\,, \tag{5.2}$$

with the normalizing factor

$$D^* = \max_P D(P,U) = -\frac{1}{2}\left[\frac{N+1}{N}\log(N+1) + \log(N) - 2\log(2N)\right]\,.$$

The evolution of visual patterns can be studied as a trajectory in the above defined entropy-complexity plane [168, 169], following the same approach used for paintings [161].

5.3 Results and Discussion

Our study starts from Dawkins' hypothesis of the meme as the basic unit of cultural evolution, in connection with the post-memetics analyses of memes as cultural signs. We studied the evolution of Internet visual memes, i.e. images with (typically) overlapped text strings over a time span of 10 years. The dataset comprises around 2 million images, that were grouped together exploiting an unsupervised Machine Learning routine (see Section 5.2). Such an extended dataset enables us to investigate some properties of this particular cultural meme. The clusters retrieved by our procedure correspond to the *templates* of the various memes. In the following, we refer to "meme" as for the template, while each image belonging to a given template is an "instance" of the meme.

To quantify the growth of memes adoption we computed their evolution rate. For each cluster, we store the creation time of its first instance. Next, per each week of sampling, we compute the number of new templates. The result is

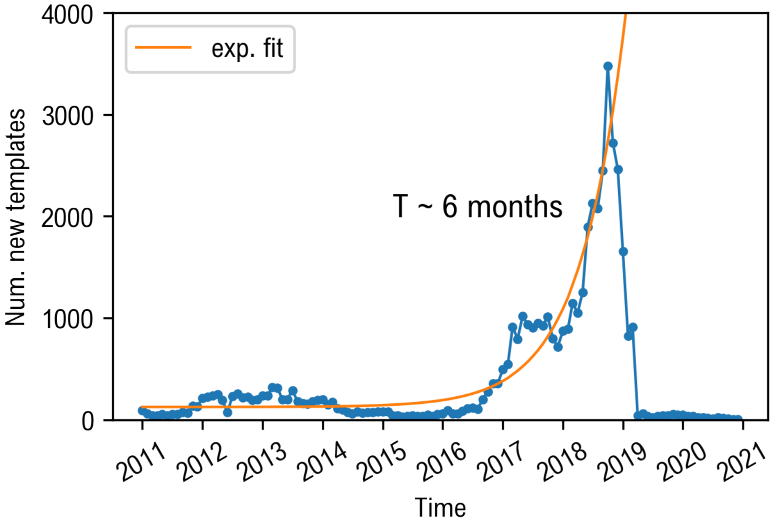


Figure 5.4. Evolution rate of internet memes. The number of new templates per each month is reported as a function of time. The growth rate is estimated through an exponential fit, with a doubling time of $T \sim 6$ months.

reported in Figure 5.4, in which the growth rate is estimated by an exponential fit, giving a doubling time for the number of templates $T \sim 6$ months. The sudden drop in the plot is a finite size effect of the dataset. Namely, for subreddits `r/memes` and `r/dankmemes` it has not been possible to download the data over 2019, due to the exponential trend in the number of memes (see Figure 5.1). The fit was performed considering the data until January 2019.

Another relevant quantity in terms of cultural evolution is the *mutation rate*. The mutation rate can be approximated by looking at the instances of each meme. For each template we computed the distribution of the differences (Δt) between the creation times of an instance and the following one. This distribution reveals the nature of growth of each cluster: a distribution skewed towards low values of Δt corresponds to a very fast and bursty growth dynamic. Conversely, larger values of Δt may reveal a more persistent template, whose instances occur more spaced in time. In Figure 5.5 the distribution of the "inter-instance" times (blue histograms, left column) is reported together with the clusters lifetime distribution (orange histograms, right column) , i.e. the time interval between the first and the last instance of a given meme. These distributions are computed for different typical cluster sizes (indicated as CS).

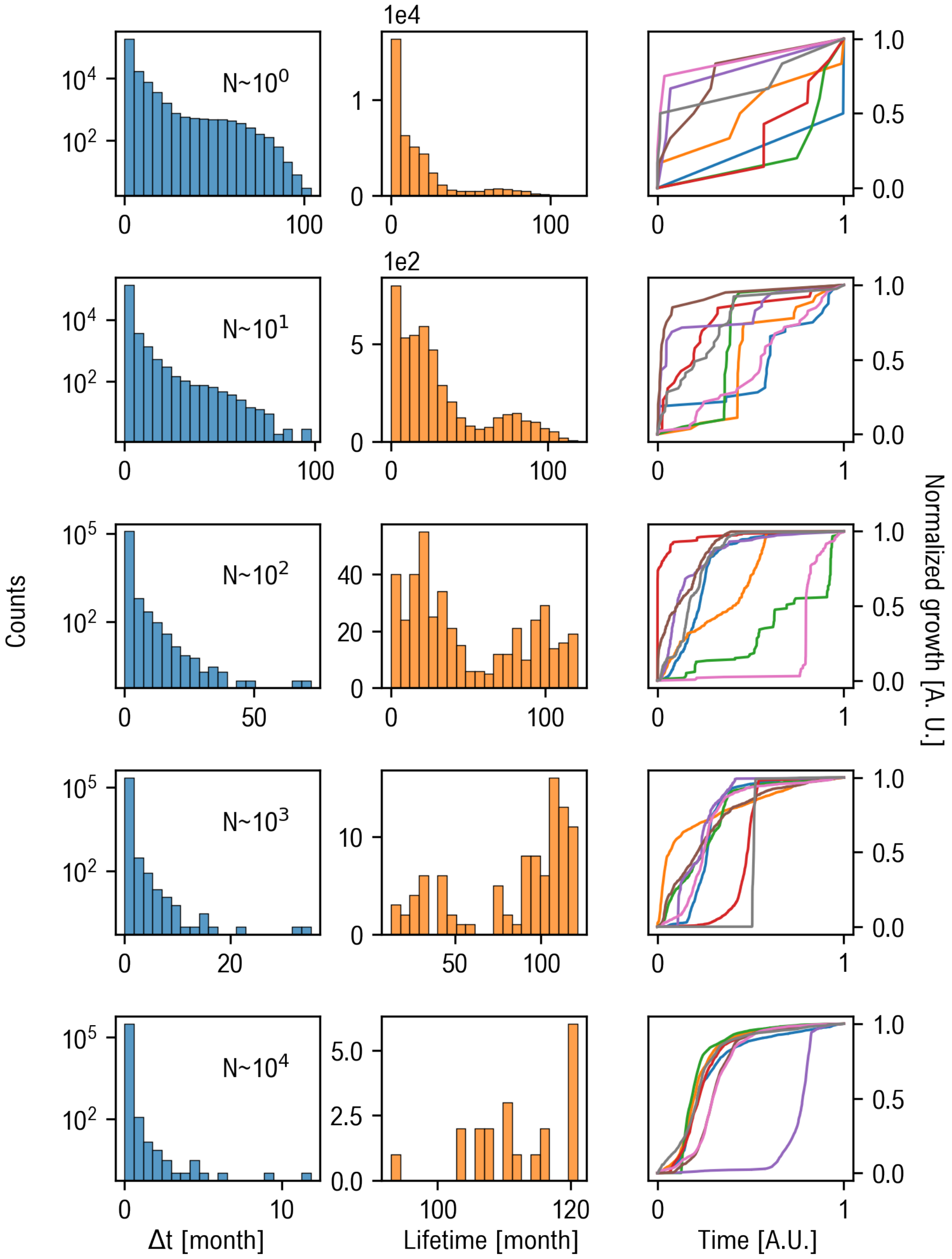


Figure 5.5. Mutation rate of memes. Left column: distribution of instances' inter-arrival times (Δt); central column: lifetime distribution of memes; right column: exemplary growth curve of meme adoption. Each row correspond to a typical size of meme cluster.

Overall, lifetime results positively correlated with cluster size. Small clusters show a heterogeneous distribution of instances' inter-arrival times, comprising both small-bursty clusters and small-slowly paced ones. The lifetime is peaked towards low values. This behaviour is also shown by very recent clusters whose actual size cannot be estimated within our dataset, as their evolution is ongoing. As the cluster size grows, we observe a shift of the inter-times distribution towards low values, unveiling faster dynamics, while the lifetime distribution is concentrated around larger values. By looking the corresponding growth curves we observe an ensemble of trajectories that tend to display a fast initial build up of popularity, followed by a slower diffusion that determines the longer lifetime values. Conversely there are also examples of memes that takes more time to reach a wide popularity. This aspect may be due to non-trivial popularity dynamics, calling for further research.

Following [161], we investigated the evolution of memes in the entropy-complexity plane. For each meme instance we computed the values of $H(P)$ and $C(P)$ and then averaged the obtained values by year. The results are reported, for each subreddit, in Figure 5.6. We observe that each community moves towards higher complexity values, except for `r/AdviceAnimals`, whose posting rules limit the natural evolution of produced memes; this effect could be also linked to the overall decrease in memes production observed for this community. Interestingly, also paintings followed a similar trajectory in the entropy-complexity plane: quite localized along the entropy axis, but shifting towards higher complexity in time (see. Figure 1 of ref. [161]).

The tendency of memes to evolve towards more complex structures can be explained considering this object as part of the emerging internet meta-language. In fact, memes are used to quickly vehicle context-specific content, which in turn evolves towards more and more specific templates. This may lead to a segregation effect, with a specific dialect depending on the community in which a meme is shared. In fact a meme created for a specific community, e.g. gaming community, does not have to be universally comprehensible across the web. This aspect leads to the use of more complex and specific patterns.

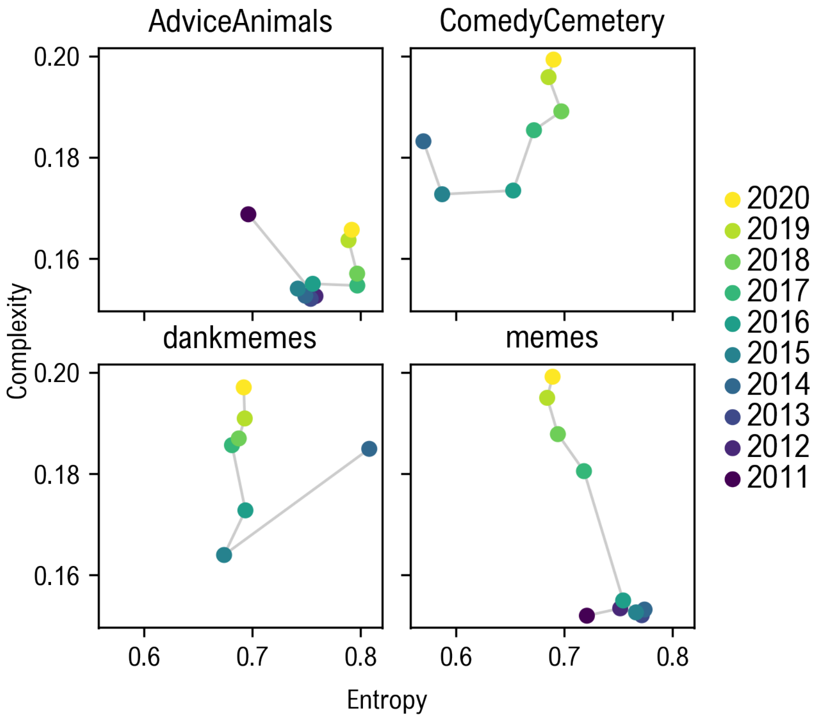


Figure 5.6. Trajectories in entropy-complexity plane for the four Reddit communities. All, except `r/AdviceAnimals` present an evolution towards higher values of complexity that resembles that of painting arts (see [161]). Each dot is the average value of entropy and complexity for each year.

5.4 Conclusion

The Internet provides an environment in which information quickly spreads and adapts to comply with users' cognitive abilities. A foundational question about memes is whether they present culturally and temporally transcendent characteristics in their organizing principles and how they evolve. Such a significant increase and spread of visual memes can be read under the light of post-memetics theories. Visual memes are favored by the rapid, fluid, continuously changing internet environment because of their simplicity, ease of handling and broad applicability in terms of subjects and situations. We find support for the hypothesis that memes are part of an emerging form of internet metalanguage: on one side, we observe an exponential growth with a doubling time around 6 months; on the other side, the complexity of memes contents increases, allowing to timely represent social trends and attitudes. Our analysis shows that memes are relational entities functioning as flexible elements of a metalanguage that de-codifies and re-codifies the cultural system. They appear as fundamental components of an organic process that affects and conditions the digital environment and produces evolving forms of visual and textual complexity.

Chapter 6

Characterizing Engagement Dynamics across Topics on Facebook

In Chapter 5, we explored the evolution of visual memes on the Internet, observing their increasing adoption as a new language in online conversations. Recommendation algorithms employed by social media platforms potentially contribute to the dissemination of new content and, therefore, of new languages by suggesting content targeted users consider of interest. This ability of platforms to expose users to new content has been playing a crucial role in shaping the popularity of various topics. However, an observed drawback of this feature is that, as the suggested content becomes viral, it aggregates users with opposing systems of beliefs, potentially sparking heated discussions that can increase user polarization. Therefore, understanding the dynamics that regulate the interplay between topic virality and the level of controversial reactions they can produce is crucial for creating safer digital environments. To fill this gap, in this Chapter, we explore the interplay between the virality of controversial topics and how they may trigger heated discussions and eventually increase users' polarization. We perform a quantitative analysis on Facebook by collecting $\sim 57M$ posts from $\sim 2M$ pages and groups between 2018 and 2022, focusing on engaging topics involving scandals, tragedies, and social and political issues. Using logistic functions, we quantitatively assess the evolution of these topics, finding similar patterns in their engagement dynamics. Finally, we show that initial burstiness may predict the rise of users' future adverse reactions regardless of the discussed topic.

6.1 The attention economy in social media era

The advent of social media platforms changed how users consume information online [172, 173, 174, 175]. The micro-blogging features on Twitter and Facebook, combined with a direct interaction between news producers and consumers, have remarkably affected how people get informed, shape their own opinions, and debate with other peers online [176, 177, 178]. Over the years, following the business model of social media platforms, news outlets and producers attempted to maximize the time spent by users on their contents [179, 180], giving birth to the concept of *attention economy* [181]. The term refers to the users' limited capability and time to process all information they interact with [182, 183, 184]. The transition toward a news ecosystem shaped on social media platforms unveiled patterns in information consumption at multiple scales [43, 68], which contributed to the emergence of the polarization phenomenon and the formation of like-minded groups called echo chambers [185, 3, 186]. Within echo chambers, characterized by homophily in the interaction network and bias in information diffusion towards like-minded peers, selective exposure [187] is a significant driver for news consumption [3]. The combination of echo chambers and selective exposure makes users more likely to ignore dissenting information [9], choosing to interact with narratives adhering to their point of view [68, 188].

Several studies explored the existence of these mechanisms in many topics concerning political elections, public health, climate change, and trustworthiness of the news sources [68, 188, 189, 24, 190, 191, 192, 72, 19, 193]. Findings indicate neither the topic nor the quality of information explains the users' opinion-formation process. Instead, several studies observed how the virality of discussions can increase the likelihood of inducing polarization, hate speech, and toxic behaviors [194, 17, 195], highlighting how recommendation algorithms may have a role in shaping the news diet of users.

Therefore, it is necessary to provide a better understanding of how user interest evolves in online debates. To achieve this goal, in this chapter we provide a quantitative assessment of the dynamics underlying user interest in news articles about different topics. We analyze the engagement patterns produced by $\sim 57M$ posts on Facebook related to ~ 300 topics, involving a total of $\sim 2M$ posting pages and groups over a period that ranges from 2018 to 2022. We first provide a quantitative assessment of topics' attention through time, extracting insightful parameters from their engagement evolution. Then, we construct a metric called the Love-Hate Score to estimate the level of controversy associated with a topic using the sentiment of users' engagement, as expressed by the normalized difference between their positive and negative reactions. Our results show that topics are generally characterized by an interest that constantly increases since the appearance of the first post. We find that topics' interactions grow with permanent intensity, even for prolonged periods, indicating how interest is a cumulative process that takes time. We statistically validate this result by comparing parameters across topic categories, discovering no differences in the evolution of the engagement. Indeed, regardless of their category, topics keep users engaged steadily over time, and their lifetime progression seems thus unrelated to its thematic field. Finally, we find that topics with sudden virality tend to occur with more controversial and heterogeneous interactions. In turn, topics with a steady evolution exhibit more positive and homogeneous reaction types. This difference in the sentiment of reactions, and the protracted duration of topics' lifetime, are both upshots consistent with the emergence of selective exposure as a driver of news consumption.

6.2 Materials and Methods

This section describes the data collection process, the topic extraction process, the models and the metrics employed in assessing collective attention.

6.2.1 Overview of the data collection process

The data collection process comprises several parts, as described in Figure 6.1. We start by creating a sample of news articles from the GDELT

Event Database [38]. Then, we process the articles' text to obtain a set of representing terms. Consequently, we apply the Louvain community detection algorithm[196] on the bipartite projection of the co-occurrence term network to identify the topics of interest. The terms representing these topics will serve as input for collecting posts from Facebook.
The data collection and analysis process are compliant with the terms and conditions [197] imposed by CrowdTangle [55]. Therefore, the results described in this chapter cannot be exploited to infer the identity of the accounts involved.

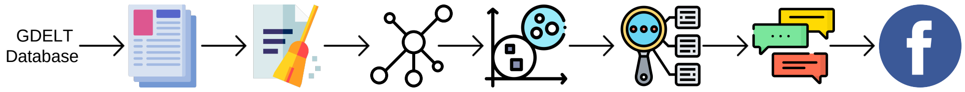


Figure 6.1. Summary of the analysis workflow followed in the current study. News articles are collected from the GDELT Database, and their corpus is extracted, cleaned and analyzed to retrieve the most representing terms. The bipartite projection of the co-occurrence network built upon these terms serves as an input for the Louvain community detection algorithm to identify keyword clusters. Independent labellers then analyze these clusters to identify the subset of words that represent the topic under consideration, which are then used on CrowdTangle to retrieve the Facebook posts relating to those events.

News Extraction from GDELT

The GDELT (Global Database of Events, Language, and Tone) Project [198], powered by Google Jigsaw, is a database of global human society which monitors the world's broadcast, print, and web news from nearly every corner of every country in more than 100 languages. It identifies the people, locations, organisations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day [199]. We gathered news articles from the GDELT 2.0 Event Database [38], which can store new world's breaking events every 15 minutes and translates the corresponding news articles in 65 languages, representing 98.4% of its daily non-English monitoring volume [38]. The analysis covers a period between 1/1/2018 and 13/5/2022, collecting 50 news articles each week for a total of $\sim 79K$.

Extracting representative keywords from news articles

To clean and extract the most representative keywords of each news article, we employed the *newspaper3k* Python package [200]. We initially extracted words from the body of the article, excluding stopwords and numbers. Then, we computed the word frequency $f(w, i)$ for each word w in article i. Finally, we sorted words in descending order according to their frequency, keeping the top 10 most frequent words.

Topic Extraction from News Article's Keywords

The list of terms with the corresponding news articles can be formalised as a bipartite graph $G = (T, A, E)$ whose partitions T and A represent the set of terms $t \in T$ and the articles $a \in A$ respectively, for which an edge $(t, a) \in E$ exists if a term t is present in an article a. By projecting graph G on its terms T we obtain an undirected graph P made up of nodes $t \in T$, which are connected if they share at least one news article.
We perform community detection on the nodes of P by employing the Louvain algorithm [196]. As a result, we obtain a set of clusters C, where each cluster $c \in C$ contains a list of keywords that are assumed to be semantically related to a topic. We then asked a pool of three human labellers to select, for each community, from two to three terms they considered the most representative to identify a topic unambiguously.

Data collection of Facebook posts

The news articles obtained from the GDELT Event Database do not contain information helpful in estimating the attention they generate online. To include the dimension of user engagement, we employ each topic's set of representative terms to collect Facebook data over a period that goes from 01/01/2018 to 05/05/2022. The data was obtained using CrowdTangle [55], a Facebook-owned tool that tracks interactions on public content from Facebook pages, groups, and verified profiles. CrowdTangle does not include paid ads unless those ads began as organic, non-paid posts that were subsequently "boosted" using Facebook's advertising tools. CrowdTangle also does not store data regarding the activity of private accounts or posts made visible only to specific groups of followers.

The collection process produced a total of $\sim 57M$ posts from $\sim 2M$ unique pages and groups, generating $\sim 8B$ interactions. The result of the data collection process is described in Table 6.1.

Total News Articles from GDELT	Total Posts from Facebook	Total Interactions	Total Groups and Pages	Number of Topics Collected	Period
79 650	57 031 026	8 015 177 602	2 224 430	296	1/1/2018 - 5/5/2022

Table 6.1. Data Breakdown of the study, including the total amount of news articles and posts collected from GDELT and Facebook respectively, together with the number of topics and the analysis period.

Topic Categorization

To provide a correspondence between topics and their area of interest, we performed a categorization activity under the following labels: Art-Culture-Sport (ACS), Economy, Environment, Health, Human Rights, Labor, Politics, Religion, Social and Tech-Science. Three human labellers carried out the activity to connect topics and categories, choosing as the representative only those categories selected by at least two of the three labellers.

6.2.2 Metrics

We begin by describing a measure for fitting the cumulative engagement evolution. Then, based on the previous step, we outline an index to evaluate the sharpness of the topic's diffusion. Finally, using Facebook's reactions, we introduce a sentiment score to assess the topic's controversy. A topic-aggregated version of the dataset containing all the metrics defined in this section can be found in the Data Breakdown Section of SI.

Fitting cumulative engagement evolution

The study of the diffusion of new ideas has been carried on through the years, starting from the Bass diffusion model [201] and then extended to a multitude of topics[202, 203, 204, 205, 206, 207, 208], indicating the relevance of s-curves in the analysis of innovation spreading. Therefore, to model the evolution of the engagement received by posts, we fit the cumulative distribution of the overall engagement (i.e., the number of likes, shares and comments) over time employing a function $f_{\alpha,\beta}(t)$, with $\alpha, \beta \in \mathbb{R}$, defined as

$$f_{\alpha,\beta}(t) = \frac{1}{1 + e^{-\alpha(t-\beta)}}. \tag{6.1}$$

From a mathematical point of view, Equation 6.1 defines a general sigmoid function that depends on the parameters α and β. The α parameter represents the slope of the function, describing the steepness of the engagement evolution. On the other hand, β is the point at which the function reaches the value 0.5 and quantifies the time required for a topic to reach half its total interactions.

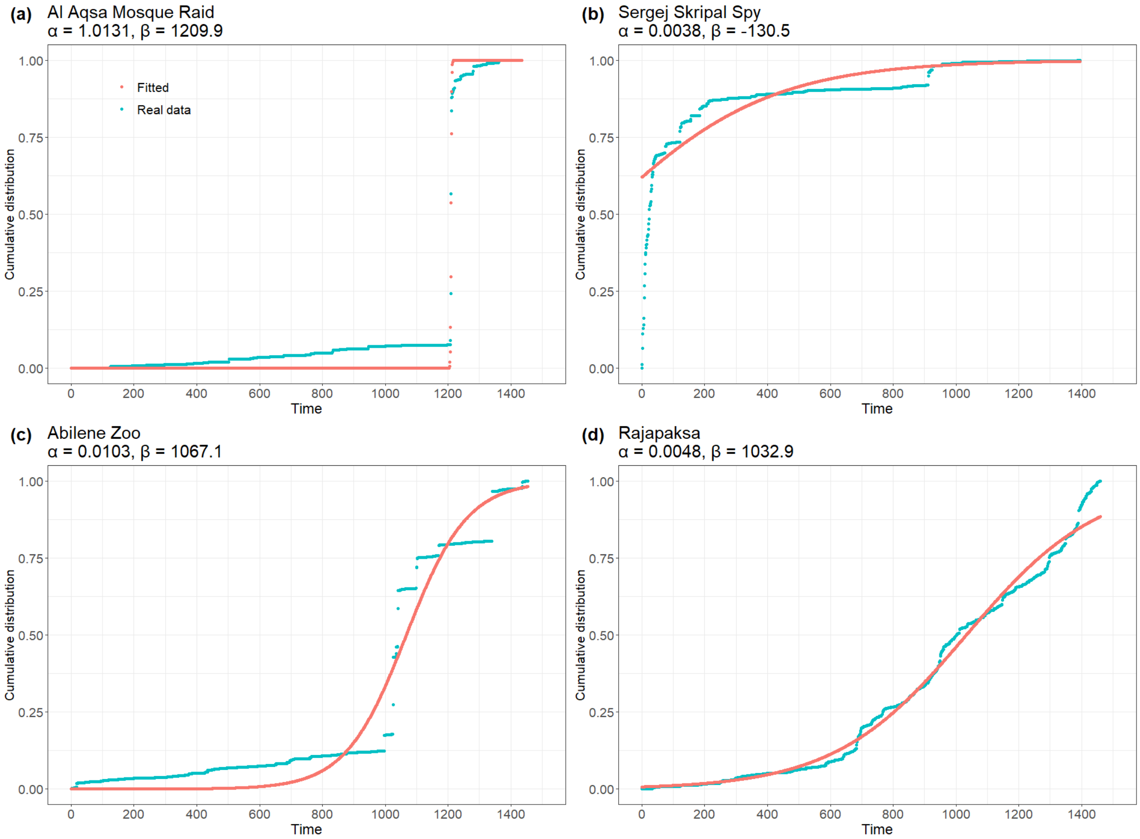


Figure 6.2. Representation of a sample of four topics employing their normalized cumulative evolution of engagements and fittings. The incidence of the α parameter can be observed in the sharpness of the fitting curves. The β parameter instead regulates the shift of the function through the x axis: the higher its value, the higher the delay from t_0 where the sigmoid produces its increment.

To provide a representation of the impact that α and β can have in topic engagement evolution, Figure 6.2 displays four topics with peculiar configurations. Figure 6.2a shows a sigmoid in which the high values of α and β produce a sharp increment relatively far from t_0. Such behaviour corresponds to those topics that require some time before gaining maximum diffusion with the public. Figure 6.2b instead provides a fit where the sigmoid produces low values for α and β, resulting in a smoother increment in the proximity of t_0 than the one described in Figure 6.2a. Finally, Figure 6.2c and 6.2d provide an example of how two curves that share similar values of β parameters can have a different evolution of their increase by slightly modifying the values for α parameter.

Speed Index

To provide a measure of how quickly the attention towards a topic reaches its saturation, we define a measure called the Speed Index $SI(f_{\alpha,\beta})$ as

$$SI(f_{\alpha,\beta}) = \frac{\int_0^T f_{\alpha,\beta}(t)dt}{T}. \tag{6.2}$$

The SI considers the joint contribution of α and β parameters, where T represents the time of the last observed value for $f_{\alpha,\beta}(t)$. Note that the SI is the mean integral value of $f_{\alpha,\beta}$, i.e. the normalised area under the curve of $f_{\alpha,\beta}$ (therefore $SI(f_{\alpha,\beta}) \in [0, 1]$). The assumption in the definition of this function relies on the fact that high-speed values are obtained by sigmoids that reach the plateau in a short time, as the behaviour represented in Figure 6.2b.

Love-Hate Score

To quantify the level of controversy that a Facebook post may produce, we define a measure called the Love-Hate (LH) Score. In line with previous works that quantified controversy from post reactions [209, 210], we define the LH Score $LH(i) \in [-1, 1]$ as

$$LH(i) = \frac{l_i - h_i}{l_i + h_i}, \tag{6.3}$$

where h_i and l_i are respectively the total number of *Angry* and *Love* reactions collected by a post i. A value of LH equal to -1 indicates that the post received only *Angry* reactions from the users, while a value equal to 1 indicates that the post received only *Love* reactions. Therefore, a value close to 0 reflects the presence of controversy on a post due to a balance of positive and negative reactions.

6.3 Results and Discussion

6.3.1 Quantifying topic engagement evolution

We first provide a quantitative assessment of the the evolution of engagement with topics on social media. To do so, we perform a Non-linear Least Squares

(NLS) regression by fitting the sigmoid function $f_{\alpha,\beta}(t)$ to the cumulative engagement gained by each topic.

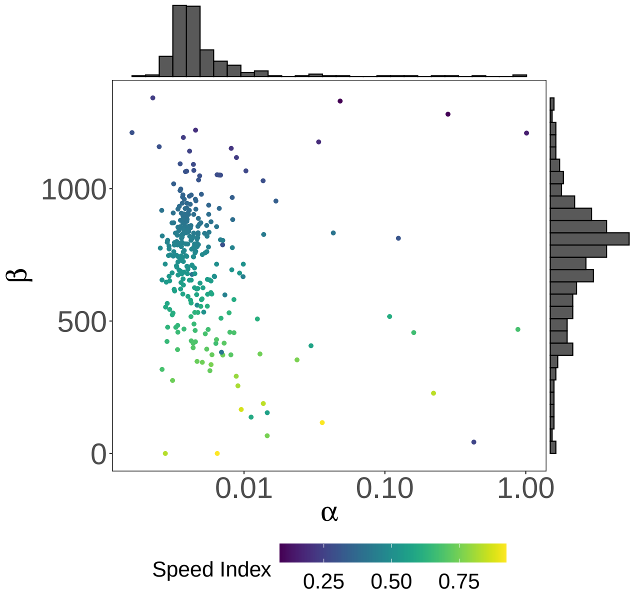


Figure 6.3. Joint distribution of α and β parameters obtained from the NLS regression for each topic. We observe that topics are generally characterized by values of α and β, which explains how user interest in a topic does not increase all of a sudden but is the result of a process that evolves over time.

The distribution of the α parameter provided in Figure 6.3 describes how the majority of topics have a value of α belonging to the $[0, 0.0047]$ interval. This result demonstrates how user interest in a topic does not suddenly increase but results from a long-term process. Instead, the distribution of the β parameter describes a prevalence of topics in the $[600, 1000]$ interval, identifying the tendency of topics to become a matter of interest with some delay w.r.t the first post covering them.

6.3.2 Evaluating the relationship between topic engagement and controversy

To quantify the interplay between users' interest in a topic and the associated level of controversy, we compute the Spearman correlation between the Speed Index and the LH Score for each topic. Results from the upper panel of

Figure 6.4 show a general negative tendency of users to react with a negative sentiment when a topic gains engagement faster ($\rho = -0.26$), leaving positive reactions to those topics that require time to obtain maximum diffusion. Results described in the lower panel of Figure 6.4 provide further characterisation of the interplay between the Speed Index and the LH Score after classifying the topics according to the four most frequent categories analyzed, i.e., Politics, Labor, Human Rights and Health. We observe how the Politics and Health categories have the lowest correlation scores ($\rho = -0.36$ and $\rho = -0.45$), providing an indication of their intrinsic polarizing attitude (see Appendix for further details about correlation coefficients). Furthermore, the correlation between α and LH Score produces similar results as with the Speed Index (see Appendix for more details).

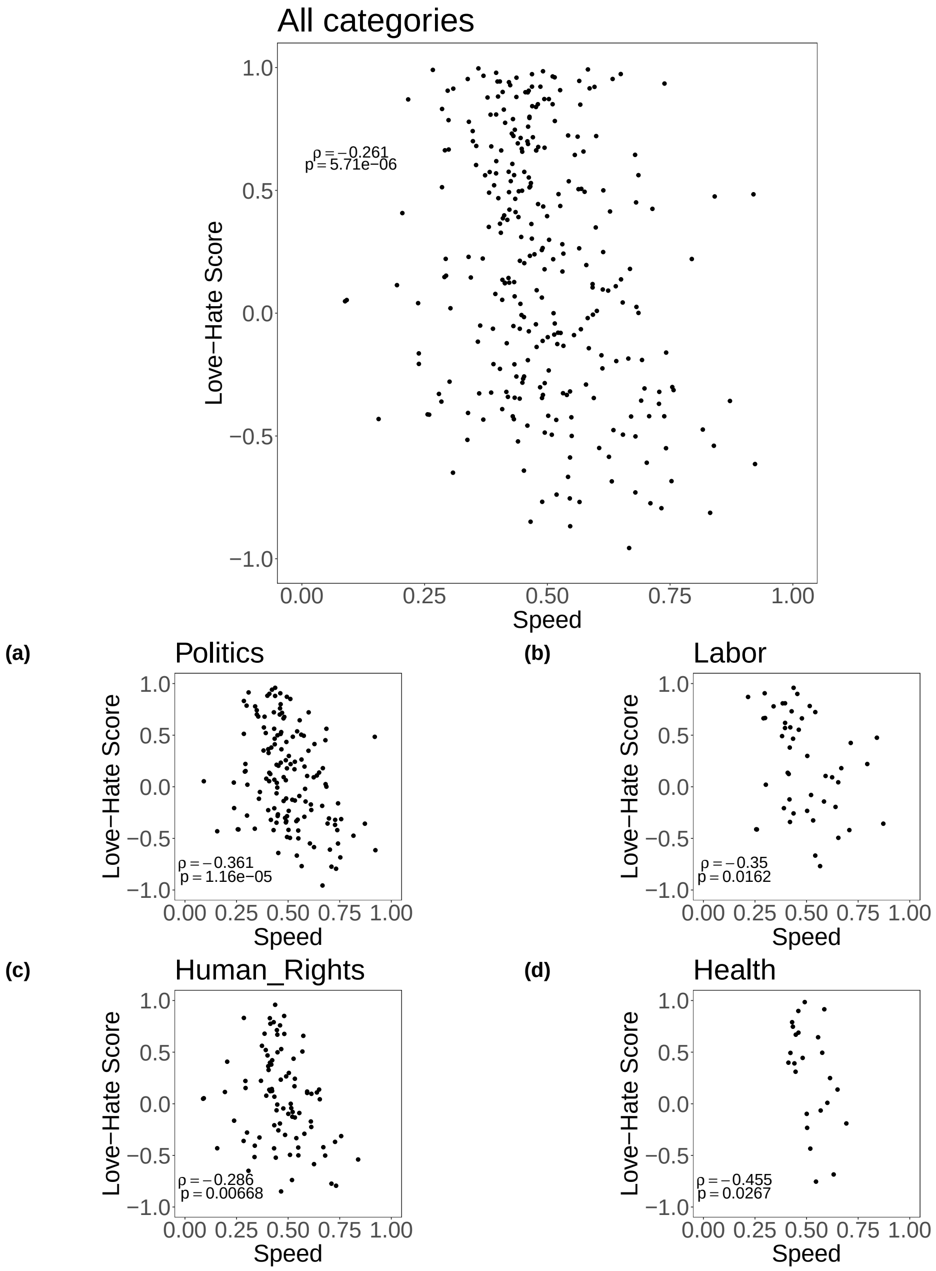


Figure 6.4. Upper panel: correlation between SI and LH score for each identified topic. Lower panel: correlation between SI and LH score for the top 4 most frequent topics. Overall, we observe how users react negatively as topics become sharply viral.

6.3.3 Assessing the differences of engagement behaviors across topic categories

To conclude our analysis, we investigate the differences in the evolution of engagement across topic categories. In particular, for each parameter distribution (α, β and SI), we apply a two-tailed Mann–Whitney U test [211] to each pair of parameters. Table 6.2 provides the percentages of the significant p-values for the four parameters. Due to the necessity to perform multiple tests, we apply a Bonferroni correction to our standard significance level of 0.05, leading to reject the null hypothesis if the p-value $p < 0.001$. Our results show that the resulting p-values from the tests do not lead to rejecting the null hypothesis. Such a result corroborates the hypothesis that, on average, users are characterized by homogeneous engagement patterns that are not influenced by the consumed topic. We further extend the statistical assessment by performing the same test between LH Score distributions of the different categories.

	α	β	**Speed Index**	**LH**
<0.001	2.22%	0%	0%	20%
>0.001	97.78%	100%	100%	80%

Table 6.2. Percentage of p-values resulting from the two-sided Mann–Whitney U test between each category employing their α, β, Speed Index and LH Score.

Conversely to engagement evolution results, the topic's category explains differences in the sentiment of reactions in 20% of cases. Such findings reveal that some categories are composed of significantly more negative and controversial topics, indicating how elicited reactions vary according to specific subjects. Understanding that some of them are more prone to induce negative feedback from users could be a proxy to introduce their related topics in the online debate.

6.4 Conclusions

In this Chapter, we perform a quantitative analysis of user interest on a total of $\sim 57M$ Facebook posts referring to ~ 300 different topics ranging from 2018 to 2022. We initially quantify the distribution of topics' engagement evolution throughout the analysis. Then, we evaluate the relationship between engagement and controversy. Ultimately, we assess the differences in engagement across different categories of topics. Our findings show that, on average, users' interest in topics does not increase exponentially right after their appearance but, instead, it grows steadily until it reaches a saturation point. From a sentiment perspective, topics that reached a plateau in their engagement evolution right after their initial appearance are more likely to collect negative/controversial reactions, whilst topics which are more steady in their growth tend to attract positive users' interactions. This result provides evidence about how recommendation algorithms should introduce topics adequately since sudden rises in topic diffusion could be related to the reinforcement of polarization mechanisms. Finally, we find no statistical difference between user interest across different categories of topics, providing evidence that, on a relatively large time window, the evolution of engagement with posts is primarily unrelated to their subject. On the contrary, we observe differences in the sentiment generated by topics with different diffusion speed, providing evidence of how people perceive the piece of content they consume online in different ways, according to how suddenly they get exposed to it.

Users' interest and engagement evolution in the online debate are both aspects of human behaviour on social media whose underlying dynamics still need to be discovered from an individual point of view. Our findings provide an aggregate perspective of the interplay between major emerging behavioral dynamics and topics' lifetime progression, deepening the relationship between diffusion patterns and users' reactions. Understanding that topics with an early burst in virality are associated with primarily adverse reactions from users may enable the identification of highly polarizing topics since their initial stage of diffusion.

The following study presents some limitations. In data collection, CrowdTangle provides only posts from public Facebook pages with more than 25K Page

Likes or Followers, public Facebook groups with at least 95K members, all US-based public groups with at least 2K members, and all verified profiles. These restrictions affected our datasets' sample and our findings' generality. Moreover, we could not access removed posts, groups, and pages, which could have been a meaningful proxy to characterize the attention dynamics of retracted content. Finally, since CrowdTangle does not provide information about users interacting with posts, we cannot assess their engagement from an individual perspective and model the possible relationship between users and topics employing a network approach.

The results obtained in this chapter may help to better understand how users consume information, improving social media moderation tools by considering both the "life-cycle" of topics and their potential controversy. Indeed, the introduction of the Speed Index and the Love-Hate Score can be exploited to identify in advance topics with the potential to collect considerable interest and generate heated debates quickly. From a news outlet and content creator perspective, understanding that specific topics may reach broader audiences and produce controversial opinions can improve the quality of the communication produced by these two types of authors.

Chapter 7

A Topology-Based Approach for Predicting Toxic Outcomes on Twitter and YouTube

In Chapter 6, we described how content belonging to viral topics is more associated with collecting controversial reactions, potentially responsible for fostering polarization dynamics due to the heated debates they produce. These debates, characterized by toxic language and antisocial behaviours, represent an open problem in the social media ecosystems, whose solution is non-trivial. Recent approaches to designing practical moderation tools focused either on the content of the conversations or the topology conversation tree. Both approaches, however, generally fail to consider the relationship between the topic discussed and the community involved, creating models that cannot enforce moderation policies on time or, on the contrary, are too preventive in limiting content. To address this gap, in this Chapter we describe a cross-platform comparison on Twitter and YouTube concerning the Italian Football League, a topic close to the Italian popular culture and a divisive topic - the Italian Political Elections. We first probe structural and conversational toxicity differences by analyzing 257K conversations (3.7M posts, 1M users) on both platforms. Then, we provide a machine learning approach that, by leveraging the previous features, identifies the presence of the following toxic comment in different stages of conversations. Our findings suggest that topics close to a community's popular culture tend to exhibit lower toxicity levels than divisive ones, with the latter producing more extended conversations that attract a broader audience. Moreover, we observe how text-based social media platforms like Twitter exhibit steady toxicity levels regardless of the topic being discussed, whilst media-based ones like YouTube report a decreasing trend as time passes. Lastly, the classifiers resulting from the conversation stage-based approach achieve state-of-the-art performances despite a restricted set of features. Furthermore, our cross-topic comparison shows that models trained on divisive topics can be generalized to other discussion arguments without causing a degradation of their performance.

7.1 Background

7.1.1 Defining toxicity on social media

The definition of online toxicity and toxic behaviors has evolved over the years due to the many disciplines it affects and the cultural factors involed. Prior work on this topic coined the term *hateful speech*, referring to any speech expressing hatred by the author against a person or people based on their identity [10]. Similar definitions from the juridical literature defined hateful speech as any form of expression that can increase harassment towards individuals or groups due to some characteristics they share or affiliation [212]. A further advance in the definition of toxicity was made in recent years by the United Nations, which formalized the concept of hate speech as "*any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language regarding a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factors*" [213]. More recently, researchers from Google Jigsaw, contextually with the introduction of their Perspective Application Programming Interface (API) [214], defined toxic content as any content characterized by "*rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion*" [215].

7.2 The rising of toxicity in social media

Social media platforms have reshaped how users inform themselves, frame the world and participate in online discussions [176, 177, 216]. Indeed, the microblogging features and the decentralized scheme proposed by these platforms provided the opportunity to be involved in an unprecedented number of debates, with the result of promoting the emergence of new ideas [217] and becoming rapidly aware of a multitude of topics [218]. Despite the potential benefits, social media are also considered responsible for fostering the spreading of online misinformation [6], selective exposure and echo chambers [3], which translate into an increasing number of heated debates [14, 13, 52]. These debates are characterized by toxic language and antisocial behaviours like cyberbullying [219], sexual harassment [12], trolling [15], and hate speech [220], potentially contributing to the rise of mental health issues [221, 222] and social division [223]. Therefore, to pursue the development of safer digital environments, it is crucial to identify early warnings of emergent toxicity and adequately moderate them. Many scholars have already faced this challenge with a mixture of approaches that ranged from the analysis of conversation cascades [13, 52, 224, 225] to Machine Learning (ML) [226, 227, 228, 50, 229, 230, 231, 232, 233]. However, little is known about the interplay between communities and the discussion topics [234, 235, 236]. This scenario raises the question of whether the closeness of a topic to the popular culture of a community may produce different toxicity dynamics than those known for being divisive, such as elections [19] or vaccination [237].

7.2.1 Conversation cascades and toxicity dynamics

Conversation cascades are an instance of the so-called *information cascades* whose properties and insights have been observed for years [225]. Despite the prior knowledge, the problem of curating online conversations has attracted increasing interest due to the societal implications it has [50, 17, 238]. Prior research efforts on this topic investigated the topological structures of conversations [239, 240] and proposed new generative models [241, 10] for their reconstruction. From a social media perspective, scholars made an extensive effort to analyze conversations and their role in anti-social behaviors like harassment [11, 12], spreading of misinformation [13] or trolling [14, 15]. Moreover, it was found that users tend to concentrate their anti-social efforts on a small number of threads [16], providing no evidence for the presence of "pure haters" [17]. From a dynamics perspective, it was observed how discussions on YouTube tend to degenerate towards increasingly toxic exchanges of views [17]. Such exchanges, however, have been demonstrated not to nourish misinformation spreading on social media [13]. Finally, a stream of work investigated the predictive power of structural content and user features to identify toxic comments and anti-social behaviors [10, 52, 242, 243], achieving important results in the selection of features to employ in the automatic identification of toxic elements.

7.2.2 Machine Learning for toxicity identification

From a ML perspective, the non-trivial task of identifying the presence of toxicity in online conversations has collected an increasing interest due to its implications for society and the technical challenge it poses. Researchers achieved promising results by applying architectures that ranged from traditional classifiers [226, 227, 228, 50, 229] to deep learning approaches, including Recurrent Neural Networks (RNN) [230] and Natural Language Processing (NLP) [229, 231, 232, 233]. Along this path, in 2017, Google Jigsaw introduced Perspective API [214, 244], a ML system that detects toxicity on online comments [245]. Despite its initial criticism [246, 247], the API was employed by multiple research works [52, 13, 248, 249], being recognized as a state-of-the-art tool in the context of online toxicity quantifying.

7.3 Preliminaries and definitions

7.3.1 Data Collection

We collect social media data concerning the 2022 Italian Political Elections and Football League. The first topic, Italian Elections, is known for being a polarising topic, especially in the case of the 2022 Italian Elections, where a strongly conservative party participated and won the elections, nourishing phenomena like echo chambers and polarization [250] and, eventually, offline disorders. Instead, the motivation for choosing the Italian Football League as a proxy for Italian popular culture is twofold. From a relevance perspective, football in Italy has the highest number of teams, thus a large geographical and media coverage, and it receives the highest number of public investments among all Italian sports [251]. From a toxicity perspective, we chose football due to its ability to spark anti-social behaviours, including tumults and brutal acts of violence [252, 253], which have the potential to be correlated with division and anti-social behaviors online.

The collection of posts and comments was performed on Twitter and YouTube to compare two regulated environments that rely on different media types, namely the text messages for Twitter and the videos for YouTube. The analysis includes all posts published from 25/08/2022 to 25/12/2022 with the corresponding comments. This period was chosen to capture the social media debate around the Italian electoral campaign, which ended on election day on September 25, 2022, and the following debate between the political parties involved once the winners were announced.

For the Football topic, we look for all posts containing at least one hashtag that refers to the Italian *Serie A* League team names and their slogans. Then, for each obtained post, we collect all the corresponding comments. The same approach was applied to the Elections topic, with the difference in the search hashtags that refer to political parties, exponents and general terms used by newspapers.

On Twitter, the data collection was performed by using Twitter API for Academic Research [254], producing a total of 3.6M posts for both topics, published by 300K users, and 8.2M Italian comments, identified by using Google's Compact Language Detector 3 (CLD3), from 550K users (see table 7.1 for further details). On YouTube, instead, posts with their comments were collected using the YouTube Data API [255], resulting in a dataset of 87K posts for both topics published by 10K channels, which produced 2.6M Italian comments, again identified with CLD3, from 381K users commenting (see table 7.1 for further details).

Social	Topic	Posts	Users Posting	Comments	Users Commenting	Percentage of labelled elements	Percentage of toxic elements
YouTube	Football	52 023	5 431	1 296 837	193 907	99.8	2.4
YouTube	Elections	35 479	5 087	1 393 369	187 791	99.6	5.2
Twitter	Football	1 404 010	120 407	1 780 583	235 385	98.6	2
Twitter	Elections	2 258 988	183 252	6 426 742	331 310	99.6	3.4
Total		3 750 500	314 177	10 897 531	948 393	99.4	3.2

Table 7.1. Data breakdown Twitter and YouTube data about Italian Football League and Elections.

7.3.2 Toxicity Labelling

In the current chapter, we refer to toxic content using the definition that Google Jigsaw Researchers provide, which identifies as toxic any content that is "*rude, disrespectful, or unreasonable language likely to make someone leave a discussion*" [215]. Consistently with the authors of this definition, the toxicity content classification is based on Google Jigsaw Perspective API [214]. Such API uses a ML model [245] to provide a score ranging from 0 to 1, indicating the probability that a reader would perceive the comment as toxic [256]. To define an appropriate threshold, we draw from the existing literature [256, 52, 13], indicating that any content with a toxicity score ≥ 0.6 is considered toxic. To assess the validity of this threshold, we also performed content classification with a threshold of 0.5 and 0.7. Among all topics and platforms, the 0.6 threshold provided the best tradeoff between the percentage of classified elements and the size of the resulting dataset to employ for the training of toxicity classifiers.

By applying Perspective API, we quantify the toxicity of the 98.6% of the total number of posts and comments in the dataset (see table 7.1 for further details). The remaining 1.4% comprises all those contents for which the model failed to produce a toxicity score. This scenario may happen with texts containing only emojis, special characters or lexical elements for which the API did not quantify their toxicity [245].

7.3.3 Conversation Cascade Reconstruction

We model a conversation cascade as a directed tree graph $T = (V, E)$, where $V = \{1, \ldots, n\}$ represents the set of nodes and $E = \{1, \ldots, m\}$ the set of links. Each node $v \in V$ can be either an original post that started the conversation, representing the tree's root, or a comment. On both platforms, the tree's root is characterized by an identifier (ID) that uniquely defines the conversation, shared by other nodes through the *conversation_id* attribute on Twitter and by the *video_id* on YouTube. The edges $e \in E$ instead represent the act of replying that links a node v_j to a node v_i, with $j > i$. For instance, the edge $e_1 = (v_1, v_2)$ means that the comment node v_2 has replied to the node v_1, which can be another comment or the root.

We implement the following procedure to reconstruct the conversation trees on each social media platform. On Twitter, we start from the root node and iterate on its children whose parent, represented by the *in_reply_to_id* attribute, corresponds to the root ID. For each identified node, we recursively look at their children with the same rationale until we reach all the tree leaves. The same procedure is applied on YouTube. However, in case of sub-conversations starting from a comment node v_i, YouTube will always indicate as v_i the parent of these nodes, despite the fact they may have replied to a child node v_j, such that $e_i = (v_i, v_j)$. Such limitations may prevent the algorithm from reconstructing the actual cascade structure. To overcome this problem, we apply a heuristic to reconstruct the tree graphs by looking at the latest comment posted by the user mentioned in a message (referring to its username indicated by *@Username*). If no username is found in the text, we indicate as the parent of the comment the root of the tree, i.e., the original post. Otherwise, we assign as the parent of the comment the ID of the most recent comment node posted by the user identified by its username in the sub-conversation. Finally, we label the nodes on both platforms based on the toxicity score of the element, as described in section 7.3.2. The resulting structure from this process is represented in fig. 7.1.

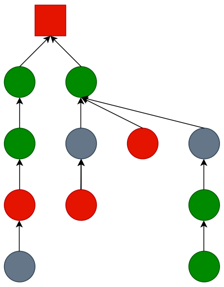


Figure 7.1. Graphical representation of a conversation tree. The root node representing the post is a square, while the children nodes (comments) are represented as circles. The nodes' colours represent the toxicity category assigned from their text. A node in green represents content whose text was identified by Perspective API with a toxicity score < 0.6, whilst a red node identifies an element with a toxicity score ≥ 0.6. Finally, grey nodes represent all those contents for which the API could not quantify their toxicity.

7.3.4 Cascade metrics

To provide a comparison between cascades, we define two categories of metrics. The first one called *structural metrics* and defined in Section 2.5, includes all the metrics related to the tree structure. The second one called *conversational metrics*, refers to additional information that is not strictly related to the topology of each conversation tree.

Conversational Metrics

Average Comment Intertime To quantify the average time, in seconds, lasting from the appearance of a comment and its successor in a conversation, we introduce a measure called Avg. Comment Intertime $CI(T)$. Given tree graph T, it is defined as

$$CI(T) = \frac{1}{n-1} \sum_{e \in E} \Delta t(e), \tag{7.1}$$

where $\Delta t(e) = t(w) - t(v)$ represents the difference between the timestamps associated to the nodes w and v, with $e = (w, v) \in E$.

Number of Unique Users The number of unique users $U(T)$ is the number of distinct users appearing in a post by posting or commenting, which is lower or equal to the Tree Size $TS(T)$, then $U(T) \leq n$.

Root Toxicity To account for the influence that the text of the initial post can have on the conversation, we assign a toxicity label to the root of each tree T, as described in section 7.3.2.

7.3.5 Permutation Test

To assess differences in the distribution of cascade metrics between different topics, we perform permutation tests whose algorithm is described in algorithm 1. For each metric, we consider the two distributions X_{ele} and Y_{foot} relative to the *Elections* and *Football* topics, keeping track of which population an observation is taken from. We begin by computing the test statistic m, defined as the absolute difference value between the mean of X_{ele} and Y_{foot}. Then, we unify the cascade distributions of two topics into a new one, called Z, and we shuffle the labels of the measures, obtaining Z^*, a set containing the same observations but (possibly) with different labels. Such operation allows us to perform the permutation tests by extracting the two shuffled distributions, i.e., X^*_{ele} and Y^*_{foot} based on their labels in Z^* and performing the absolute difference for their mean m^*. We repeat the procedure 1000 times and, as a result, we compute the probability that the test statistics m^*, observed in our null model, is higher (in absolute value) than m. We decide to use the permutation test since it can reduce the effects of imbalances in the sample sizes that may interfere with other tests, such as the Kolmogorov-Smirnov (KS) test.

7.3.6 Toxicity comment prediction in a conversation

Content moderation algorithms play a crucial role in the maintenance of online ecosystems. On the one hand, they must promptly limit the diffusion of harmful content. At the same time, too much limitation can prevent the emergence of vibrant discussions, impacting freedom of speech. Recent approaches to designing effective moderation [10, 52, 242, 243] tools focused on structural aspects of the conversations without effectively considering the relationship between the topic discussed and the community involved. To

Algorithm 1 Permutation test algorithm to assess statistical differences in the cascade metrics of two topics.

Input: Two topic metric distributions X_{ele} and Y_{foot}, where each measure posses a label identifying its provenience

Parameter: N, number of permutations

Output: p, the p-value resulting from the permutation test

```
c = 0
N = 1000
Calculate the test statistic $m = |\overline{X_{ele}} - \overline{Y_{foot}}|$
$Z = X_{ele} \cup Y_{foot}$ (maintaining the label of each observation)
let $i = 1$
while $i \leq N$ do
    $Z^*$ = shuffle the labels of observation in $Z$
    Extract ${X_{ele}}^*$ and ${Y_{foot}}^*$ from $Z^*$ according to their label in $Z^*$
    $m^* = |\overline{{X_{ele}}^*} - \overline{{Y_{foot}}^*}|$
    if $m^* \geq m$ then
        $c = c + 1$
    end if
    $i = i + 1$
end while
$p = \frac{c}{N}$
return $p$
```

address this gap, we propose a ML approach that differs from the current literature for two main reasons. First, we aim to provide a minimal yet effective feature set based on previously computed cascade metrics. Second, since it is known that structural feature importance is subjected to decaying as the tree size grows [225], we implement 4 different classifiers, each trained with comments belonging to specifc stages of a conversation. In terms of toxicity, we hypothesise such a solution will capture its evolution in the different stages of a conversation.

Dataset creation

Computing cascade metrics at comment-level We begin the dataset creation procedure by reconstructing, for each topic and platform, the conversation cascades as described in section 7.3.3. During the reconstruction, we filter out all those conversations with less than one comment to ensure the existence of at least a pair of toxic/non-toxic comments. Next, we compute the evolution of the features described in section 7.3.4 at the insertion time of each comment.

Creating a dataset for the toxicity prediction task In ML tasks involving cascades, it is mandatory to account for the decaying importance of their features as the size grows [225, 224, 257]. If not, the predictions produced by models trained on these data may be biased from the tree's current state. Drawing from previous approaches [258], we apply a dataset creation strategy that performs a logarithmic binning on the cascade size. Indeed, each unfolded conversation is split into four intervals, i.e., $(1, 10)$, $(10, 100]$, $(100, 1000]$, $(1000, 10000]$, according to the position assigned to a comment by entering in the conversation (comment index). This approach allows the creation of subsets that describe the different stages at which a conversation evolves, potentially helping the emergence of topological or conversational dynamics.

To optimize the separation between toxic and non-toxic elements, on each subset, we retain only those comments with a toxicity score provided by Perspective API less than 0.2, representing elements with a low presence of toxic language and greater or equal to 0.6, representing the toxic elements.

For each conversation in a subset, we create a pair of comments that include a toxic/non-toxic element until all toxic comments have a unique counterpart. However, to account for all those toxic comments without a counterpart, we randomly assign them a non-toxic element chosen from the subset in the exam. Then, we extract the features of both comments from all pairs, obtaining a cascade snapshot from a structural and conversational perspective when a toxic and non-toxic comment in the different conversations is posted. Finally, we end the dataset creation by performing an 80/20 split to obtain the train and test sets for the model training and testing phase.

Model training

To predict the occurrence of a toxic comment in a conversation, we implement an ensemble approach that consists of four ML sub-models, each specialized for a specific conversation stage as described in section 7.3.6. We train these models on a set of structural and conversational features, defined in section 7.3.4, to capture the different aspects that can bring to the production of toxic content in a conversation. We implement several ML-supervised models to identify the consistency of results and the most suitable model for this task, namely Logistic Regression (LR) models, Random Forests (RF), Decision Trees (DT), AdaBoost (AB), Support Vector Machines (SVM) and Gradient Boosted Regression Trees (GBRT). For each model, we tune its hyper-parameters through a 10-fold CV. The best model is refitted on the entire training set based on its accuracy score. For each dataset interval, we choose the best model with the highest F1 score, considering the Accuracy score in the case of a draw.
To estimate the predictive power of singular features, we proceed as follows. We first compute the F1 score s obtained by fitting the model m on the original dataset X. Next, we randomly shuffle its values for each feature $j \in [1, P]$ of the dataset, where P is the total number of features. For every shuffle $k \in [1, 10]$, we fit the model m on the dataset $\tilde{X}_{j,k}$ with the j-th column shuffled, obtaining a new score $s_{k,j}$. The importance of the feature i_j is defined as

$$i_j = s - \frac{1}{10}\sum_{k}^{10} s_{k,j}. \tag{7.2}$$

7.4 Results

7.4.1 Toxicity Evolution

We begin the analysis by comparing the toxicity evolution for the Italian Football League, representing a topic close to the Italian community, and the 2022 Italian Political Elections, representing a divisive topic. fig. 7.2(a) represents the average toxicity scores observed for each topic and social media platform during the analysis period. We observe that conversations about Italian Elections display higher toxicity levels than those about Italian Football. Indeed, on Twitter, Elections conversations produce an average daily toxicity score of 0.18 compared to the 0.09 for Football. The same behavior is found on YouTube, where the Elections topic attracts more toxicity than Football, with an average score 0.22 against the 0.13 of its counterpart. This result complies with the toxicity labelling results described in table 7.1 in which, on both social media, Elections contents have the highest percentage of toxic elements. We statistically assess this result by applying the KS test on both topic distributions for each social, obtaining a p-value $p < 0.05$ for both cases. Ultimately, we provide the first evidence of how the topic of Football produces conversations characterized by a lower presence of toxic language compared to political Elections.

Next, we quantify the rate at which toxicity evolved during the analysis period, assessing whether essential events in a community, like the Italian Elections voting day on September 25th, 2022, can reduce toxicity in its corresponding community. To achieve this goal, we estimate the evolution of toxicity on each topic through Ordinary Least Squares (OLS) regression models, defined as $\text{Toxicity}_t = \beta_0 + \beta_1 \text{Date}_t$.

Results from the fitting procedure show that YouTube is characterized by a decreasing trend of the toxicity scores for both topics ($\beta_1 = -2.59 \times 10^{-4}$ Elections and $\beta_1 = -5 \times 10^{-4}$ for Football), whilst Twitter presents a stationary trend for the Elections topic ($\beta_1 = 5.06 \times 10^{-5}$) and an increasing one for Football ($\beta_1 = 1.59 \times 10^{-4}$).

In terms of differences found coinciding with the voting day, fig. 7.2(b) reports a toxicity decrease of -21.98% on Football and -3.57% on YouTube, whilst on Twitter we note an increase of 3.03% for Football and a decrease of -0.42% for Elections. However, by conducting a KS test on the sample concerning the pre and post-event periods, we observed that the only significative, adverse changes in toxicity happened on YouTube with p-value < 0.05 for both topics against the 0.22 and 0.35 in the case of Elections and Football topics on Twitter.

To conclude our analysis, we look at the possible factors affecting previously reported toxicity trends. To do so, we compute the Pearson correlation score between the toxicity score and a set of proxies related to the content traffic volume and the user's behaviour, namely the number of posts (*Posts*), comments (*Comments*), and the number of users commenting (*Users Commenting*) and the comment they produce (*User Comments*). From the results reported in table 7.2, we observe that, on YouTube, the evolution of toxicity is positively linked with all the proxies introduced, identifying the role of the content volume in the production of online hate for both topics. More specifically, the positive correlation between the number of comments and users involved provides evidence of how online toxicity is closely associated with the length of discussions - represented by the number of comments - and with the commenting activity of users - represented by the number of user comments. On Twitter, toxicity in Football conversations appears to be linked to the number of posts generated about the topic, without being influenced by the commenting perspective. For the Elections topic instead, results confirm what was observed on YouTube, i.e., toxicity has a strict direct relationship with the commenting activity.

Ultimately, we provide evidence of how popular culture topics like Football tend to attract less hate than those inherently divisive, such as political elections. From a social media perspective, we observe how the primary content type of a platform may affect how toxicity evolves. Indeed, on Twitter, known for being a text-based social media, we report no significant changes in toxicity in the observed timespan. On YouTube instead, where videos are the primary source of information, we note a decreasing trend on the entire analysis period, which, on a comparison between the pre and post-election period, results in a significant reduction in the amount of hate circulating on both topics.

7.4.2 Structural analysis

We continue our comparison by investigating how the structure of conversations diverges according to their topic and platform. We first compute a set of structural metrics, described in section 7.3.4. Then, we assess the statistical validity of the obtained distributions using a permutation test, described in section 7.3.5, with a Bonferroni correction to account for multiple comparisons,

Twitter				
Topic	**Posts**	**Comments**	**Users Commenting**	**User Comments**
Football	0.53	0.01	-0.03	-0.04
Elections	0.04	0.38	0.39	0.28
YouTube				
Topic	**Posts**	**Comments**	**Users Commenting**	**User Comments**
Football	0.40	0.61	0.56	0.77
Elections	0.34	0.40	0.50	0.56

Table 7.2. Pearson correlation scores between average daily metrics concerning the toxicity scores, the number of posts and comments as well as the number of users commenting and how many times they commented daily.

considering p-values less than 0.00625 (0.05/8) as significant. fig. 7.3 reports the Complementary Cumulative Distribution Functions (CCDFs) computed on the previous cascade metrics for both topics and social media. We observe how, on Twitter, the Elections topic tends to attract bigger (*Tree Size*), wider (*Max Width*) and deeper (*Max Depth*) conversations. From a content perspective instead, Elections conversations are more likely to carry more toxic tweets (*Toxicity Ratio*) than those from Football. Conversely, users reading a Football conversation from the root have more chance to find a toxic comment earlier than In the Elections one (*Avg. Toxicity Distance*). The p-value of the statistical tests evidences how *Max Width* and *Number of unique users* are the only metrics on Twitter having no differences despite the topic.

7.4.3 Predicting the following toxic comment in a conversation

To conclude our analysis, we predict the toxicity of the following toxic comment in a conversation. Our results show that GBRT models achieve the highest performance on most configurations, whose results are reported in section 7.4.2. We report results containing the $(1, 10]$ interval for the sake of completeness, but we do not include them in the discussion of the results. The reason is that newborn conversations with few comments may not have established proper conversational dynamics yet, therefore not representing an adequate asset for toxicity predictors. The F1 scores reported for the Elections topic range between $[0.72, 0.78]$ on Twitter and $[0.70, 0.76]$ on YouTube. For Football instead, F1 scores range between $[0.79, 0.84]$ on Twitter and $[0.77, 0.84]$ on YouTube. Next, we create a baseline by training each model on datasets obtained by unifying all intervals for each topic-platform combination. The resulting metrics unveil how, in all configurations, the $(10, 100]$ interval produces greater or equal F1 scores than the baseline, providing evidence of how accounting for the different stages of a conversation may produce models with better performance and, therefore, with the ability to keep digital ecosystems safer.

Next, we investigate the generalizing power of models concerning the topics they were trained from. To do so, we perform a cross-topic evaluation for each social media: each stage model is trained on one topic and tested on its counterpart. section 7.4.2 displays the result of this comparison, where we observe a twofold scenario. On YouTube, training on Football data and testing against the Elections test set decreased F1 score by an average of 7%. The same result is observed even by training on Elections data and testing against the Football test set, with an average decrease of F1 score equal to 9%. Conversely, on Twitter, we observe a twofold effect. Whilst training on the Football comment and testing on Elections produced an average reduction of the F1 score equal to 8%, the opposite scenario produced an average increase of the metric equal to 8%. Such a result indicates that topics like Football, whose conversations are less toxic and participated, cannot generalize toxicity dynamics occurring in divisive topics like the Elections, resulting in a drop in performance. Instead, the models trained on cascades with a more articulated structure, like the Elections ones, tend to better generalize unknown observations in their feature space, achieving higher performance on a cross-topic benchmark. Finally, we assess the importance of each employed metric in this prediction task by measuring how the $F1$ would be impacted if a feature is removed. Results displayed in fig. 7.5 show, as expected, that the toxicity ratio

(*Toxicity Ratio*) is the most significant feature for predicting the toxicity of a comment, leading to an average reduction of 22% in the F1 score on both platforms, followed by the average toxicity distance (*Avg. Toxicity Distance*) (2%) and the assortativity (*Assortativity*) (1%). This result describes how combining cascade features with domain-specific information can be relevant in predicting harmful content.

7.5 Conclusion

In this chapter, we proposed a Twitter and YouTube comparison between the Italian soccer championship, a topic close to Italian popular culture, and the 2022 Italian general election, a divisive topic. We first assessed their differences in toxicity evolution, understanding which factors induce changes in “rude, disrespectful, or unreasonable" speech. Then, we compared conversations from a topological perspective by employing a set of structural metrics typical of cascades. Finally, we employed a ML approach, which, by creating four sub-models accounting for the different stages of a conversation, predicted the presence of the following toxic comment in a conversation. Our findings suggest that topics close to a community’s popular culture tend to exhibit lower toxicity levels than divisive ones, with the latter producing longer conversations that attract a broader audience. From a content perspective, we observed how text-based social media platforms like Twitter account for steady toxicity levels despite the topic being discussed, whilst media-based ones like YouTube report a decreasing trend as time passes. From a structural perspective, conversations from the Elections are broader, more toxic and involve more users. Moreover, the classifiers resulting from the stage-based approach achieved state-of-the-art results despite a minimal set of features, with models from early stages of conversations performing as well as those trained on the entire datasets.

Our study presents some limitations. The first limitation relates to the language of the conversation - Italian - which cannot easily generalize the findings reported to other languages. Moreover, results may suffer from a limited number of topics, platforms employed, and the period length. Nonetheless, global studies comprehending several platforms, topics and languages are rare due to several restrictions in the data-gathering process. Furthermore, our analysis relies on content that may have been moderated by Twitter and YouTube, resulting in discussions that only partially reflect the actual scenario that users experience in real time.

In future works, we aim to generalize by extending the topic choice and the list of platforms to perform the analysis, including unregulated ecosystems. Finally, to advance the quality of predictions, we also aim to define newer structural and conversational metrics to include in our models.

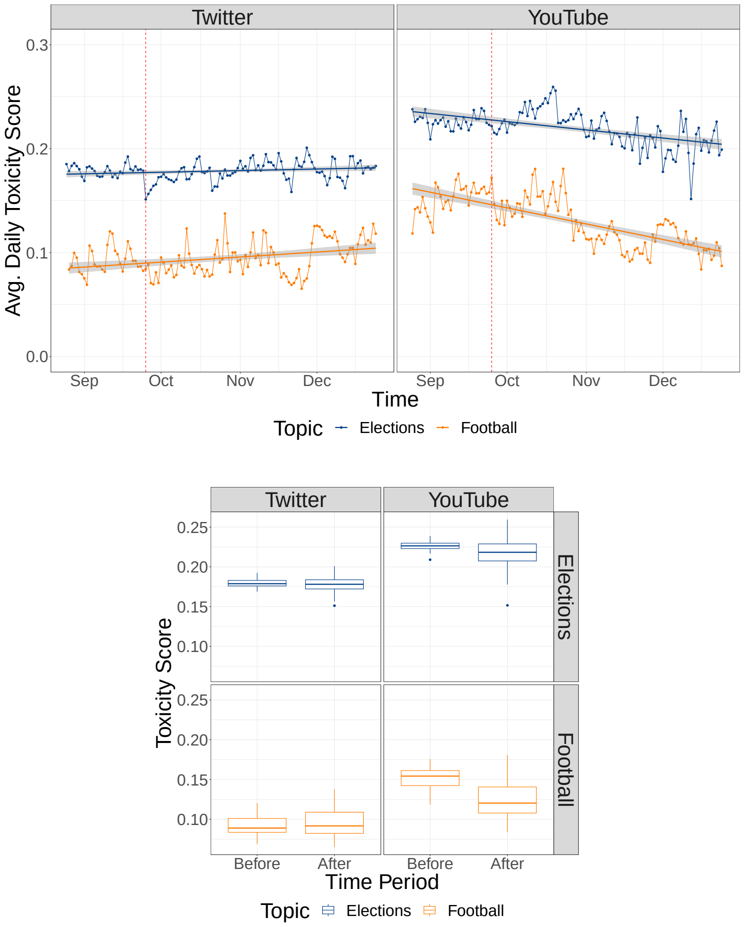


Figure 7.2. Left panel: average daily toxicity score reported on Twitter (left) and YouTube (right). The straight horizontal lines represent the linear fit performed on each trend. The red vertical line represents the date of the voting day for the Italian Elections (September 25, 2022). Right panel: toxicity score distributions for each social media and topic before and after the date concerning the Italian Elections voting.

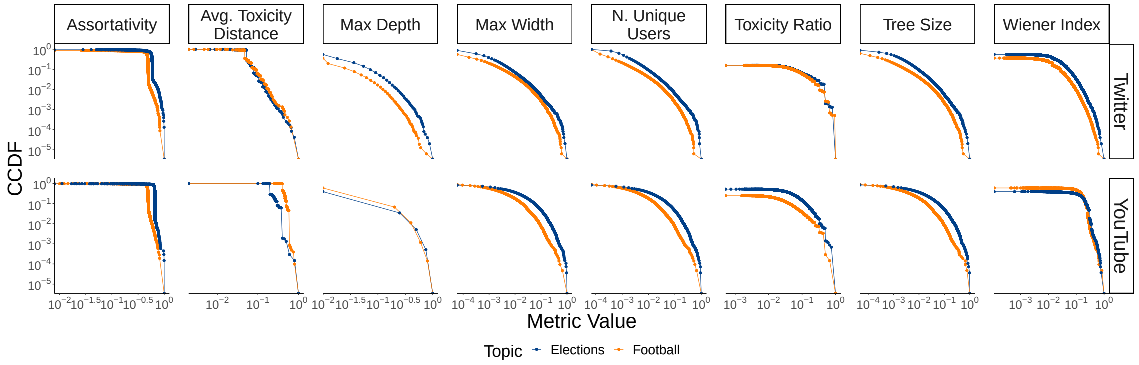


Figure 7.3. CCDFs of the standardized cascade metrics for Twitter (top) and YouTube (bottom).

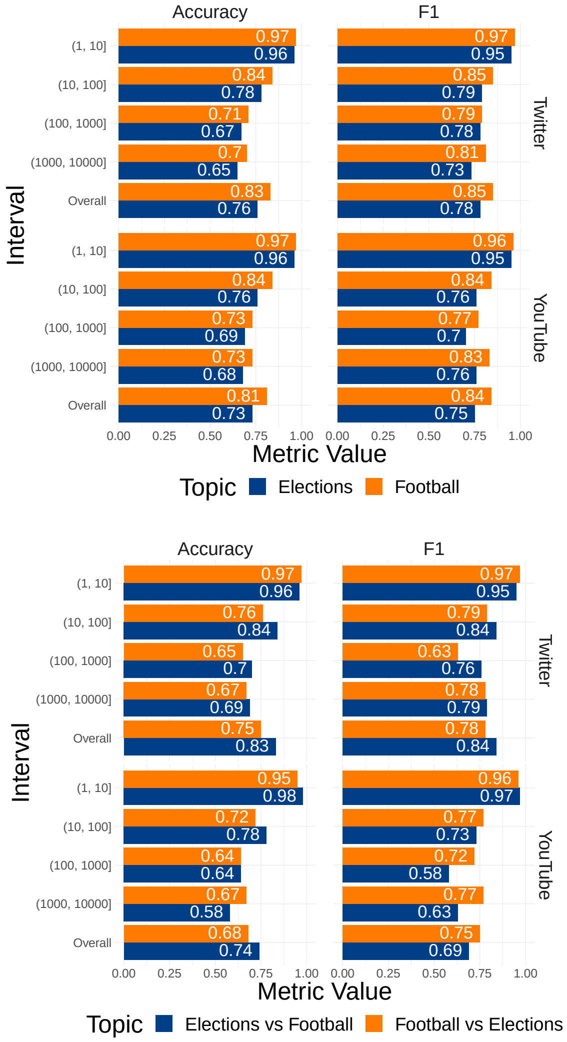


Figure 7.4. Left panel: prediction results of the GBRT model trained on intervals from each social media and topic. Right: Prediction results from a cross-topic comparison on each social media. We observe how performing out-of-topic prediction reduces prediction scores.

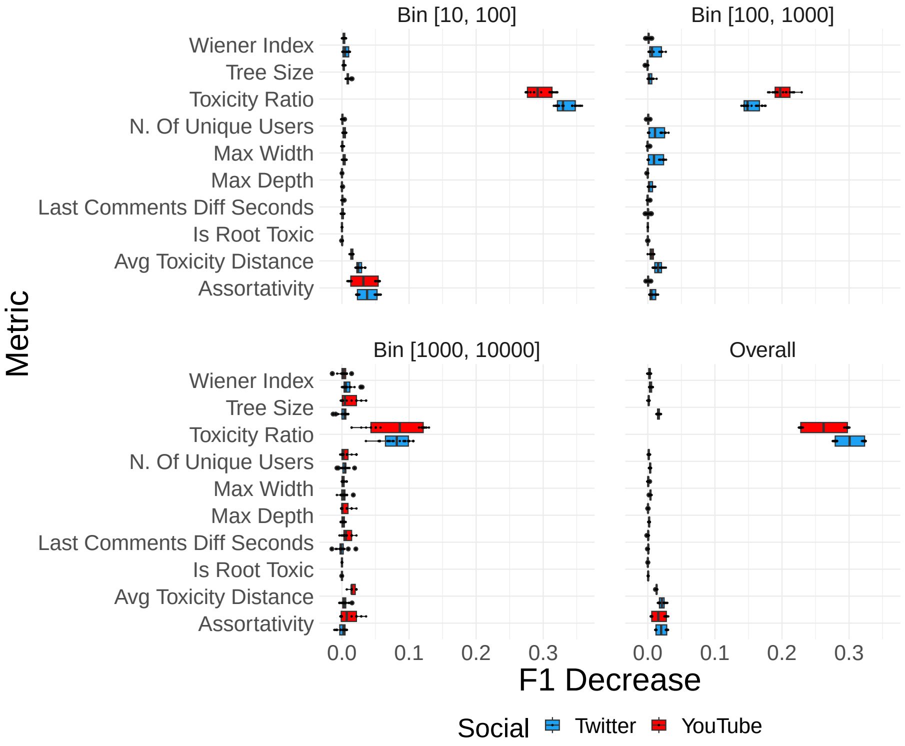


Figure 7.5. Representation of the importance of the features employed in the model, quantified by the average drop in F1 score corresponding to removing a specific feature.

Chapter 8

Toxicity in online conversations

In Chapter 7, we observed how controversial topics are more likely to produce conversations whose toxicity levels are higher than those closer to the culture of the commenting communities. This result confirms the allegedly known role of algorithms in the promotion and mitigation of content circulating online, avoiding promotion segregation and antisocial behaviors. However, research fails to understand how hate speech evolves on a broader scale, therefore lacking generalizable patterns which are common among platforms. To fill this gap, in this Chapter, we extend the results of the previous one with a comprehensive analysis of conversation dynamics across diverse social media platforms, specifically focusing on invariant patterns of toxic content across different platforms. Drawing from an extensive dataset spanning eight platforms over 30 years—from the advent of Usenet to contemporary platforms—our findings show consistent conversational patterns and user activity evolution irrespective of platform, topic, or era. Notably, long conversations consistently exhibit higher toxicity. Contrary to popular belief, toxic language does not invariably lead people to leave a conversation, and toxicity does not necessarily escalate as discussions evolve. Our results suggest that one of the main drivers of the observed dynamics is the ideological polarization of opinions among users, which may lead to more lively and hostile discussions. Remarkably, the trajectories of online toxicity have remained stable over three decades despite the advent of diverse platforms and evolving societal norms. By identifying consistent patterns of human behaviour beyond platform-specific features, our findings suggest directions to advance moderation policies on social media, including early interventions in discussions that may escalate toxicity.

8.1 Background

The digital age has seen an unprecedented rise in online participation, predominantly due to the proliferation of social media platforms [259, 260]. These platforms have smoothly integrated into our lives, serving as channels for information, entertainment, and personal communication [261]. However, concerns are growing about the potential pitfalls of such platforms for democracy [262, 263, 264, 265], and remedies seem far from effective [263, 264]. In fact, social media might exacerbate issues like polarization [138, 25, 266], spread misinformation [267, 68], and even encourage antisocial behaviors [5, 268]. Online users, indeed, may tend to seek information that is most aligned with their pre-existing beliefs, ignoring dissenting viewpoints [9, 269] and joining clusters of like-minded individuals[270, 271], where shared narratives may be collectively shaped and reinforced [68]. This "echo chamber" effect and related heightened polarization may vary across different social media platforms [3], primarily designed to prioritize user engagement instead of accurate information dissemination. In the face of such complex dynamics, the challenges posed by digital communication demand both careful scrutiny and interventions. This encompasses addressing the ramifications of misinformation, ideological polarization, and the fostering of uncivil behavior. One such ramifications is the deterioration in the quality of public debate, towards which increasing concerns have been raised. A facet of this discourse gaining considerable traction is the exploration of harmful language on social media and its broader repercussions both online and offline [272, 273, 274]. The burgeoning interest in this area aligns with recent advancements in machine learning models adept at identifying toxic language [275, 276, 277]. Numerous studies on online toxicity and related phenomena have been conducted so far, but most of them focus on specific platforms and discussion topics [278, 279, 280], whereas the broader, cross-platform ones are still limited in size and scope [268, 281, 282], creating a patchy picture that does not entirely capture the overall trends and mechanisms within the online debate dynamics. Moreover, this approach does not adequately take into account either the differences between platforms or the fact that users are often active on more than one of them, thus preventing the examination of invariant trends that can help distinguish what may be induced by algorithms from what instead represents inherent tendencies of human interaction. This fragmentation makes it challenging to ascertain whether certain beliefs regarding the role of toxicity in online conversations hold on a global scale or whether they may represent misconceptions. For example, do online discussions inevitably devolve into toxic exchanges, or are there different dynamics governing toxic conversations versus non-toxic ones? Moreover, have these dynamics changed over time? With a well-grounded understanding of such general properties of online discussions, efforts to mitigate related issues may be directed, aiding the design of better solutions and policy interventions. For instance, identifying common toxicity patterns across platforms and time periods can better inform moderation policies and practices on social media platforms, allowing platform-specific problems to be discerned from general independent behavior, thus finding more targeted solutions to prevent user exposure to toxic content. Here, we perform a comparative analysis of the online debate and its associated toxicity across all three dimensions that define it: time, hosting platform, and subject matter. In conducting our study and performing a comprehensive data analysis, which spans eight different platforms and encompasses more than 500 million comments, we base our understanding of toxicity on the definition derived from the Perspective API, a standard-bearer in the field of automatic toxic speech detection, which characterizes it as “a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion". However, we also demonstrate that this interpretation aligns consistently with alternative tools, yielding comparable results. Notice that different definitions of toxicity can be found in the literature (e.g., [283, 284, 233, 285, 286]), and that, due to the complex nature of the subject, the effectiveness and limitations of current machine learning-based automated toxicity detection systems has been debated in recent years [287, 235, 288]. Nonetheless, automatic systems are the only viable option for conducting large-scale analyses such as those presented here, and by focusing on aggregated results, they allow to capture of relevant macroscopic trends that would otherwise remain hidden. Indeed, the comprehensive approach in this study seeks to provide a more accurate and nuanced understanding of some of the complex dynamics that govern online conversations, shedding light on how these dynamics change over time, differ across platforms, and are influenced by the topics being discussed. By comparing multiple platforms, specific features, and patterns that stay consistent across various online environments can be noticed. These shared trends suggest that some patterns of online human interaction are steady, regardless of the surroundings. In the following examination of toxicity, we identify contributing factors and test common beliefs about how it affects conversations, adding measurable data to this critical discussion. Indeed, our analysis shows that – on average – people drop out or stay in a conversation regardless of its toxicity, that longer conversations contain a higher percentage of toxic exchanges, and that toxicity itself may not be a determining factor in such phenomena, but rather emerge as a byproduct of the ideological distance between participants in a discussion.

8.2 Materials and Methods

8.2.1 Data Collection

Facebook

We employed datasets from previous works that covered discussions about Vaccines [136], News [47] and Brexit [289]. Data were collected using Facebook Graph API, and all contents were anonymized as described in [3]. For the Vaccine topic, the resulting dataset contains $\sim 2M$ contents from Groups and Pages in a period that ranges from 1/1/2010 to 1/7/2017. For the News topic instead, the dataset contains $\sim 365M$ pieces of content between 1/1/2010 and 1/8/2016 that covered the English news ecosystem from several news outlets. To analyze the leaning of the authors we also used $\sim 4.5B$ likes the users posted on the pages of those outlets. Finally, for the Brexit topic, the dataset contains $\sim 500K$ pieces of content from 1/12/2015/ to 1/7/2016.

Gab

The entire feed of the platform ($\sim 34M$ comments) from its launch on 10/8/2016 to 29/10/2018 (when Gab went temporarily offline) was downloaded from the Pushift.io archive (https://files.pushshift.io/gab/).

Reddit

Data were collected from the Pushift.io archive (https://redditsearch.io/) on a period that ranged from 1/1/2018 to 31/12/2022. For each topic, whenever possible, we referred to its most representative subreddit. As a result of this operation, we obtained $\sim 800k$ contents from the *r/conspiracy* subreddit for the Conspiracy topic. For the Vaccines topic, we collected $\sim 60K$ contents from the *r/VaccineDebate* subreddit, focusing mainly on the COVID-19 vaccine debate. For the News topic, we collected $\sim 600K$ contents from the *r/News* subreddit from 1/1/2018 to 31/12/2018. For the Climate change topic, we collected $\sim 300K$ contents from the *r/environment* subreddit. Finally, for the Science topic, we collected $\sim 600K$ contents from the *r/science* subreddit.

Datasets in Appendix We collected data from the largest general Q&A subreddit, *r/AskReddit*, between 1/2/2021 and 15/2/2021 ($\sim 2.5M$ contents), from the general conversational subreddit *r/Iama* between 1/1/2020 and 31/12/2022 ($\sim 500K$ contents), and from the *r/movies* film discussion subreddit from 1/1/2020 to 1/1/2023 ($\sim 12M$ contents).

Telegram

We built a list of 14 channels, associating each of them with one of the topics considered in the study. For each channel, we manually collected messages with their related comments. As a result, from the 4 channels associated with the News topic (*news_notiziae, news_ultimora, news_edizionestraordinaria, news_covidultimora*), we obtained $\sim 2M$ comments from posts between 1/8/2019 and 31/12/2022. For the Politics topic, instead, the 2 channels (*politics_besttimeline, politics_polmemes*) assigned produced $\sim 0.7M$ messages and comments between 1/8/2017 and 31/12/2022. Finally, the 8 channels assigned to the Conspiracy topic (*conspiracy_bennyjhonson, conspiracy_tommyrobinsonnews, conspiracy_britainsfirst, conspiracy_loomeredofficial, conspiracy_thetrumpistgroup, conspiracy_trumpjr, conspiracy_pauljwatson, conspiracy_iononmivaccino*) produced $\sim 0.9M$ comments between 1/4/2018 and 31/12/2022.

Datasets in Appendix We collected the cryptocurrency related *Watcher-Guru* channel in a time range: 23/7/2021-22/6/2023, for a total of $\sim 150K$ contents.

Twitter

We employed a list of datasets from previous works that include discussions about Vaccines [193], Climate Change [89] and News [13] topics. For the Vaccine topic, we collected a total of $\sim 56M$ posts and comments from 1/1/2010 to 31/12/2022. For the News topic, we extended the dataset employed in [13] by collecting all threads composed of less than 20 comments, obtaining a total of $\sim 9.5M$ pieces of content on a time range that goes from 1/1/2020 to 30/11/2022. Finally, for the Climate change topic, we collected $\sim 0.9M$ posts and comments between 1/1/2020 and 31/12/2023. Notice that the retrieved conversations were obtained by performing a keyword search based on the Conference of the Parties (COP), i.e. content was selected if it contained the term “cop2x", with $x \in \{0, \ldots, 6\}$.

Datasets in Appendix We collected original posts with their comments about the *Game of Thrones* series and *NASA*. Concerning the first, we retrieved all original tweets containing the hashtag #gameofthrones from January 1st 2019 to April 15th 2019, and their comments, for a total of $\sim 441K$ contents. For NASA, we collected all tweets published by the official NASA account on Twitter from January 1st 2019 to December 31st 2021, and their comments, for a total of $\sim 343K$ contents.

Usenet

We collected data for the Usenet discussion system by querying the Usenet Archive (https://archive.org/details/usenet?tab=about). We selected a list of topics considered adequate to contain a large, broad and heterogeneous number of discussions involving active and populated newsgroups. Ultimately, we chose Conspiracy, Politics and News as topic candidates for our analysis.

For the Conspiracy topic, we collected $\sim 320K$ pieces of content between 1/9/1994 and 31/12/2005 from the *alt.conspiracy* newsgroup. Then, for the Politics topic, we collected $\sim 3.2M$ pieces of content between 1/6/1992 and 31/12/2005 from the *alt.politics* newsgroup. Finally, for the News topic, we collected $\sim 750K$ pieces of content between 1/6/1992 and 31/12/2005 from the *alt.news* newsgroup. Finally, we collected all the conversations from the newsgroup *talk* from 1/2/1989 to 31/12/2005 for a total of $\sim 2.3M$ contents.

Voat

Data analysis was performed relying on a Voat dataset which has been employed by other researchers in the past [60]. It covers the entire lifetime of the platform, from 8/1/2013 to 25/12/2020, including a total of $\sim 16.2M$ contents posted from $\sim 113K$ users in $\sim 7.1K$ subverses (the equivalent of a subreddit for Voat). We associated some topics of interest from the entire corpus with specific subverses. For the Conspiracy topic, we collected $\sim 1M$ contents from the *greatawakening* subverse between 1/1/2018 and 31/12/2020. For the Politics topic, we collected $\sim 1M$ contents from the *politics* subverse between 1/6/2014 and 1/12/2020. Finally, for the News topic, we collected $\sim 1.4M$ contents from the *news* subverse between 1/11/2013 and 31/12/2020.

Datasets in Appendix We collected all contents in a time range from two Q&A generic subverses, namely $\sim 600K$ contents from the *askvoat* subverse between 19/6/2014 and 25/12/2020, and $\sim 1.4M$ contents from the *whatever* subverse from 31/5/2015 to 25/12/2020.

YouTube

Data from YouTube relied on previous works that included contents about the Vaccine [193], News [13], and Climate Change [25] topics. The data were collected through a keyword search based on the YouTube Data API (`https://developers.google.com/youtube/v3`). For the Vaccine topic, we collected data between 1/1/2020 and 1/10/2020, obtaining $\sim 2.5M$ comments to videos. We employed keywords related to the most discussed and used COVID-19 vaccines brands at the time of the data download, namely {*Sinopharm, CanSino,* *Janssen, Johnson&Johnson, Novavax, CureVac, Pfizer, BioNTech, AstraZeneca, Moderna*}. For the News topic, data were collected between 1/2/2006 and 31/1/2022, gathering all videos and comments from a list of news outlets provided by Newsguard. As a result of this operation, we collected $\sim 20M$ pieces of content. Finally, for the Climate change topic, we collected $\sim 1M$ videos and comments between 1/3/2014 to 1/2/2022. Notice that the retrieved data were obtained by performing a keyword search based on the Conference of the Parties (COP), i.e. content was selected if it contained the term "cop2x", with $x \in \{0, \ldots, 6\}$.

Datasets in Appendix Data collection on YouTube was performed using the YouTube Data API. We collected videos that contain at least one search term from the following list of keywords related to football: {*SerieA, SerieATim, VAR, Napoli, ForzaNapoliSempre, RangersNapoli, Atalanta, GoAtalantaGo, ForzaAtalanta, Milan, ACMilan, SempreMilan, Udinese, ForzaUdinese, AlèUdin, Inter, IMInter, ForzaInter, Forza Lazio, La Lazio, SSLazio, CMonEagles, ASRoma, Juventus, JuventusFC, Juve, ForzaJuve, IlTorino, Forza Torino, SFT, Salernitana, forzagranata, Fiorentina, forzaviola,ACFFiorentina, IlBologna, ForzaBologna, ForzaBFC, WeAreOne, IlSassuolo, Forza Sassuolo, ForzaSasol, LEmpoli, ForzaEmpoli, EmpoliFC, EmpoliFootballChannel, HellasVerona, Hellas, DaiVerona, HVFC, LoSpezia, ForzaSpezia, SpeziaCalcio, IlLecce, ForzaLecce, avantilecce, Cremonese, SolAmAi, ForzaGrigiorossi, DaiCremo, Sampdoria, FORZADORIA, Il Monza, Forza Monza, ACMonza, Monza, InsiemealMonza*}. Then, we retrieved all the comments for each video, resulting in a dataset of more than 16.8K videos which produced a total of ∼ 1M comments. Consistently with previously mentioned keywords search, we collected another dataset using only the keyword {*carbonara*}. We gathered more than 4300 videos in the window ranging between 05/01/2018 and 30/06/2023 which received $\sim 700K$ comments.

Content Moderation Policies

Content moderation policies are guidelines that online platforms use to monitor the content that users post on their sites. Platforms have different

Platform and topic	Time range	Comments	Threads	Users	Toxicity
Facebook brexit	2015/12/31-2016/07/29	464765	4241	252157	0.07
Facebook news	2009/09/09-2016/08/18	364696312	6845663	60140204	0.06
Facebook vaccines	2010/01/02-2017/07/17	2081160	153138	388553	0.05
Gab other	2016/08/10-2018/10/29	32623269	18648850	292365	0.09
Reddit climatechange	2018/01/01-2022/12/12	70648	5057	26521	0.07
Reddit conspiracy	2018/01/01-2022/12/08	777393	35092	92678	0.07
Reddit news	2018/01/01-2018/12/31	389582	7798	109860	0.09
Reddit science	2018/01/01-2022/12/11	549543	28330	211546	0.01
Reddit vaccines	2018/01/21-2022/11/06	66457	4539	5192	0.04
Telegram conspiracy	2019/08/30-2022/12/20	1833538	32852	150251	0.14
Telegram news	2018/04/09-2022/12/20	829621	28387	16716	0.02
Telegram politics	2017/08/04-2022/12/19	612800	27994	6132	0.04
Twitter climatechange	2020/01/01-2023/01/10	9659826	126481	3560651	0.07
Twitter news	2020/01/01-2022/11/29	9511143	90947	2000909	0.06
Twitter vaccines	2010/01/23-2023/01/25	56145723	130049	12788237	0.08
Usenet alt.politics	1992/06/29-2005/12/31	2658418	626099	209930	0.09
Usenet conspiracy	1994/09/01-2005/12/31	292280	74975	49259	0.05
Usenet news	1992/12/05-2005/12/31	689463	184059	82779	0.09
Usenet talk	1989/02/13-2005/12/31	2241399	350071	163454	0.06
Voat conspiracy	2018/01/09-2020/12/25	1024812	99953	27667	0.10
Voat news	2013/11/21-2020/12/25	1397955	170801	88454	0.19
Voat politics	2014/06/19-2020/12/25	1083932	143103	66441	0.19
YouTube climatechange	2014/03/16-2022/02/28	846355	9022	467250	0.07
YouTube news	2006/02/13-2022/02/08	20617098	108352	5180300	0.08
YouTube vaccines	2020/01/31-2021/10/24	2713372	13614	1000635	0.04

Table 8.1. Dataset breakdown of the study.

goals and audiences, and their moderation policies may vary greatly, with some placing more emphasis on free expression and others prioritizing safety and community guidelines.

Facebook and YouTube have strict moderation policies that prohibit hate speech, violence, and harassment [290]. To address harmful content, Facebook follows a Remove, Reduce, Inform strategy and uses a combination of human reviewers and artificial intelligence (AI) to enforce its policies [291]. YouTube has a similar set of community guidelines regarding hate speech policy which covers a wide range of behaviors such as vulgar language [292] and harassment [293]. The platform clearly states that hate speech and violence against individuals or groups based on various attributes are not allowed [294]. YouTube too employs a mix of AI algorithms and human reviewers to enforce its guidelines [295].

Twitter also has a comprehensive content moderation policy and specific rules against hateful conduct [296, 297]. They use automation [298] and human review in the moderation process [299]. Twitter's content policies have remained unchanged since Elon Musk's take over, except for ceased enforcing their COVID-19 misleading information policy since November 23, 2022. Their policy enforcement has faced criticism for inconsistency [300].

Reddit falls somewhere in between regarding how strict is its moderation policy. Reddit's content policy has eight rules, including prohibiting violence, harassment, and promoting hate based on identity or vulnerability [301, 302]. Reddit relies heavily on user reports and volunteer moderators, hence they could be considered more lenient than Facebook, YouTube, and Twitter regarding enforcing rules. In October 2022, Reddit announced they intend to update their enforcement practices to apply automation in content moderation [303].

In contrast, Telegram, Gab, and Voat take a more hands-off approach with fewer restrictions on content. Telegram has ambiguity in its guidelines which arises from the use of broad or subjective terms and it can lead to different interpretations[304]. Although they mentioned they may use automated algorithms to analyze messages, Telegram relies mainly on users to report a range of content, such as violence, child abuse, spam, illegal drugs, personal details, and pornography [305]. According to Telegram's privacy policy, reported content may be checked by moderators, and if it is confirmed to be in violation of their terms, temporary or permanent restrictions may be imposed on the account [306]. Gab's Terms of Service allow all speech protected under the First Amendment to the U.S. Constitution, while unlawful content is removed. They state that they do not review material before it is posted on their website, and they cannot guarantee prompt removal of illegal content after it has been posted [307]. Voat was once known as a “free-speech" alternative to Reddit and allowed content even if it may be considered offensive or controversial [60].

Usenet is a decentralized online discussion system created in 1979. Because of its decentralized nature, Usenet has been difficult to moderate effectively, and it has a reputation for being a place where controversial and even illegal content can be posted without consequence. Each individual group on Usenet can have its own moderators, who are responsible for monitoring and enforcing their group's rules, and there is no single set of rules that applies to the entire platform [308].

8.2.2 Logarithmic binning and conversation sizes

Due to the heavy tailed distributions of conversation sizes, to plot the figures and make the relative analyses we used a logarithmic binning, dividing each dataset into 21 bins. To lessen the impact of outliers and ensure a minimal adequate number of points for robust statistics, we iteratively changed the left bound of the last bin so that it contains at least $N = 50$ elements (we were able to set $N = 100$ in the case of Facebook news, due to its larger size). Specifically, the size of the largest thread was changed to that of the second last largest and the binning recalculated accordingly, until the last bin contained at least N points. Participation analysis, for each dataset we selected only those conversations that belong to the $[0.7, 1]$ interval of the normalized logarithmic binning of thread sizes. This interval ensures that the conversations are sufficiently long and that we have a substantial number of threads on which to compute participation in each dataset. Participation trends were then calculated for these datasets by dividing each thread - chronologically ordered sequence of comments - into 20 equal percentile intervals. In Table 9.12 in Appendix is reported a breakdown of the resulting datasets.

8.2.3 Toxicity detection and validation of Perspective API

The problem of detecting toxicity in highly debated, to the point that there is currently no agreement on the very definition of "toxic speech"(see e.g. [287] for a sociolinguistic and anthropological account). A toxic comment can be regarded as one that include obscene or derogatory language [283], that employs harsh, abusive language and personal attacks [284], or contains extremism, violence and harassment [235], just to give a few example. Even though toxic speech should in principle be distinguished from hate speech, which is commonly more related to targeted attacks that denigrate a person or a group on the basis of attributes such as race, religion, gender, sex, sexual orientation etc. [309], it sometimes may also be used as an umbrella term [310, 311]. This lack of agreement is a direct reflection of the challenging and inherent subjective nature of the concept of toxicity. The complexity of the topic makes it particularly difficult to assess the reliability of Natural Language Processing (NLP) models for automatic toxicity detection, despite the recent years dramatic improvements in the field. Modern NLP models, such as Perspective API, leverage word embedding techniques to build representations of words as vectors in a high-dimensional space, in which a metric distance should reflect the "conceptual" distance among words, thus providing linguistic context. Then, machine learning models are trained to detect toxicity based on huge amounts of annotated data, that is, pieces of text to which one or more human annotators assign a categorization label. As a result these classifiers learn to recognize patterns and features such as combinations of words or sentence structures typically used in toxic exchanges. The main criticism towards this approach to toxicity detection is that these models have limited capabilities to take into account the context in which a conversation develops [288, 235], and that they suffer from biases inherited from the annotators and the peculiarities of the datasets on which their training process is based [310, 312]. It is argued that determining whether a text is toxic requires a wealth of knowledge of the context outside of the specific content, such as the personal characteristics and motivations of the toxic source and the target recipient, the conversation participants relationships and group membership, and the general tone of the discussion [235]. For instance, what one group (e.g. ethnic, age) may perceive as toxic content, others may view as completely acceptable [312] – a factor that, at training time, may also insert biases in the classificator, if the pool of human labelers is not selected with sufficient cultural heterogeneity; Toxic content can also be concealed in indirect allusions, memes, and inside jokes that are directed towards particular groups and intended for specific audiences. These are challenges in NLP. The word embeddings provide current classifiers with a rich linguistic context that enables the recognition of a vast variety of patterns that are proper to toxic expression, but the aforementioned requirements are clearly out of the scope of automatic detection models. We acknowledge these limitations. However, any attempts at conducting a large-scale analysis of toxicity online such as in this study has to be based on automatic detection models. We add that when focusing on aggregated results from hundreds of millions of pieces of text, the highlighted shortcomings should dilute in the statistics, allowing to identify meaningful macroscopic patterns and trends. In this study we employed Perspective API, which is currently a state-of-the-art toxicity classifier available for labelling large amounts of data. Due to the limitations mentioned above (for a criticism of Perspective API see [313]), we validated our results by performing a comparative analysis using two other toxicity detectors: Detoxify (`https://github.com/unitaryai/detoxify`), which is similar to Perspective, and IMSYPP, a classifier developed for a European Project on hate speech [5] (`https://huggingface.co/IMSyPP`). In Table 9.19 in Appendix are reported the percentages of agreement among the three models in classifying $100K$ comments taken at random from each of our datasets. For Detoxify we used the same binary toxicity threshold (0.6) used with Perspective. Even though IMSYPP is based on a different definition of toxicity [5], the results show decent general agreement. Moreover, we perform the core analyses of this study using all classifiers on a further, vast and heterogeneous dataset. As shown in Figure 9.14 and in Figure 9.15 in Appendix, the results regarding toxicity increase with conversation size and users' participation and toxicity are quantitatively very similar, but most importantly qualitatively identical, providing evidence that our findings are robust with respect to the choice of the toxicity detector employed. Finally, we tested that our findings do not change if a different threshold for toxicity is chosen. In the main analysis, the analyses adopted the threshold value suggested by Perspective API, 0.6, which could be considered conservative, to prevent false positives. Figure 9.17 in Appendix show that our findings hold perfectly even if a reasonable, less conservative toxicity threshold (0.5) is used. We conclude this section by reporting that Perspective API works with multiple languages, and for this study we selected all the available languages spoken in the European and American continents, namely English, Spanish, French, Portuguese, German, Italian, Dutch, Polish, and Swedish, plus Russian. Detoxify is also multilingual, whereas IMSYPP can be applied to English and Italian text.

Polarization and user leaning attribution

Our measure of controversy within a conversation relies on an estimate of the degree of political partisanship of users involved in a discussion. This quantity is thus tightly linked to the political science concept of political polarization, intended as the divergence of political attitudes away from the center and towards ideological extremes [314]. Within this definition, a distinction is often made between the so-called ideological polarization, pertaining to the divide in terms of political viewpoints, and affective polarization, which instead refers to the emergence of positive emotions toward members of one's own faction and to the contextual hostility toward those of the opposite one [315, 316]. In this study, when we speak of polarization, we refer to the former, as the following description of the users' leaning attribution procedure should clarify. On online social media, the individual leaning of a user toward a topic can be inferred via the content produced, or the endorsement shown toward specific content. In this study, we considered the endorsement of users to news outlets whose political leaning has been evaluated by trustworthy external sources. While not without limitations – that we address below -, this is a standard approach used in several studies, which has become a common and established practice in the field of social media analysis due to its practicality and effectiveness in providing a broad understanding of political dynamics on these online platforms [3, 317, 318, 96, 45] We labeled news outlets with a political score based on the information reported by Media Bias/Fact Check (MBFC) (https://mediabiasfactcheck.com), integrating with the equivalent information from Newsguard (NG) (https://www.newsguardtech.com/). MBFC is an independent fact-checking organization that rates news outlets on the basis of the reliability and of the political bias of the contents they produce and share. Similarly, NG is a tool created by an international team of journalists that provides trust and political bias scores of news outlets. Following standard methods employed in literature [3, 317], we calculated the individual leaning of a user $l \in [-1, 1]$ as the average of the leaning scores $l_c \in [-1, 1]$ attributed to each of the content it produced/shared, where l_c results from a mapping of the news organizations political scores provided by MBFC and NG (respectively: [left, center-left, center, center-right, right] and [far left, left, not aligned, right, far right] to $[-1, -0.5, 0, 0.5, 1]$. Our datasets have different structures, so we had to evaluate user leanings in different ways. For Facebook News, we assigned a leaning score to users that posted a like at least three times, and commented at least three times under news outlet pages that have a political score. For Twitter News a leaning was assigned to users who posted at least 15 comments under scored news outlet pages. For Twitter Vaccines and Gab we considered users that shared at least three times content produced by scored news outlet pages.

One limitation of the procedure described above is that it is not guaranteed that commenting or sharing materials from politically aligned sources always means adhering to the source point of view, as one could engage with content from opposing political viewpoints for critical discussion. While true, research suggests that that is relatively rare compared to the sharing of content that aligns with one's own views, especially in politically charged discussions (see e.g. [319, 320, 18]). Also, social media leaning attribution procedures capture users that openly share their political leanings, leaving out most "passive" ones. This is inevitable, because when no information about users is available, only those who express their opinions in some way can be profiled. However, focusing on active users provides valuable insights into the discourse among those who are most engaged and influential in online platforms.

8.2.4 Burst analysis

We applied the Kleinberg burst detection algorithm [321] to all conversations with at least 50 comments in a dataset considering up to 5000 randomly selected conversations with such number of comments. To obtain comparable results, and after making sure that in the vast majority of cases the peak of activity occurs in the first 24 hours of a thread, we considered only comments posted in this time interval for each thread. Usenet was not included because of the much longer response times between comments. Labelled as discrete positive values, higher levels of burstiness represent higher activity segments. To avoid considering flat density phases, threads with maximum burst level equal to 2 are excluded from this analysis. We performed a Mann-Whitney U test [322], with Bonferroni correction for multiple testing, between the fraction of toxic comments in three activity phases: during the peak of activity and at the highest levels before and after. Table 9.14 in Appendix shows the corrected p-values of each test, at 0.99 confidence level, with H1 indicated in the column header. An example of distribution of frequency of toxic comments in threads at the three phases of a conversation considered (pre-peak, peak and post-peak) is reported in Figure 8.4c).

8.2.5 Toxicity Detection on Usenet

As pointed out in sec. Toxicity detection and Perspective API, automatic detector tools build their “knowledge" of what is toxic based on the annotated sets of data they are trained on. Since Perspective API is trained mostly on texts from recent or current years and its human labelers adhere to our contemporary cultural norms, in principle there is no guarantee that its application to texts produced several decades ago can be considered entirely valid, because linguistic and cultural changes may affect its reliability. Therefore, although our dataset only goes back 20-30 years, we provide here a discussion on the viability of the application of Perspective API to Usenet, and a validation analysis. As far as cultural factors are concerned, it is clearly evident from the public debate that our contemporary (Western) society is much more sensitive and attentive regarding toxicity than it was a few decades ago, especially regarding topics of gender, race, sexual orientations, disabilities, and related issues. What is considered toxic nowadays is detected on Usenet as it is on more recent platforms, with the possible inclusion of some comments that would have not been perceived as toxic by the cultural standards of the time, which is of no concern for our analysis. On the other hand, changes in linguistic features may have some repercussions: There may be words and locutions that were frequently used in the 1990s that instead appear sparsely in today’s language, making Perspective potentially less effective in classifying short texts that contain them. We then proceeded to evaluate the impact that such a possible scenario could have on our results in the following way. In light of the above considerations, we consider texts labelled as toxic as correctly classified; instead, we assume there is a fixed probability p that a comment may be incorrectly labelled as non toxic. We thus randomly select a fraction p of non toxic comments, change their label to toxic, and perform the core analysis of the study on the modified data. We consider $p_w = \{1, 3, 5, 7, 9, 11, 13, 15\}\%$. For each one of those values, the analysis is repeated for a large number N_s of random selections of the set of comments considered wrong. The only dataset that is sensitive to a small but non-negligible probability of error $(1-2$

comments every 100) is Usenet conspiracy . However, in this case the slopes of the linear regression of all simulated trends were always positive.

8.2.6 Participation and toxicity in short conversations

Our analysis of the interplay between users' participation in a conversation and its toxicity content is focused in capturing dynamics pertaining to participated and/or long-lasting discussions. However, one may argue that restricting the analysis to long threads may result in leaving out discussion that end very early due to toxicity, thus pre-selecting the cases in which users are not immediately scared away by a highly toxic comment happening towards the beginning of a conversation. To rule out this possibility, we took all short conversations composed of 6 to 20 comments and, for each dataset, computed the Kernel Density Estimation of the toxicity score of the last three comments in the thread, and of all the other comments preceding them (e.g. for a 15 comment thread we take the mean toxicity score of the first 12 and that of the last 3). As shown in Figure 9.18 and 9.19 in Appendix, for each dataset the pair of densities are almost indistinguishable, meaning that, on average, in short conversations the last comments are not significantly more toxic than those preceding them, indicating that the potential effects mentioned above do not undermine our conclusions. Regarding our analysis of longer threads, we notice here that the participation quantity can give rise to similar trends in various cases. For instance, high participation can be achieved because many users take part in the conversation, but also with small groups of users in which everyone is equally contributing over time. Or, in very large discussions the contributions of individual outliers may remain hidden. By measuring participation these and other borderline cases may not be distinct from the statistically highly more likely discussion dynamics, but ultimately this lack of discriminative power does not have any implications on our findings nor on the validity of the conclusions we draw.

8.2.7 Data availability

Facebook, Twitter and YouTube data are made available in accordance with their respective terms of use. IDs of comments used in this analysis are provided at Open Science Framework (https://doi.org/10.17605/osf.io/fq5dy). For the remaining platforms (Gab, Reddit, Telegram, Usenet, Voat), all the necessary information to recreate the datasets used in this study can be found in section “Data Collection".

8.3 Results

To obtain a comprehensive picture of online social media conversations, we analyze a dataset of more than 500 million posts and comments from eight major platforms, covering diverse topics and spanning over three decades (see Table 8.1). This dataset offers heterogeneity and enables a historical comparison with Usenet, a precursor of modern social media. We restrict our analysis of Usenet to the period before 2006 when Facebook, Twitter, and YouTube emerged.

This section provides a general overview of online conversations related to our topic of interest, focusing on user activity and thread size metrics. We define a conversation (or a thread) as a sequence of comments that follow an initial post chronologically. We observe that, across all platforms, both user activity (Figure 8.1a) and thread size (Figure 9.10a in Appendix) exhibit heavy-tailed distributions with similar shapes.

One of the common features of emergent collective behavior in large complex systems is the presence of heavy-tailed distributions [323]. This phenomenon has been observed in various studies on social media platforms.. Indeed, our analysis reveals that the emergent macroscopic patterns of online conversations, such as the distribution of users’ and threads’ activity and lifetime, are robust and invariant across different social network settings (e.g., moderation policies, population size, historical context) and conversation topics.

We analyze user activity within a conversation across all platforms by defining a measure of participation during the evolution of a thread. For this purpose, we only take threads composed of many comments greater than a threshold that depends on the dataset and ensures sufficient length and divide each of them by a fixed number of equal intervals, each containing the same number of chronologically ordered comments (i.e., 0-5% of the thread, 5-10%, etc.). To compute the measure, for each of these intervals, we calculate the ratio of the number of unique users to the number of comments present in the interval and average it over all the considered threads. Smaller values of participation thus indicate that fewer unique users are producing the same number of comments in a segment of the conversations. In contrast, values close to one mean that the homogeneity of users activity is maximal. We find that across all datasets, the participation of users through the evolution of conversations is almost always clearly or slightly decreasing, meaning that as a conversation goes on, fewer users tend to dwell on it, but more actively (see Figure 9.11 in Appendix). Regarding shape and values, trends in participation for the different topics considered are consistent for each platform, except for Telegram and Usenet, in which users' participation shows a higher dependence on the conversation scope. Togliere il paragrafo qua sotto e mettere un'altra figura al posto della 1b (da giustificare in rebuttal) (forse mettere proprio le persistence). Interestingly, though the user's participation exhibits similar evolution, it shows distinctive trends towards the beginning and end of conversations, while appearing approximately linear otherwise. Thus, in order to better visualize this peculiar behavior, which changes depending on the platform, we subtracted participation from its linear component, as shown in Figure 8.1b.

8.3.1 Conversation size and toxicity

We now focus on the toxicity of online conversations. To detect the use of toxic language, we rely on Google's Perspective API [286], which is currently the most suitable toxicity classifier and is being extensively used in recent literature [280, 279, 51, 324, 325] As anticipated, Perspective API defines as toxic "A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion". Based on this definition, the classifier assigns a toxicity score in the [0,1] range to a piece of text that can be interpreted as an estimate of the likelihood that someone will perceive the text as toxic. We assign this toxicity score to all comments in our dataset and label them as toxic (non-toxic) if they are above (below) a score equal to 0.6. Perspective suggests this threshold value, especially for automatic moderation purposes, because it is meant to guarantee, with a good margin that more than half the readers would perceive the comment as toxic. Also, notice that this threshold can be considered conservative, as a study involving human toxicity labelers found that a slightly smaller threshold for their case of interest was suitable [326]. As previously mentioned, using an automatic detector allows for large-scale analysis at the cost of losing some of the nuances of toxic speech due to adhering to a specific definition of toxicity and the limited ability to consider the conversational context. In sec. Toxicity Detection and Validation of Perspective API we discuss at length these issues, and we also show that the findings presented in the analysis do not change qualitatively if different tools and datasets are employed or if a lower binary toxicity threshold is used.

Each dataset's percentage of toxic comments is reported in Table 8.1. This value is always below 10% except for the conspiracy channel of Telegram and for all Voat datasets, which show a remarkably higher presence of toxic comments. This is arguably due to the absence of automatic content moderation policies for the latter. However, one can notice that on platforms that are similar in this respect, toxicity may vary widely depending on the topic of discussion (Telegram) and be comparable with more moderated platforms when considering the entire discussion feed (Gab). Considering the activity of users, the fraction of users that post only or almost only toxic comments is extremely low for each dataset (in the range between 10^{-3} and 10^{-4}). The number of users versus the fraction of toxic comments they posted in a dataset decreases sharply, in an exponential fashion. The same holds for the fraction of toxic comments within conversations, even though these distributions appear to vary more among platforms and topics (the complementary cumulative distribution functions are shown in Figure 9.12 in Appendix).

One feature that can make online conversations very different is their size. It is reasonable to expect that a discussion's dynamics and complexity can depend on its length - which in online conversations is also obviously correlated with the number of people participating. Indeed, a recent study showed the importance of grouping conversation cascades by size to analyze their structural differences properly [327]. We decide to follow a similar approach and, for each conversation, we consider its size, defined as the number of comments it comprises and the percentage of toxic comments. For each dataset, we group conversations by size through a logarithmic binning with a fixed number of intervals (see sec. Logarithmic binning and conversation sizes) and plot the

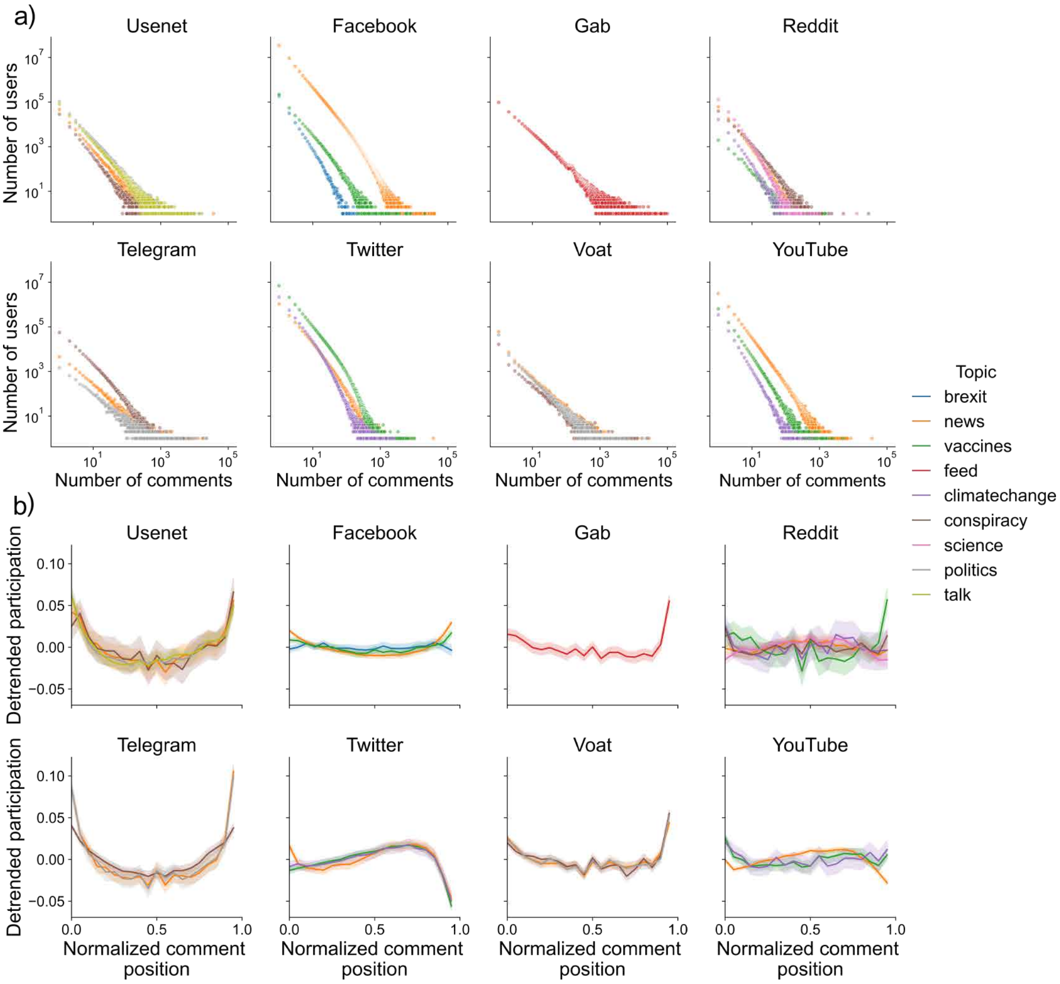


Figure 8.1. General characteristics of online conversations. a. Distributions of user activity in terms of comments posted for each platform and each topic (color-coded legend on the side). **b.** Detrended mean participation of users along threads. Trends are reported with their 95% confidence intervals. To better visualize pattern similarities, each trend has been subtracted its mean value. The x-axis represents the normalized position of comment intervals in the threads.

trend of the mean fraction of toxic comments in threads versus the thread length intervals. As shown in Figure 8.2, the resulting trends are almost all increasing, showing that independently of platform, time, or topic, the longer the conversation, the more toxic it tends to be. We checked the increase of the trends by both performing linear regression and applying the Mann-Kendall test – a non-parametric test assessing the presence of a monotonic upward or downward tendency – to ensure the statistical significance of our results (see Table 9.10). To further validate these outcomes, we shuffled the toxicity labels of the comments, finding that when data are randomized, trends are almost always non-increasing. Of the three toxicity trends that are statistically ambiguous according to the Mann-Kendall's test (Reddit news, Reddit vaccines, and Usenet talk), only the first has a slope that is not significantly different from the mean of the distribution of slopes resulting from randomizations, the others being much more than five standard deviations away from it; the only decreasing trend is Usenet alt.politics. In addition, we checked that the results are not sensitive to the chosen number of size intervals in the logarithmic binning by computing the same trends for different values to find that they do not qualitatively change. The results of this analysis are summarized in Table 9.10 in Appendix, and for the extended set of data (see Appendix) in Figure 9.14 and Table 9.18. Finally, we studied the toxicity content of conversations versus their lifetime - i.e., the time elapsed between the original post and the last comment. However, in this case, most trends are flat, and there is no indication that toxicity is generally associated with the duration of a conversation (see Figure 9.13 in Appendix).

8.3.2 Conversation evolution and toxicity

In the previous sections, we examined the toxicity level of online conversations after they ended. In this section, we explore how toxicity evolves in a conversation and how it affects the discussion dynamics.

The perception that online interactions deteriorate over time into toxic exchanges is widespread today and, perhaps surprisingly, was so even in the early days of the World Wide Web [328, 329]. Similarly, one should expect that when unbearable levels of toxicity are reached, the conversation is effectively over and should be terminated. This is coherent with the definition of toxic language adopted by Perspective API, which implies that it reduces the chance of a person staying in a conversation.

Although these generally accepted assumptions cannot be incontrovertibly verified, quantitative evidence can be derived to test their validity. In order to do that, we proceed as we did to calculate participation, i.e., we take sufficiently long threads and divide each of them by a fixed number of equal intervals,

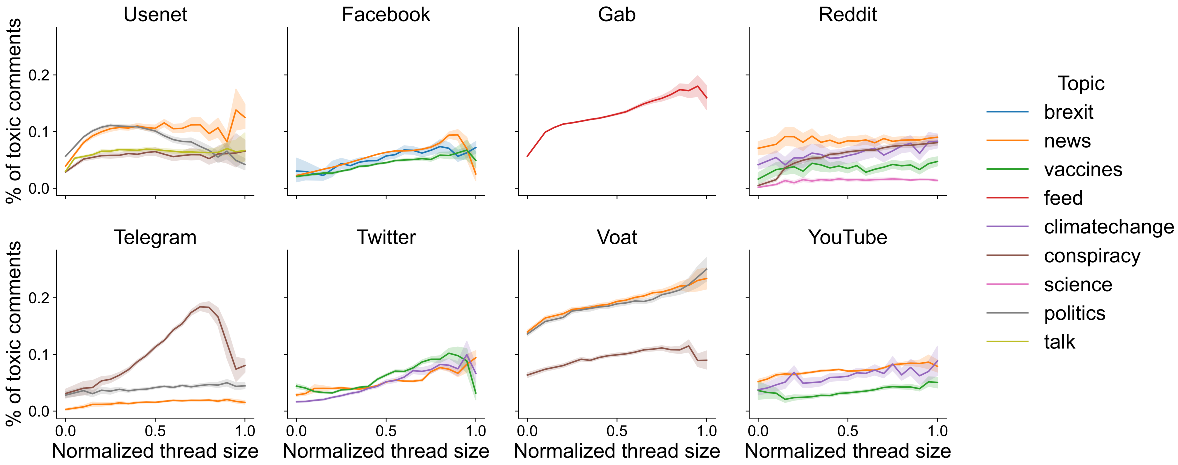


Figure 8.2. Toxicity increases with conversation size. Fraction of toxic comments in conversations versus conversation size, for each dataset (color-coded legend on the side). Trends represent the mean toxicity over each size interval and their 95% confidence interval. Size ranges are normalized to allow for a visual comparison of the different trends.

compute the percentage of toxic comments for each of these intervals, average it over all threads, and plot the trend in toxicity through the unfolding of the conversations. We find that the average toxicity level is mostly stable and does not show a general distinctive behavior across all datasets around the final part of threads (see Figure 8.3a) and c), and Figure 9.11 in Appendix). They indicate that toxicity does not always increase toward the end of a conversation. Notice that a similar observation was made in [328], but referring only to Reddit. Furthermore, the reasonable, intuitive assumption that toxicity is likely to lead people to leave a conversation does not seem correct, despite the fact that this feature is also part of the definition of toxicity on which the detector tool employed here is based. This can be seen by checking the relationship between trends in user participation - a quantity related to the number of users in a discussion at some point - and toxicity. The fact that the former almost always decreases, while the latter remains stable during conversations (see Figure 8.3c) indicates that toxicity does not affect participation in conversations, thus suggesting that people - on average - drop out of discussions regardless of the toxicity of the exchanges. We calculated Pearson's correlation between user participation and toxicity trends for each dataset to support this hypothesis. As shown in Figure 8.3b, the resulting correlation coefficient values are very heterogeneous, thus not indicating the presence of the expected strong anti-correlation between the two trends.

We compared user participation in regular and toxic conversations to validate this analysis further. To separate these two sets, we calculated the toxicity distribution T_i of long threads in each dataset i - where, as usual, toxicity stands for a fraction of toxic comments in a thread, and we labeled as toxic conversations whose toxicity exceeds $t_i = \mu(T_i) + \sigma(T_i)$, with $\mu(T_i)$ being mean and $\sigma(T_i)$ the standard deviation of T_i; all the other conversations are considered regular. Then, for each dataset, we computed user participation in toxic and non-toxic threads and calculated the Pearson's correlation between these pairs of trends to find strongly positive and robust correlations in all cases (Figure 8.3e). In addition, the pairwise differences between the slopes of the participation trends resulting from linear regressions are very close to zero, indicating an absence of significant differences in the rate at which conversations are abandoned (Figure 8.3f). Therefore, user behavior in toxic and non-toxic conversations shows almost identical patterns in terms of participation. This reinforces our finding that toxicity, on average, does not appear to impact the likelihood of people leaving a conversation.

Ultimately, despite the limitations of these classifiers, what we found highlights an unexpected feature of online toxicity dynamics, providing insights into how users interact in the face of toxic language. The analyses presented here and in the previous section were repeated with a lower toxicity classification threshold (0.5), to show that the outcomes are robust to other reasonable choices of this parameter (see Figure 9.17 and Table 9.20 in Appendix). Results were further validated with additional datasets (see Appendix for further information).

8.3.3 Controversy and toxicity

We explore the intricate dynamics of online conversations, focusing on the interplay between toxicity and engagement. Our exploration centers on two

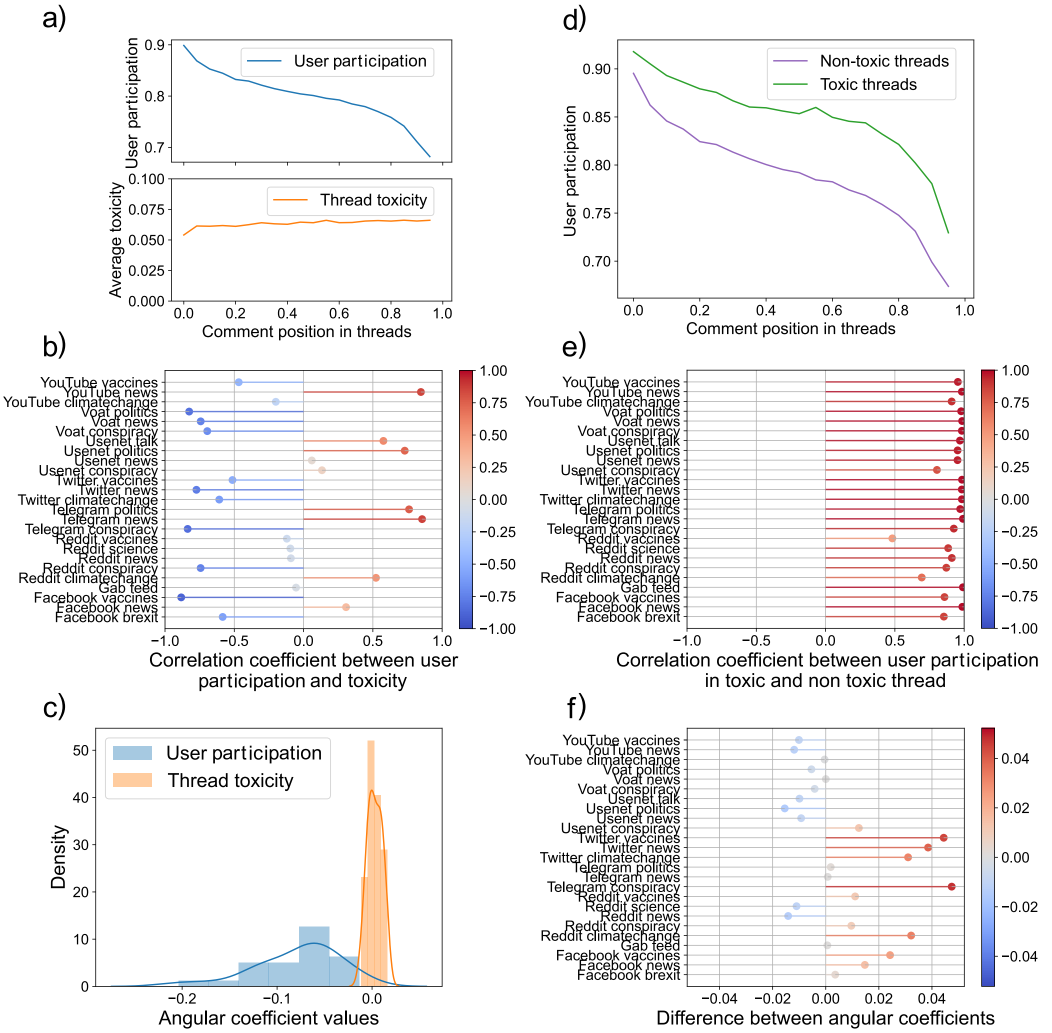


Figure 8.3. Participation of users is not dependent on toxicity. a. Examples of a typical trend in averaged user participation (top) and toxicity (bottom) versus normalized position of comment intervals in the threads (Twitter News dataset). **b.** Pearson's correlation coefficients between user participation and toxicity trends for each dataset. **c.** Density distribution of toxicity and participation trend slopes, as resulting from linear regression. **d.** An example of user participation in toxic and non-toxic thread sets (Twitter News dataset). **e.** Pearson's correlation coefficients between users participation in toxic and non-toxic thread sets, for each dataset. **f.** Difference between toxic and non-toxic thread sets participation slopes resulting from linear regression.

fundamental questions: what compels individuals to remain in toxic conversations, and why toxic exchanges are more frequent in longer conversations?

These questions are not easily answered, as online interactions are influenced by a multitude of factors that interact in complex ways. As such, we adopt an exploratory approach, formulating hypotheses based on the features of the conversations in our dataset.

One factor that may influence toxicity and engagement is the controversy surrounding the topic under discussion. Controversial topics may give rise to more heated and protracted debates and an increase in toxic language. In this view, toxicity is not a cause but a consequence of controversy. Pursuing this line of inquiry, we identified proxies for the level of controversy in conversations and examined their correlation with toxicity and conversation size. Concurrently, we investigated the relationship between toxicity and engagement.

For our initial investigation, we operate under the assumption that controversy is more likely to emerge when individuals with opposing views engage in active discourse. Exploiting the peculiarities of our data, we can infer the political leanings of a subset of users in the Facebook News, Twitter News, Twitter Vaccines, and Gab Feed datasets. This was achieved by examining the endorsements expressed toward news outlets whose political inclinations have been independently assessed. Table 9.13 in Appendix shows a breakdown of the datasets. As a result, those users were labeled with a leaning score $l \in [-1, 1]$, -1 being left and $+1$ right-leaning. We then selected threads with at least 10 different labeled users, in which at least 10% of comments (with a minimum of 20) were produced by labeled users and assigned to each of these comments the same leaning score of those who posted them. In this setting, the level of controversy within a conversation is assumed to be captured by a measure accounting for the spread of the political leaning of the participants and their activity in the conversation, in our case, the standard deviation of the distribution $\sigma(l)$ of comments possessing a leaning score: the higher $\sigma(l)$, the greater the level of debate and controversy in a thread. We observe that our measure of controversy is based on ideological polarization, understood here as a division along a political spectrum that is supposed to reflect the degree of disagreement of conceptions, views, and attitudes toward a wide range of socially relevant issues.

We analyzed the relationship between controversy and toxicity in online conversations of different sizes. Figure 8.4.a shows that controversy increases with the size of conversations in all datasets, and its trends are highly correlated with the corresponding trends in toxicity (see also Table 9.13 in Appendix). This supports our hypothesis that controversy and toxicity are closely related in online discussions. To validate this result, we used another measure of controversy previously reported in the literature: the standard deviation of the sentiment $\sigma(s)$. Indeed, in [317] controversial discussions were shown to exhibit a clearly higher variance of sentiment expressed in comments, which is an indication of a stronger level of debate. Note that this quantity is informative in that it measures the breadth of the sentiment spectrum in a conversation, a characteristic that cannot be correlated with toxicity, unlike, for example, the simple amount of negative sentiment present in a discussion. We performed sentiment analysis using a dictionary-based method [330] on all threads containing at least ten comments and found that $\sigma(s)$ also increases significantly with conversation size in almost all datasets (see Table 9.11 in Appendix). We applied randomization to our data to rule out any possible systematic effect that could affect the observed trends.

These outcomes indicate that controversy can be a factor in lengthening conversations, but, even though less likely, in principle, so does toxicity. If so, one might expect that toxic comments generated more engagement on average than non-toxic ones. Using the mean number of likes/upvotes, as a proxy of engagement, we have an indication that this may not be the case. In Figure 8.4b), we can see that the trend in likes/upvotes versus toxicity of comments is never increasing past the toxicity score threshold (0.6).

Finally, to complement our analysis, we inspected the relationship between toxicity and users' activity within conversations, measured as a level of density of comments in time. To do so, we employed a robust method for activity burst detection [321] that, upon reconstructing the density profile of a temporal stream of elements, separates the stream into different levels of activity and assigns each element to the level to which it belongs. Focusing on each conversation's two highest activity levels, we calculated the fraction of toxic comments $f_t \in [0, 1]$ before, during, and after the peak of activity. By comparing the distributions of the number of conversations versus f_t for the three intervals, we found that these distributions are statistically different from one another in almost all cases (see Table 9.14 in Appendix and Figure 8.4c). In all datasets

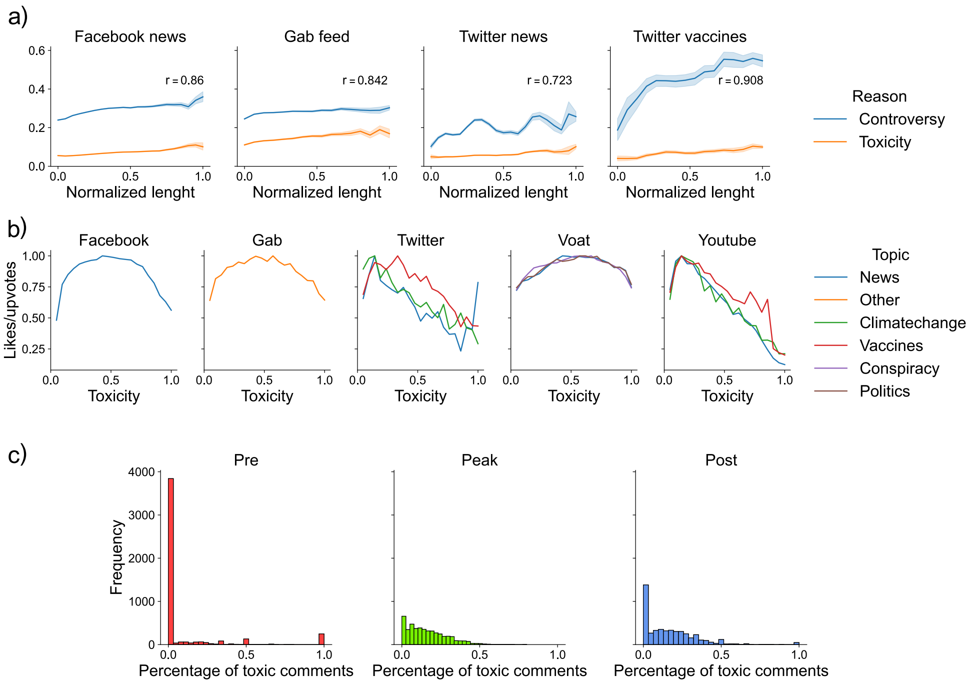


Figure 8.4. Controversy and toxicity in conversations a. Mean controversy ($\overline{\sigma}(l)$) and toxicity versus thread size (log-binned and normalized) for the Facebook News, Twitter News, Twitter Vaccines, and Gab Feed datasets. Here toxicity is calculated in the same conversations in which controversy could be computed (see Table 9.13 in Appendix). Pearson's correlation coefficients r between controversy and toxicity trends are reported in the plots; in Table 9.13 in Appendix the relative Spearman's and Kendall's correlation coefficients can be found too. Trends are reported with their 95% confidence interval. **b.** Likes/upvotes versus toxicity (binned). **c.** An example (Voat politics dataset) of distribution of frequency of toxic comments in threads before (left), at (center), and after (right) the peak of activity (i.e. density of comments posted in time).

but one, distributions are consistently shifted towards higher toxicity at the peak of activity, compared with the previous phase. These results suggest that toxicity is likely a consequence of increased engagement of users in a conversation rather than a contributing feature.

8.4 Discussion

This study delves into one of the most prominent and persistent characteristics of online discussions: toxic behavior, defined here as rude, disrespectful, or unreasonable conduct. Our analysis suggests that toxicity is neither a deterrent to user involvement nor an engagement amplifier; rather, it tends to emerge when exchanges become more frequent and may be a product of opinion polarization. Our findings suggest that the polarization of users' opinion - intended as the degree of opposed partisanship of users in a conversation - may play a more crucial role than toxicity in shaping the evolution of online discussions. Of course, other factors may influence toxicity and engagement, such as the specific subject of the conversation, the presence of influential users or 'trolls', the time and day of posting, as well as cultural or demographic aspects, such as users' average age or geographic location. However, when people encounter views that contradict their own, they may react with hostility and contempt. Our analysis suggests that this, in turn, may create a cycle of negative emotions and behaviors that fuels toxicity. Additionally, we show that some online conversation dynamics have mainly remained consistent over the past three decades, despite the evolution of social media platforms and norms.

Our study has some limitations that we acknowledge and discuss. First, we use political leaning as a proxy for general leaning, which may only capture some of the nuances of online opinions. However, political leaning represents a broad spectrum of opinions across different topics, and it correlates well with other dimensions of leaning, such as news preferences, vaccine attitudes, and stance on climate change [331, 25]. We could not assign a political leaning to users to analyze controversies on all platforms. Still, those considered - Facebook, Gab, and Twitter - represent different populations and moderation policies, and the combined data accounts for nearly 90% of the contents in our entire dataset. We used sentiment analysis to strengthen our analysis of the controversy, which we know has limitations. Nevertheless, the outcomes of the randomization tests we performed corroborate the meaningfulness and statistical robustness of the observed trends. We remark that we studied toxicity using a binary threshold, which does not allow us to distinguish between different levels of toxicity or different linguistic features. Our analysis approach is based on breadth and heterogeneity. As such, it may raise concerns about potential reductionism due to the comparison of different data sets from different sources and time periods. We acknowledge that each discussion thread, platform, and context has its unique characteristics and complexities that might be diminished when homogenizing data. However, we aim not to capture the full depth of every discussion but to identify and highlight general patterns and trends in online toxicity across platforms and time. The quantitative approach employed in our study enables us to uncover these overarching principles and patterns that may otherwise remain hidden in the individual complexities of each dataset. Finally, we observe that we can only provide descriptive insights into the dynamics governing online platforms without establishing clear causal relationships or making normative judgments.

Notwithstanding these limitations, the study significantly contributes to our comprehension of online social behavior. Through the new and extensive dataset presented here, critical aspects of the online platform ecosystem and fundamental dynamics of users' interactions can be explored. In addition, we have provided insights that a comparative approach such as the one followed here can prove invaluable in discerning the degree to which platform feed algorithms impact the online dynamics that characterize other sensitive issues, such as the formation of polarization and the spread of misinformation. The resulting outcomes have multiple potential impacts. Revealing persistent toxicity patterns across different platforms and time periods can better inform moderation policies and practices on social media platforms. In fact, our results suggest that toxic behavior online is not merely a product of the specific features or rules of individual platforms but is a broader issue that requires a comprehensive, cross-platform approach. Furthermore, the participation of users in toxic conversations suggests that a simple approach to removing toxic comments may not be sufficient to prevent users' exposure to such content. This indicates a need for more sophisticated moderation techniques to manage conversations' dynamics, including early interventions in discussions that show signs of becoming toxic. In addition, our findings offer evidence that automatic toxicity detection models trained using data from one platform may be relevant to other platforms due to the observed homogeneity. However, this should be carefully validated for each case due to potential user demographics and platform features variations. Future works may delve deeper into the role of controversy and the interplay between that and other factors that can give rise to toxicity. It would also be valuable to examine the impact of different moderation practices on the evolution of toxic discussions.

Chapter 9

Final remarks and future direction

In this thesis, we described our research work focusing on the interplay between misinformation and toxicity in online conversations. In the first part of the thesis, we provided an overview of the role of online misinformation in the spreading of conspiracy theories among regulated and unregulated social media. Then, we performed a quantitative comparison between Twitter and Gab to assess the effect of moderation policies concerning the debate around the COVID-19 pandemic. In the second part, instead, we focused on conversations in the digital landscape. We initially investigated how language in online conversations changed due to the introducing of new forms of expressions, such as visual memes. Then, we analyzed the engagement dynamics that these contents produce in relationship to their virality, providing a framework of how users react to newer elements concerning their diffusion. Finally, we end the research by investigating toxicity dynamics that these reactions may arise in online conversations, together with the design of solutions that mitigate the antisocial behaviors that may arise from these debates.

Our results show how moderation plays an undeniable role in contrasting the spreading of misinformation, creating moderated ecosystems where reliable content is predominant, whilst its absence nourishes the circulation of questionable news, corroborated by the echo chamber effect. We then observed how these echo systems are characterized by new forms of expressions, employing visual memes as representative, whose popularity, as well as entropy and complexity of the element proposed, have been dramatically increased since their first appearance on Reddit. Despite this, we observe that when the popularity of content is achieved relatively quickly from its first appearance, the responding audience may lack accordance in the sentiment expressed to these new elements. On the contrary, when the popularity of contents, and therefore topics, is obtained with more steady growth, the corresponding audience produces more homogeneous reactions. Lastly, by focusing on the conversations controversial reactions may induce under-posted content, we observe that toxicity grows as the debate is prolonged, providing evidence of how polarization is one of the main drivers of toxicity. The experiments conducted to limit online toxicity unveiled how accounting for the different stages that a conversation enter as it evolves, together with its topology, produce comparable results with the current state of the art despite the limited number of feature requested.

The different aspects that characterize this study raise a series of research questions that must be addressed. On a general note, the concept of polarization employed in the study still needs a proper measure that may be applied regardless of the topic, platform and interactions in the exam. Furthermore, the results reported in the study are confined to the online world due to the need for real-world data concerning these effects. However, the results of conspiracy theories in the online world have been observed more frequently. To this extent, further advancements in closing the gap between the online and offline world will benefit society.

From a conspiracy perspective, current research may benefit from longitudinal studies that characterize the evolution of conspiracy theories to improve the current moderation policies to provide more timely enforcement in mitigating the diffusion of these narratives. Moreover, the concept of echo platforms and their implications in online society is relatively new and must be investigated. From a toxicity perspective, instead, future studies may introduce the controversy produced by topics through their engagement and growth to identify the content more likely to spark heated debates. These insights may also be applied to the moderation of these conversations to contrast online toxicity, possibly introducing architectural advancements of machine learning models that contribute to more effective and ethical moderation.

Bibliography

[1] S. Sagiroglu and D. Sinanc, “Big data: A review,” in *2013 international conference on collaboration technologies and systems (CTS)*. IEEE, 2013, pp. 42–47.

[2] “Organization, w. h. director-general’s remarks at the media briefing on 2019 novel coronavirus on 8 february 2020.” [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf?sfvrsn=195f4010_6

[3] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, “The echo chamber effect on social media,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, p. e2023301118, 2021.

[4] C. R. Sunstein, “The law of group polarization,” *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, no. 91, 1999.

[5] M. Cinelli, A. Pelicon, I. Mozetič, W. Quattrociocchi, P. K. Novak, and F. Zollo, “Dynamics of online hate and misinformation,” *Scientific Reports*, vol. 11, no. 1, p. 22083, 2021.

[6] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, “Echo chambers: Emotional contagion and group polarization on facebook,” *Scientific reports*, vol. 6, no. 1, p. 37825, 2016.

[7] K. H. Jamieson and J. N. Cappella, *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

[8] R. K. Garrett, “Echo chambers online?: Politically motivated selective exposure among internet news users,” *Journal of computer-mediated communication*, vol. 14, no. 2, pp. 265–285, 2009.

[9] F. Zollo, A. Bessi, M. Del Vicario, A. Scala, G. Caldarelli, L. Shekhtman, S. Havlin, and W. Quattrociocchi, “Debunking in a world of tribes,” *PloS one*, vol. 12, no. 7, 2017.

[10] V. Gómez, H. J. Kappen, and A. Kaltenbrunner, “Modeling the structure and evolution of discussion cascades,” in *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, 2011, pp. 181–190.

[11] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, “Detection of harassment on web 2.0,” *Proceedings of the Content Analysis in the WEB*, vol. 2, no. 0, pp. 1–7, 2009.

[12] A. G. Chowdhury, R. Sawhney, R. Shah, and D. Mahata, “# youtoo? detection of personal recollections of sexual harassment on social media,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2527–2537.

[13] A. Quattrociocchi, G. Etta, M. Avalle, M. Cinelli, and W. Quattrociocchi, “Reliability of news and toxicity in twitter conversations,” in *Social Informatics*, F. Hopfgartner, K. Jaidka, P. Mayr, J. Jose, and J. Breitsohl, Eds. Cham: Springer International Publishing, 2022, pp. 245–256.

[14] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Anyone can become a troll: Causes of trolling behavior in online discussions,” in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 1217–1230.

[15] J. Hannan, “Trolling ourselves to death? social media and post-truth politics,” *European Journal of Communication*, vol. 33, no. 2, pp. 214–226, 2018.

[16] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Antisocial behavior in online discussion communities,” in *Proceedings of the inter-*

national aaai conference on web and social media, vol. 9, no. 1, 2015, pp. 61–70.

[17] M. Cinelli, A. Pelicon, I. Mozetič, W. Quattrociocchi, P. K. Novak, and F. Zollo, “Dynamics of online hate and misinformation,” *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[18] A. Bessi and E. Ferrara, “Social bots distort the 2016 us presidential election online discussion,” *First monday*, vol. 21, no. 11-7, 2016.

[19] A. Bovet and H. A. Makse, “Influence of fake news in twitter during the 2016 us presidential election,” *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[20] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on twitter during the 2016 u.s. presidential election,” *Science*, vol. 363, no. 6425, pp. 374–378, 2019. [Online]. Available: https://science.sciencemag.org/content/363/6425/374

[21] “Who director-general’s opening remarks at the media briefing on covid-19 - 11 march 2020.”

[22] U. Nations, “What is climate change?” 2023. [Online]. Available: https://www.un.org/en/climatechange/what-is-climate-change

[23] V. of Humanity, “Conflict trends in 2023: A growing threat to global peace.” [Online]. Available: https://www.visionofhumanity.org/conflict-trends-in-2023-a-growing-threat-to-global-peace/

[24] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, “The covid-19 social media infodemic,” *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[25] M. Falkenberg, A. Galeazzi, M. Torricelli, N. Di Marco, F. Larosa, M. Sas, A. Mekacher, W. Pearce, F. Zollo, W. Quattrociocchi *et al.*, “Growing polarization around climate change on social media,” *Nature Climate Change*, pp. 1–8, 2022.

[26] F. Pierri, L. Luceri, N. Jindal, and E. Ferrara, “Propaganda and misinformation on facebook and twitter during the russian invasion of ukraine,” in *Proceedings of the 15th ACM Web Science Conference 2023*, 2023, pp. 65–74.

[27] J. Grimmelmann, “The virtues of moderation,” *Yale JL & Tech.*, vol. 17, p. 42, 2015.

[28] C. Kiene and B. M. Hill, “Who uses bots? a statistical analysis of bot usage in moderation teams,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–8. [Online]. Available: https://doi.org/10.1145/3334480.3382960

[29] M. Horta Ribeiro, J. Cheng, and R. West, “Automated content moderation increases adherence to community guidelines,” in *Proceedings of the ACM Web Conference 2023*, ser. WWW ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2666–2676. [Online]. Available: https://doi.org/10.1145/3543507.3583275

[30] J. Seering, R. Kraut, and L. Dabbish, “Shaping pro and anti-social behavior on twitch through moderation and example-setting,” in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 111–125.

[31] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[32] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, “Fairness and abstraction in sociotechnical systems,” in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 59–68.

[33] N. TeBlunthuis, B. M. Hill, and A. Halfaker, “Effects of algorithmic flagging on fairness: quasi-experimental evidence from wikipedia,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–27, 2021.

[34] L. Wang and H. Zhu, “How are ml-based online content moderation systems actually used? studying community size, local activity, and disparate treatment,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 824–838.

[35] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[36] T. Gebru, “Race and gender,” *The Oxford handbook of ethics of aI*, pp. 251–269, 2020.

[37] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, “Saving face: Investigating the ethical concerns of facial recognition auditing,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 145–151.

[38] “Gdelt 2.0: Our global world in realtime,” https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/, Accessed 2023.

[39] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann *et al.*, “Computational social science.” *Science (New York, NY)*, vol. 323, no. 5915, pp. 721–723, 2009.

[40] D. Schoch, C.-h. Chan, C. Wagner, and A. Bleier, “Computational reproducibility in computational social science,” *arXiv preprint arXiv:2307.01918*, 2023.

[41] E. Bakshy, S. Messing, and L. A. Adamic, “Exposure to ideologically diverse news and opinion on facebook,” *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[42] B. Owens, “Replication failures in psychology not due to differences in study populations,” *Nature*, vol. 19, 2018.

[43] M. Cinelli, E. Brugnoli, A. L. Schmidt, F. Zollo, W. Quattrociocchi, and A. Scala, “Selective exposure shapes the facebook news diet,” *PloS one*, vol. 15, no. 3, p. e0229129, 2020.

[44] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[45] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Science vs conspiracy: Collective narratives in the age of misinformation," *PloS one*, vol. 10, no. 2, p. e0118093, 2015.

[46] M. Del Vicario, F. Zollo, G. Caldarelli, A. Scala, and W. Quattrociocchi, "Mapping social dynamics on facebook: The brexit debate," *Social Networks*, vol. 50, pp. 6–16, 2017.

[47] A. L. Schmidt, F. Zollo, M. Del Vicario, A. Bessi, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "Anatomy of news consumption on facebook," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. 3035–3039, 2017.

[48] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy on social media," *ACM Transactions on Social Computing*, vol. 1, no. 1, pp. 1–27, 2018.

[49] P. Barberá, "Social media, echo chambers, and political polarization," p. 34, 2020.

[50] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[51] A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal, "Identifying toxicity within youtube video comment," in *Social, Cultural, and Behavioral Modeling*. Springer International Publishing, 2019, pp. 214–223.

[52] M. Saveski, B. Roy, and D. Roy, "The structure of toxic conversations on twitter," in *Proceedings of the Web Conference 2021*, 2021, pp. 1086–1097.

[53] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of covid-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, 2021.

[54] G. Bonaccorsi, F. Pierri, M. Cinelli, A. Flori, A. Galeazzi, F. Porcelli, A. L. Schmidt, C. M. Valensise, A. Scala, W. Quattrociocchi *et al.*, "Economic and social consequences of human mobility restrictions under covid-19," *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15 530–15 535, 2020.

[55] C. Team. (Accessed 2023) Crowdtangle. https://www.crowdtangle.com/.

[56] CrowdTangle, "Crowdtangle," 2023. [Online]. Available: https://help.crowdtangle.com/en/articles/3873721-crowdtangle-search-faq

[57] D. Paresh, "Reddit is already on the rebound," 2023.

[58] J. Barnes, "Twitter ends its free api: Here's who will be affected," 2023.

[59] Twitter, "Academic research program." [Online]. Available: https://developer.twitter.com/en/solutions/academic-research

[60] A. Mekacher and A. Papasavva, "" i can't keep it up anymore." the voat.co dataset," *arXiv preprint arXiv:2201.05933*, 2022.

[61] YouTube, "Youtube data api." [Online]. Available: https://developers.google.com/youtube/v3

[62] S. Zannettou, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringini, and J. Blackburn, "What is gab: A bastion of free speech or an alt-right echo chamber," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 1007–1014.

[63] E. Spafford, "The usenet," in *The User's Directory of Computer Networks*. Elsevier, 1990, pp. 386–390.

[64] I. Archive, "Usenet directory listing." [Online]. Available: https://archive.org/download/usenet-net

[65] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," *EPJ Data Science*, vol. 8, no. 1, pp. 1–13, 2019.

[66] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, “The structural virality of online diffusion,” *Management Science*, vol. 62, no. 1, pp. 180–196, 2016.

[67] M. E. Newman, “Mixing patterns in networks,” *Physical review E*, vol. 67, no. 2, p. 026126, 2003.

[68] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, “The spreading of misinformation online,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.

[69] C. R. Sunstein, “Democracy and filtering,” *Communications of the ACM*, vol. 47, no. 12, pp. 57–59, 2004.

[70] C. A. Vermeule and C. R. Sunstein, “Conspiracy theories: causes and cures,” *Journal of Political Philosophy*, 2009.

[71] K. M. Douglas, J. E. Uscinski, R. M. Sutton, A. Cichocka, T. Nefes, C. S. Ang, and F. Deravi, “Understanding conspiracy theories,” *Political Psychology*, vol. 40, pp. 3–35, 2019.

[72] S. C. Briand, M. Cinelli, T. Nguyen, R. Lewis, D. Prybylski, C. M. Valensise, V. Colizza, A. E. Tozzi, N. Perra, A. Baronchelli *et al.*, “Infodemics: A new challenge for public health,” *Cell*, vol. 184, no. 25, pp. 6010–6014, 2021,
* The article surveys the issue of infodemics and investigates the limits of representing information spreading as a viral process.

[73] S. van der Linden, “Misinformation: susceptibility, spread, and interventions to immunize the public,” *Nature Medicine*, pp. 1–8, 2022,
** The study investigates the drivers of misinformation spreading on social network and countermeasures to boost psychological immunity to misinformation.

[74] A. Acerbi, “Cognitive attraction and online misinformation,” *Palgrave Communications*, vol. 5, no. 1, pp. 1–7, 2019.

[75] Y. Zhang, L. Wang, J. J. Zhu, and X. Wang, “Conspiracy vs science: A large-scale analysis of online discussion cascades,” *World wide web*, vol. 24, no. 2, pp. 585–606, 2021.

[76] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[77] M. Conti, D. Lain, R. Lazzeretti, G. Lovisotto, and W. Quattrociocchi, “It’s always april fools’ day!: On the difficulty of social network misinformation classification via propagation features,” in *2017 IEEE Workshop on Information Forensics and Security (WIFS)*. IEEE, 2017, pp. 1–6.

[78] J. L. Juul and J. Ugander, “Comparing information diffusion mechanisms by matching on cascade size,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 46, 2021,
* The article deeply investigate statistical differences between information cascades of reliable and unreliable news.

[79] J.-W. van Prooijen, J. Ligthart, S. Rosema, and Y. Xu, “The entertainment value of conspiracy theories,” *British Journal of Psychology*, vol. 113, no. 1, pp. 25–48, 2022.

[80] A. L. Schmidt, F. Zollo, A. Scala, C. Betsch, and W. Quattrociocchi, “Polarization of the vaccination debate on facebook,” *Vaccine*, vol. 36, no. 25, pp. 3606–3612, 2018.

[81] G. Etta, M. Cinelli, A. Galeazzi, C. M. Valensise, M. Conti, and W. Quattrociocchi, “News consumption and social media regulations policy,” *arXiv preprint arXiv:2106.03924v1*, 2021.

[82] E. Brugnoli, M. Cinelli, W. Quattrociocchi, and A. Scala, “Recursive patterns in online echo chambers,” *Scientific Reports*, vol. 9, no. 1, pp. 1–18, 2019.

[83] D. A. Broniatowski, J. Gu, A. M. Jamison, and L. C. Abroms, “Evaluating the efficacy of facebook’s vaccine misinformation content removal policies,” *arXiv preprint arXiv:2202.02172*, 2022.

[84] K.-C. Yang, F. Pierri, P.-M. Hui, D. Axelrod, C. Torres-Lugo, J. Bryden, and F. Menczer, "The covid-19 infodemic: Twitter versus facebook," *Big Data & Society*, vol. 8, no. 1, p. 20539517211013861, 2021.

[85] M. T. Bastos and D. Mercea, "The brexit botnet and user-generated hyperpartisan news," *Social science computer review*, vol. 37, no. 1, pp. 38–54, 2019.

[86] M. Stella, E. Ferrara, and M. De Domenico, "Bots increase exposure to negative and inflammatory content in online social systems," *Proceedings of the National Academy of Sciences*, vol. 115, no. 49, pp. 12 435–12 440, 2018.

[87] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[88] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi, "Coordinated behavior on social media in 2019 uk general election," *arXiv preprint arXiv:2008.08370*, 2020.

[89] M. Falkenberg, A. Galeazzi, M. Torricelli, N. Di Marco, F. Larosa, M. Sas, A. Mekacher, W. Pearce, F. Zollo, W. Quattrociocchi *et al.*, "Growing climate polarisation on social media," *arXiv preprint arXiv:2112.12137*, 2021.

[90] M. Cinelli, S. Cresci, A. Galeazzi, W. Quattrociocchi, and M. Tesconi, "The limited reach of fake news on twitter during 2019 european elections," *PloS one*, vol. 15, no. 6, p. e0234689, 2020.

[91] J. Flamino, A. Galeazzi, S. Feldman, M. W. Macy, B. Cross, Z. Zhou, M. Serafino, A. Bovet, H. A. Makse, and B. K. Szymanski, "Shifting polarization and twitter news influencers between two us presidential elections," *arXiv preprint arXiv:2111.02505*, 2021.

[92] E. Ferrara, "What types of covid-19 conspiracies are populated by twitter bots?" *arXiv preprint arXiv:2004.09531*, 2020.

[93] M. Faddoul, G. Chaslot, and H. Farid, "A longitudinal analysis of youtube's promotion of conspiracy videos," *arXiv preprint arXiv:2003.03318*, 2020.

[94] C. Buntain, R. Bonneau, J. Nagler, and J. A. Tucker, "Youtube recommendations and effects on sharing across online social platforms," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–26, 2021,

* The article reports on the effect of content moderation highlighting cross-sharing behavior on multiple social media platforms.

[95] D. Röchert, G. Neubaum, B. Ross, and S. Stieglitz, "Caught in a networked collusion? homogeneity in conspiracy-related discussion networks on youtube," *Information Systems*, vol. 103, p. 101866, 2022.

[96] A. Bessi, F. Zollo, M. Del Vicario, M. Puliga, A. Scala, G. Caldarelli, B. Uzzi, and W. Quattrociocchi, "Users polarization on facebook and youtube," *PloS one*, vol. 11, no. 8, p. e0159641, 2016.

[97] N. Di Marco, M. Cinelli, and W. Quattrociocchi, "Infodemics on youtube: Reliability of content and echo chambers on covid-19," 2021. [Online]. Available: https://arxiv.org/abs/2106.08684

[98] H. Hosseinmardi, A. Ghasemian, A. Clauset, M. Mobius, D. M. Rothschild, and D. J. Watts, "Examining the consumption of radical content on youtube," *Proceedings of the National Academy of Sciences*, vol. 118, no. 32, 2021.

[99] S. Wu and P. Resnick, "Cross-partisan discussions on youtube: Conservatives talk to liberals but liberals don't talk to conservatives," in *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*, vol. 15, 2021.

[100] G. De Francisci Morales, C. Monti, and M. Starnini, "No echo in the chambers of political interactions on reddit," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[101] M. Samory and T. Mitra, "'the government spies using our webcams': The language of conspiracy theories in online discussions," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, nov 2018.

[102] S. Phadke, M. Samory, and T. Mitra, "Characterizing social imaginaries and self-disclosures of dissonance in online conspiracy discussion communities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–35, 2021.

[103] ——, "What makes people join conspiracy communities? role of social factors in conspiracy engagement," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW3, pp. 1–30, 2021.

[104] M. Aliapoulios, E. Bevensee, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, and S. Zannettou, "A large open dataset from the parler social network," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 943–951.

[105] A. Papasavva, J. Blackburn, G. Stringhini, S. Zannettou, and E. D. Cristofaro, ""is it a qoincidence?": An exploratory study of qanon on voat," in *Proceedings of the Web Conference 2021*, 2021, pp. 460–471.

[106] M. Hoseini, P. Melo, F. Benevenuto, A. Feldmann, and S. Zannettou, "On the globalization of the qanon conspiracy theory through telegram," *arXiv preprint arXiv:2105.13020*, 2021.

[107] A. Sipka, A. Hannak, and A. Urman, "Comparing the language of qanon-related content on parler, gab, and twitter," *arXiv preprint arXiv:2111.11118*, 2021.

[108] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," *Scientific Reports*, vol. 10, no. 1, Oct 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-73510-5

[109] W. H. Organization, "Coronavirus disease (covid-19) pandemic." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019?adgroupsurvey={adgroupsurvey}

[110] A. Bessi, F. Zollo, M. Del Vicario, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Trend of narratives in the age of misinformation," *PloS one*, vol. 10, no. 8, 2015.

[111] W.-Y. S. Chou, A. Oh, and W. M. Klein, "Addressing health-related misinformation on social media," *Jama*, vol. 320, no. 23, pp. 2417–2418, 2018.

[112] J. Zarocostas, "How to fight an infodemic," *The lancet*, vol. 395, no. 10225, p. 676, 2020.

[113] T. Sharot and C. R. Sunstein, "How people decide what they want to know," *Nature Human Behaviour*, vol. 4, no. 1, pp. 14–19, Jan 2020. [Online]. Available: https://doi.org/10.1038/s41562-019-0793-1

[114] L. Kim, S. M. Fast, and N. Markuzon, "Incorporating media data into a model of infectious disease transmission," *PLOS ONE*, vol. 14, no. 2, pp. 1–13, 02 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0197646

[115] C. Viboud and A. Vespignani, "The future of influenza forecasts," *Proceedings of the National Academy of Sciences*, vol. 116, no. 8, pp. 2802–2804, 2019. [Online]. Available: https://www.pnas.org/content/116/8/2802

[116] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Transactions on the Web (TWEB)*, vol. 13, no. 2, pp. 1–22, 2019.

[117] M. Cinelli, M. Conti, L. Finos, F. Grisolia, P. K. Novak, A. Peruzzi, M. Tesconi, F. Zollo, and W. Quattrociocchi, "(mis)information operations: An integrated perspective," *Journal of Information Warfare*, vol. 18, no. 3, pp. 83–98, 2019. [Online]. Available: https://www.jstor.org/stable/26894683

[118] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proceedings of the National Academy of*

Sciences, vol. 113, no. 3, pp. 554–559, 2016. [Online]. Available: https://www.pnas.org/content/113/3/554

[119] J. Berger and H. Perez, “Occasional paper the islamic state’s diminishing returns on twitter: How suspensions are limiting the social networks of english-speaking isis supporters,” *GW Program on Extremism and El Akkad, Omar (2012)“Why Twitter’s censorship plan is better than you think”. The Globe and Mail. Berger, Morgan (2015)“Defining and describing the population of ISIS supporters on Twitter”. The Brookings Institution*, 2016.

[120] S. Hughes and L. Vidino, “Isis in america: From retweets to raqqa,” *Program on Extremism, George Washington University.[online] https://cchs. gwu. edu/sites/cchs. gwu. edu/files/downloads/ISIS (Rev. 03.03. 2016)*, 2015.

[121] A. A. Siegel, “Online hate speech,” *Social Media and Democracy*, p. 56, 2019.

[122] M. B. Check, “Mbfc.” [Online]. Available: https://mediabiasfactcheck.com/

[123] N. Technologies, “Ng.” [Online]. Available: https://www.newsguardtech.com/

[124] E. Chen, K. Lerman, and E. Ferrara, “Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set,” *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e19273, 2020.

[125] E. Chen, K. Lerman, and E. Ferrara, “Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set - accounts,” 2020. [Online]. Available: https://github.com/echen102/COVID-19-TweetIDs/blob/master/accounts.txt

[126] E. Chen, K. Lerman, and E. Ferrara, “Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus

twitter data set - keywords," 2020. [Online]. Available: https://github.com/echen102/COVID-19-TweetIDs/blob/master/keywords.txt

[127] L. Bozarth, A. Saraf, and C. Budak, "Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 us presidential nominees," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 48–59.

[128] G. Weld, M. Glenski, and T. Althoff, "Political bias and factualness in news sharing across more than 100,000 online communities," *arXiv preprint arXiv:2102.08537*, 2021.

[129] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[130] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, vol. 54, no. 3, pp. 426–482, 1943. [Online]. Available: http://www.jstor.org/stable/1990256

[131] Twitter, "Updates to our work on covid-19 vaccine misinformation." [Online]. Available: https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.html

[132] J. Weedon, W. Nuland, and A. Stamos, "Information operations and facebook," *Retrieved from Facebook: https://fbnewsroomus. files. wordpress. com/2017/04/facebook-and-information-operations-v1. pdf*, 2017.

[133] R. Peto and J. Peto, "Asymptotically efficient rank invariant test procedures," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 2, pp. 185–207, 1972.

[134] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in Big Data*, vol. 2, p. 13, 2019.

[135] M. Glenski, T. Weninger, and S. Volkova, "Identifying and understanding user reactions to deceptive and trusted social news sources," *arXiv preprint arXiv:1805.12032*, 2018.

[136] A. L. Schmidt, F. Zollo, A. Scala, C. Betsch, and W. Quattrociocchi, “Polarization of the vaccination debate on facebook,” *Vaccine*, vol. 36, no. 25, pp. 3606–3612, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0264410X18306601

[137] C. R. Sunstein, *# republic.* Princeton University Press, 2017.

[138] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, “Exposure to opposing views on social media can increase political polarization,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9216–9221, 2018.

[139] R. Dawkins, *28. The Selfish Gene.* Princeton University Press, 2014, pp. 140–142. [Online]. Available: https://doi.org/10.1515/9781400848393-029

[140] T. W. Deacon, “Editorial: Memes as signs: The trouble with memes (and what to do about it).” *Semiotic Review of Books*, vol. 10, no. 3, pp. 1–3, 1999.

[141] T. A. Sebeok and M. Danesi, *The Forms of Meaning.* DE GRUYTER, 12 2000. [Online]. Available: https:/doi.org/10.1515/9783110816143

[142] S. Cannizzaro, “Internet memes as internet signs: A semiotic view of digital culture,” *Sign Syst. Stud.*, vol. 44, no. 4, pp. 562–586, 12 2016. [Online]. Available: https://doi.org/10.12697%2Fsss.2016.44.4.05

[143] I. Fomin, “Memes, genes, and signs: Semiotics in the conceptual interface of evolutionary biology and memetics,” *Semiotica*, vol. 2019, no. 230, pp. 327–340, 10 2019. [Online]. Available: https:/doi.org/10.1515/sem-2018-0016

[144] K. Distin, *The Selfish Meme.* Cambridge University Press, Dec. 2004. [Online]. Available: https://doi.org/10.1017/cbo9780511614286

[145] J. Leskovec, L. Backstrom, and J. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and*

data mining - KDD '09. ACM Press, 2009. [Online]. Available: https://doi.org/10.1145/1557019.1557077

[146] D. Ienco, F. Bonchi, and C. Castillo, "The meme ranking problem: Maximizing microblogging virality," in *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 2010, pp. 328–335.

[147] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics," in *Proceedings of the 20th international conference on World wide web - WWW '11*. ACM Press, 2011. [Online]. Available: https://doi.org/10.1145/1963405.1963503

[148] C. Bauckhage, "Insights into internet memes," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, 2011.

[149] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Truthy," *Proceedings of the 20th international conference companion on World wide web - WWW '11*, 2011. [Online]. Available: http://dx.doi.org/10.1145/1963192.1963301

[150] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention," *Scientific Reports*, vol. 2, no. 1, Mar. 2012. [Online]. Available: https://doi.org/10.1038/srep00335

[151] L. Weng, F. Menczer, and Y.-Y. Ahn, "Predicting successful memes using network and community structure," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014.

[152] E. Ferrara, M. JafariAsbagh, O. Varol, V. Qazvinian, F. Menczer, and A. Flammini, "Clustering memes in social media," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, Aug. 2013. [Online]. Available: https://doi.org/10.1145/2492517.2492530

[153] M. Coscia, "Average is boring: How similarity kills a meme's success," *Scientific Reports*, vol. 4, no. 1, Sep. 2014. [Online]. Available: https://doi.org/10.1038/srep06477

[154] A. Dang, A. Moh'd, A. Gruzd, E. Milios, and R. Minghim, "A visual framework for clustering memes in social media," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, Aug. 2015. [Online]. Available: https://doi.org/10.1145/2808797.2808830

[155] O. Tsur and A. Rappoport, "Don't let me be# misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, no. 1, 2015.

[156] L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng, "Information evolution in social networks," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, Feb. 2016. [Online]. Available: https://doi.org/10.1145/2835776.2835827

[157] A. Dubey, E. Moro, M. Cebrian, and I. Rahwan, "Memesequencer: Sparse matching for embedding image macros," 2018.

[158] S. Zannettou, T. Caulfield, J. Blackburn, E. D. Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil, "On the origins of memes by means of fringe web communities," 2018.

[159] D. M. Beskow, S. Kumar, and K. M. Carley, "The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning," *Information Processing & Management*, vol. 57, no. 2, p. 102170, Mar. 2020. [Online]. Available: https://doi.org/10.1016/j.ipm.2019.102170

[160] M. Perc, "Beauty in artistic expressions through the eyes of networks and physics," *Journal of The Royal Society Interface*, vol. 17, no. 164, p. 20190686, Mar. 2020. [Online]. Available: https://doi.org/10.1098/rsif.2019.0686

[161] H. Y. D. Sigaki, M. Perc, and H. V. Ribeiro, "History of art paintings through the lens of entropy and complexity," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. E8585–E8594, Aug. 2018. [Online]. Available: https://doi.org/10.1073/pnas.1800083115

[162] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, “The anatomy of reddit: An overview of academic research,” in *Dynamics On and Of Complex Networks III*. Springer International Publishing, 2019, pp. 183–204. [Online]. Available: https://doi.org/10.1007%2F978-3-030-14683-2_9

[163] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, “Scan: Learning to classify images without labels,” in *Proceedings of the European Conference on Computer Vision*, 2020.

[164] R. J. G. B. Campello, D. Moulavi, and J. Sander, “Density-based clustering based on hierarchical density estimates,” in *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2013, pp. 160–172. [Online]. Available: https://doi.org/10.1007/978-3-642-37456-2_14

[165] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, “A high-bias, low-variance introduction to machine learning for physicists,” *Physics Reports*, vol. 810, pp. 1–124, 5 2019. [Online]. Available: https://doi.org/10.1016/j.physrep.2019.03.001

[166] L. Rokach and O. Maimon, “Clustering methods,” in *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag, pp. 321–352. [Online]. Available: https://doi.org/10.1007/0-387-25465-x_15

[167] C. Bandt and B. Pompe, “Permutation entropy: A natural complexity measure for time series,” *Physical Review Letters*, no. 17, Apr. 2002. [Online]. Available: https://doi.org/10.1103/physrevlett.88.174102

[168] R. López-Ruiz, H. Mancini, and X. Calbet, “A statistical measure of complexity,” *Physics Letters A*, vol. 209, no. 5-6, pp. 321–326, Dec. 1995. [Online]. Available: https://doi.org/10.1016/0375-9601(95)00867-5

[169] H. V. Ribeiro, L. Zunino, E. K. Lenzi, P. A. Santoro, and R. S. Mendes, “Complexity-entropy causality plane as a complexity measure for two-dimensional patterns,” *PLoS ONE*, vol. 7, no. 8, p. e40689, Aug. 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0040689

[170] L. Zunino and H. V. Ribeiro, “Discriminating image textures with the multiscale two-dimensional complexity-entropy causality plane,” *Chaos, Solitons & Fractals*, vol. 91, pp. 679–688, Oct. 2016. [Online]. Available: https://doi.org/10.1016/j.chaos.2016.09.005

[171] O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes, “Distinguishing noise from chaos,” *Physical Review Letters*, vol. 99, no. 15, Oct. 2007. [Online]. Available: https://doi.org/10.1103/physrevlett.99.154102

[172] T. Yasseri, P. Gildersleve, and L. David, “Collective memory in the digital age,” *arXiv preprint arXiv:2207.01042*, 2022.

[173] G. Lazaroiu, “The role of social media as a news provider,” *Review of Contemporary Philosophy*, vol. 13, pp. 78–84, 2014.

[174] A. N. Ahmad, “Is twitter a useful tool for journalists?” *Journal of Media Practice*, vol. 11, no. 2, pp. 145–155, 2010.

[175] D. Notarmuzi, C. Castellano, A. Flammini, D. Mazzilli, and F. Radicchi, “Universality, criticality, and complexity of information propagation in social media,” *Nature Communications*, vol. 13, no. 1, pp. 1–8, 2022.

[176] J. Brown, A. J. Broderick, and N. Lee, “Word of mouth communication within online communities: Conceptualizing the online social network,” *Journal of interactive marketing*, vol. 21, no. 3, pp. 2–20, 2007.

[177] R. Kahn and D. Kellner, “New media and internet activism: from the ‘battle of seattle’to blogging,” *New media & society*, vol. 6, no. 1, pp. 87–95, 2004.

[178] S. C. McGregor, “Social media as public opinion: How journalists use social media to represent public opinion,” *Journalism*, vol. 20, no. 8, pp. 1070–1086, 2019.

[179] R. Jaakonmäki, O. Müller, and J. V. Brocke, “The impact of content, context, and creator on user engagement in social media marketing,” in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[180] P. M. D. Gangi and M. M. Wasko, "Social media engagement theory: Exploring the influence of user engagement on social media usage," *Journal of Organizational and End User Computing (JOEUC)*, vol. 28, no. 2, pp. 53–73, 2016.

[181] H. A. Simon, "Designing organizations for an information-rich world," *Computers, Communications, and the Public Interest*, vol. 72, p. 37, 1971.

[182] S. C. Kies, "Social media impact on attention span," *Journal of Management & Engineering Integration*, vol. 11, no. 1, pp. 20–27, 2018.

[183] K. Holt, A. Shehata, J. Strömbäck, and E. Ljungberg, "Age and the effects of news media attention and social media use on political interest and participation: Do social media function as a leveller?" *European Journal of Communication*, vol. 28, no. 1, pp. 19–34, 2013.

[184] S. Brooks, "Does personal social media usage affect efficiency and well-being?" *Computers in Human Behavior*, vol. 46, pp. 26–37, 2015.

[185] S. Flaxman, S. Goel, and J. M. Rao, "Filter bubbles, echo chambers, and online news consumption," *Public Opinion Quarterly*, vol. 80, no. S1, pp. 298–320, 2016.

[186] J. A. Cookson, J. Engelberg, and W. Mullins, "Echo chambers," *The Review of Financial Studies*, vol. 36, no. 2, pp. 450–500, 2023.

[187] J. T. Klapper, *The Effects of Mass Communication*, 1960.

[188] A. Bessi, A. Scala, L. Rossi, Q. Zhang, and W. Quattrociocchi, "The economy of attention in the age of (mis)information," *Journal of Trust Management*, vol. 1, no. 1, pp. 1–13, 2014.

[189] D. Mocanu, L. Rossi, Q. Zhang, M. Karsai, and W. Quattrociocchi, "Collective attention in the age of (mis)information," *Computers in Human Behavior*, vol. 51, pp. 1198–1204, 2015.

[190] G. Etta, A. Galeazzi, J. R. Hutchings, C. S. J. Smith, M. Conti, W. Quattrociocchi, and G. V. D. Riva, "Covid-19 infodemic on facebook and

containment measures in italy, united kingdom, and new zealand," *PloS One*, vol. 17, no. 5, p. e0267022, 2022.

[191] M. Falkenberg, A. Galeazzi, M. Torricelli, N. D. Marco, F. Larosa, M. Sas, A. Mekacher, W. Pearce, F. Zollo, W. Quattrociocchi, and A. Baronchelli, "Growing polarization around climate change on social media," *Nature Climate Change*, pp. 50–60, 2022.

[192] C. Candia, C. Jara-Figueroa, C. Rodriguez-Sickert, A.-L. Barabási, and C. A. Hidalgo, "The universal decay of collective memory and attention," *Nature Human Behaviour*, vol. 3, no. 1, pp. 82–91, 2019.

[193] C. M. Valensise, M. Cinelli, M. Nadini, A. Galeazzi, A. Peruzzi, G. Etta, F. Zollo, A. Baronchelli, and W. Quattrociocchi, "Lack of evidence for correlation between covid-19 infodemic and vaccine acceptance," 2021.

[194] F. Tahmasi, L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, ""go eat a bat, chang!": On the emergence of sinophobic behavior on web communities in the face of covid-19," in *Proceedings of the Web Conference 2021*, 2021, pp. 1122–1133.

[195] *Social Media and Democracy: The State of the Field, Prospects for Reform.* Cambridge University Press, 2020.

[196] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[197] C. Team. (Accessed 2023) Understanding and citing crowdtangle data. https://help.crowdtangle.com/en/articles/5103495-understanding-and-citing-crowdtangle-data.

[198] GDELT. (Accessed 2023) The gdelt project. https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/.

[199] K. Leetaru and P. A. Schrodt, "Gdelt: Global data on events, location, and tone, 1979–2012," in *ISA Annual Convention*, vol. 2, 2013, pp. 1–49.

[200] L. Ou-Yang. (Accessed 2023) Newspaper3k. https://newspaper.readthedocs.io/en/latest/.

[201] F. M. Bass, "A new product growth model for consumer durables," *Management Science*, vol. 15, no. 5, pp. 215–227, 1969.

[202] G. D. Tarde, *The Laws of Imitation*, 1903.

[203] E. M. Rogers, "New product adoption and diffusion," *Journal of Consumer Research*, vol. 2, no. 4, pp. 290–301, 1976.

[204] A. Grubler, *The Rise and Fall of Infrastructures: Dynamics of Evolution and Technological Change in Transport*, 1990.

[205] C. Perez, *Technological Revolutions and Financial Capital*, 2003.

[206] L. Robinson, *Changeology: How to Enable Groups, Communities and Societies to Do Things They've Never Done Before*, 2012.

[207] O. Kanjanatarakul and K. Suriya, "Comparison of sales forecasting models for an innovative agro-industrial product: Bass model versus logistic function," *The Empirical Econometrics and Quantitative Economics Letters*, vol. 1, no. 4, pp. 89–106, 2012.

[208] B. Spann, E. Mead, M. Maleki, N. Agarwal, and T. Williams, "Applying diffusion of innovations theory to social networks to understand the stages of adoption in connective action campaigns," *Online Social Networks and Media*, vol. 28, p. P100201, 2022.

[209] J. Beel, T. Xiang, S. Soni, and D. Yang, "Linguistic characterization of divisive topics online: Case studies on contentiousness in abortion, climate change, and gun control," *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 32–42, 2022.

[210] J. Hessel and L. Lee, "Something's brewing! early prediction of controversy-causing posts from discussion features," *arXiv preprint arXiv:1904.07372*, 2019.

[211] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, pp. 50–60, 1947.

[212] M. Herz and P. Molnár, *The content and context of hate speech: Rethinking regulation and responses.* Cambridge University Press, 2012.

[213] A. Guterres *et al.*, "United nations strategy and plan of action on hate speech," *Taken from: https://www. un. org/en/genocideprevention/documents/U*, no. 20Strategy, 2019.

[214] G. Jigsaw, "Perspective api," 2022. [Online]. Available: https://perspectiveapi.com/

[215] Jigsaw, "Toxicity," *The Current*, no. 3, 2023. [Online]. Available: https://jigsaw.google.com/the-current/toxicity/

[216] W. Quattrociocchi, G. Caldarelli, and A. Scala, "Opinion dynamics on interacting networks: media competition and social influence," *Scientific reports*, vol. 4, no. 1, p. 4938, 2014.

[217] J. I. Criado, R. Sandoval-Almazan, and J. R. Gil-Garcia, "Government innovation through social media," pp. 319–326, 2013.

[218] G. Etta, E. Sangiorgio, N. Di Marco, M. Avalle, A. Scala, M. Cinelli, and W. Quattrociocchi, "Characterizing engagement dynamics across topics on facebook," *Plos one*, vol. 18, no. 6, p. e0286150, 2023.

[219] C. E. Robertson, N. Pröllochs, K. Schwarzenegger, P. Pärnamets, J. J. Van Bavel, and S. Feuerriegel, "Negativity drives online news consumption," *Nature Human Behaviour*, pp. 1–11, 2023.

[220] A. Matamoros-Fernández and J. Farkas, "Racism, hate speech, and social media: A systematic review and critique," *Television & New Media*, vol. 22, no. 2, pp. 205–224, 2021.

[221] M. C. Parent, T. D. Gobble, and A. Rochlen, "Social media behavior, toxic masculinity, and depression." *Psychology of Men & Masculinities*, vol. 20, no. 3, p. 277, 2019.

[222] K. Saha, E. Chandrasekharan, and M. De Choudhury, "Prevalence and psychological effects of hateful speech in online college communities," in *Proceedings of the 10th ACM conference on web science*, 2019, pp. 255–264.

[223] C. M. Valensise, M. Cinelli, and W. Quattrociocchi, "The dynamics of online polarization," *arXiv preprint arXiv:2205.15958*, 2022.

[224] R. Rotabi, K. Kamath, J. Kleinberg, and A. Sharma, "Cascades: A view from audience," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 587–596.

[225] F. Zhou, X. Xu, G. Trajcevski, and K. Zhang, "A survey of information cascade analysis: Models, predictions, and recent advances," *ACM Comput. Surv.*, vol. 54, no. 2, mar 2021. [Online]. Available: https://doi.org/10.1145/3433000

[226] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, no. 1, 2013, pp. 1621–1622.

[227] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & internet*, vol. 7, no. 2, pp. 223–242, 2015.

[228] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 29–30.

[229] S. Malmasi and M. Zampieri, "Challenges in discriminating profanity from hate speech," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 2, pp. 187–202, 2018.

[230] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association

for Computational Linguistics, Jun. 2016, pp. 88–93. [Online]. Available: https://aclanthology.org/N16-2013

[231] E. Wulczyn, N. Thain, and L. Dixon, “Ex machina: Personal attacks seen at scale,” in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1391–1399.

[232] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[233] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval),” *arXiv preprint arXiv:1903.08983*, 2019.

[234] H. Almerekhi, H. Kwak, B. J. Jansen, and J. Salminen, “Detecting toxicity triggers in online discussions,” in *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, ser. HT ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 291–292. [Online]. Available: https://doi.org/10.1145/3342220.3344933

[235] A. Sheth, V. L. Shalin, and U. Kursuncu, “Defining and detecting toxicity on social media: context and knowledge are key,” *Neurocomputing*, vol. 490, pp. 312–318, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221018087

[236] J. M. Pérez, F. M. Luque, D. Zayat, M. Kondratzky, A. Moro, P. S. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano, and V. Cotik, “Assessing the impact of contextual information in hate speech detection,” *IEEE Access*, vol. 11, pp. 30 575–30 590, 2023.

[237] S. L. Wilson and C. Wiysonge, “Social media and vaccine hesitancy,” *BMJ global health*, vol. 5, no. 10, p. e004206, 2020.

[238] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, “You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech,” *Proc. ACM*

Hum.-Comput. Interact., vol. 1, no. CSCW, dec 2017. [Online]. Available: https://doi.org/10.1145/3134666

[239] S. Gonzalez-Bailon, A. Kaltenbrunner, and R. E. Banchs, "The structure of political discussion networks: A model for the analysis of online deliberation," *Journal of Information Technology*, vol. 25, no. 2, pp. 230–243, 2010. [Online]. Available: https://doi.org/10.1057/jit.2010.2

[240] L. Backstrom, J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil, "Characterizing and curating conversation threads: expansion, focus, volume, re-entry," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 13–22.

[241] R. Kumar, M. Mahdian, and M. McGlohon, "Dynamics of conversations," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 553–562. [Online]. Available: https://doi.org/10.1145/1835804.1835875

[242] S. Levy, R. E. Kraut, J. A. Yu, K. M. Altenburger, and Y.-C. Wang, "Understanding conflicts in online conversations," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2592–2602. [Online]. Available: https://doi.org/10.1145/3485447.3512131

[243] M. Raghavendra, K. Sharma, S. Kumar *et al.*, "Signed link representation in continuous-time dynamic signed networks," *arXiv preprint arXiv:2207.03408*, 2022.

[244] A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman, "A new generation of perspective api: Efficient multilingual character-level transformers," *arXiv preprint arXiv:2202.11176*, 2022.

[245] G. Jigsaw, "Perspective api - attributes & languages," 2022. [Online]. Available: https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

[246] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving google's perspective api built for detecting toxic comments," *arXiv preprint arXiv:1702.08138*, 2017.

[247] E. Jain, S. Brown, J. Chen, E. Neaton, M. Baidas, Z. Dong, H. Gu, and N. S. Artan, "Adversarial text generation for google's perspective api," in *2018 international conference on computational science and computational intelligence (CSCI)*. IEEE, 2018, pp. 1136–1141.

[248] G. Russo, L. Verginer, M. H. Ribeiro, and G. Casiraghi, "Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning," *arXiv preprint arXiv:2209.09803*, 2022.

[249] C. S. Czymara, S. Dochow-Sondershaus, L. G. Drouhot, M. Simsek, and C. Spörlein, "Catalyst of hate? ethnic insulting on youtube in the aftermath of terror attacks in france, germany and the united kingdom 2014–2017," *Journal of Ethnic and Migration Studies*, vol. 49, no. 2, pp. 535–553, 2023.

[250] "Digital bridge: Italian election — midterm polarization — android fallout," *Politico*, 2022. [Online]. Available: https://www.politico.eu/newsletter/digital-bridge/italian-election-midterm-polarization-android-fallout/

[251] B. IFIS, "Osservatorio sullo sport system italiano," 2022.

[252] T. Jones, "Beyond the violence, the shocking power the ultras wield over italian football," *The Guardian*, 2018.

[253] N. Squires, "Italian riot police clash with football fans on the rampage in naples," *The Telegraph*, 2023.

[254] Twitter, "Twitter api for academic research," 2023. [Online]. Available: https://developer.twitter.com/en/products/twitter-api/academic-research

[255] YouTube, "Youtube data api," 2023. [Online]. Available: https://developers.google.com/youtube/v3

[256] Jigsaw, “About the scores,” 2023. [Online]. Available: https://developers.perspectiveapi.com/s/about-the-api-score?language=en_US

[257] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, “Information diffusion through blogspace,” in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 491–501.

[258] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, “Can cascades be predicted?” in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 925–936.

[259] G. Pallis, D. Zeinalipour-Yazti, and M. D. Dikaiakos, “Online social networks: status and trends,” *New directions in web data management 1*, pp. 213–234, 2011.

[260] S. Greenwood, A. Perrin, and M. Duggan, “Social media update 2016,” *Pew Research Center*, vol. 11, no. 2, pp. 1–18, 2016.

[261] T. Aichner, M. Grünfelder, O. Maurer, and D. Jegeni, “Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019,” *Cyberpsychology, behavior, and social networking*, vol. 24, no. 4, pp. 215–222, 2021.

[262] S. González-Bailón, D. Lazer, P. Barberá, and et al., “Asymmetric ideological segregation in exposure to political news on facebook,” *Science*, vol. 381, no. 6656, pp. 392–398, 2023.

[263] A. Guess, N. Malhotra, J. Pan, and et al., “How do social media feed algorithms affect attitudes and behavior in an election campaign?” *Science*, vol. 381, no. 6656, pp. 398–404, 2023.

[264] ——, “Reshares on social media amplify political news but do not detectably affect beliefs or opinions,” *Science*, vol. 381, no. 6656, pp. 404–408, 2023.

[265] B. Nyhan, J. Settle, E. Thorson, and et al., “Like-minded sources on facebook are prevalent but not polarizing,” *Nature*, pp. 137–144, 2023.

[266] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, "Social media, political polarization, and political disinformation: A review of the scientific literature," *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

[267] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[268] Y. Lupu, R. Sear, N. Velásquez, R. Leahy, N. J. Restrepo, B. Goldberg, and N. F. Johnson, "Offline events and online hate," *PLOS ONE*, vol. 18, no. 1, pp. 1–14, 2023.

[269] R. K. Garrett, "Echo chambers online?: Politically motivated selective exposure among internet news users," *Journal of Computer-Mediated Communication*, vol. 14, no. 2, pp. 265–285, 2009.

[270] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Echo chambers: Emotional contagion and group polarization on facebook," *Scientific Reports*, vol. 6, no. 1, p. 37825, Dec 2016.

[271] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 913–922.

[272] N. F. Johnson and et al., "The online competition between pro-and anti-vaccination views," *Nature*, vol. 582, no. 7811, pp. 230–233, 2020.

[273] K. Saha, E. Chandrasekharan, and M. De Choudhury, "Prevalence and psychological effects of hateful speech in online college communities," vol. 2019, 2019.

[274] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, and H. M. H. López, "Internet, social media and online hate speech. systematic review," *Aggression and Violent Behavior*, vol. 58, p. 101608, 2021.

[275] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, 2018.

[276] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: A review," *IEEE Access*, vol. 9, pp. 88 364–88 376, 2021.

[277] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation (extended version)," *Information Systems*, vol. 105, p. 101584, 2022.

[278] H. Almerekhi, H. Kwak, and J. Jansen, "Investigating toxicity changes of cross-community redditors from 2 billion posts and comments," *PeerJ Computer Science*, vol. 8, p. e1059, 2022.

[279] Y. Xia, H. Zhu, T. Lu, P. Zhang, and N. Gu, "Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW2, 2020.

[280] A. Sipka, A. Hannak, and A. Urman, "Comparing the language of qanon-related content on parler, gab, and twitter," in *14th ACM Web Science Conference 2022*, ser. WebSci '22. Association for Computing Machinery, 2022, p. 411–421.

[281] S. Baele, L. Brace, and D. Ging, "A diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem," *Terrorism and Political Violence*, pp. 1–24, 2023.

[282] N. B. Noor, N. Yousefi, B. Spann, and N. Agarwal, "Comparing toxicity across social media platforms for covid-19 discourse," *arXiv preprint arXiv:2302.14270*, 2023.

[283] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of*

the International AAAI Conference on Web and Social Media, vol. 11, 2017.

[284] V. Kolhatkar, H. Wu, L. Cavasso, E. Francis, K. Shukla, and M. Taboada, "The sfu opinion and comments corpus: A corpus for the analysis of online news comments," *Corpus Pragmatics*, vol. 4, pp. 155–190, 2020.

[285] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a lexicon of abusive words–a feature-based approach," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1046–1056.

[286] A. Lees, V. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman, "A new generation of perspective api: Efficient multilingual character-level transformers," 08 2022, pp. 3197–3207.

[287] M. Castelle, "The linguistic ideologies of deep abusive language classification," in *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 2018, pp. 160–170.

[288] W. Yin and A. Zubiaga, "Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media," *Online Social Networks and Media*, vol. 30, p. 100210, 2022.

[289] M. Del Vicario, F. Zollo, G. Caldarelli, A. Scala, and W. Quattrociocchi, "Mapping social dynamics on facebook: The brexit debate," *Social Networks*, vol. 50, pp. 6–16, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378873316304166

[290] Facebook, "Facebook community standards," https://transparency.fb.com/policies/community-standards/hate-speech/, 03 2023.

[291] G. Rosen and T. Lyons, "Remove, reduce, inform: New steps to manage problematic content," https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/, 04 2019.

[292] YouTube, "Vulgar language policy," https://support.google.com/youtube/answer/10072685?, 03 2023.

[293] ——, "Harassment & cyberbullying policies," https://support.google.com/youtube/answer/2802268, 03 2023.

[294] ——, "Hate speech policy," https://support.google.com/youtube/answer/2801939, 03 2023.

[295] ——, "How does youtube enforce its community guidelines?" https://www.youtube.com/intl/en_us/howyoutubeworks/policies/community-guidelines/#enforcing-community-guidelines, 03 2023.

[296] Twitter, "The twitter rules," https://help.twitter.com/en/rules-and-policies/twitter-rules, 03 2023.

[297] ——, "Hateful conduct," https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, 02 2023.

[298] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, p. 2053951719897945, 2020.

[299] Twitter, "Our range of enforcement options," https://help.twitter.com/en/rules-and-policies/enforcement-options, 03 2023.

[300] V. Elliott and C. Stokel-Walker, "Twitter's moderation system is in tatters," http://bit.ly/3nFnUpd, 11 2022.

[301] Reddit, "Reddit content policy," https://www.redditinc.com/policies/content-policy, 03 2023.

[302] ——, "Promoting hate based on identity or vulnerability," https://www.reddithelp.com/hc/en-us/articles/360045715951, 03 2023.

[303] A. Malik, "Reddit acqui-hires team from ml content moderation startup oterlu," https://tcrn.ch/3yeS2Kd, 10 2022.

[304] Telegram, "Terms of service," https://telegram.org/tos, 03 2023.

[305] P. Durov, “The rules of @telegram prohibit calls for violence and hate speech. we rely on our users to report public content that violates this rule.” 10 2017. [Online]. Available: https://twitter.com/durov/status/917076707055751168?lang=en

[306] Telegram, “Telegram privacy policy,” https://telegram.org/privacy, 03 2023.

[307] GAB, “Terms of service,” https://gab.com/about/tos, 01 2023.

[308] C. Salzenberg and G. Spafford, “What is usenet?” https://www0.mi.infn.it/~calcolo/W_is_usenet.html, 11 1995.

[309] A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella, “Thirty years of research into hate speech: topics of interest and their evolution,” *Scientometrics*, vol. 126, pp. 157–179, 2021.

[310] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith, “Annotators with attitudes: How annotator beliefs and identities bias toxic language detection,” *arXiv preprint arXiv:2111.07997*, 2021.

[311] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos, “Toxicity detection: Does context really matter?” *arXiv preprint arXiv:2006.00998*, 2020.

[312] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, “The risk of racial bias in hate speech detection,” in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 1668–1678.

[313] L. Piedras, L. Rosenblatt, and J. Wilkins, “Critical perspectives: A benchmark revealing pitfalls in perspectiveapi,” *arXiv preprint arXiv:2301.01874*, 2023.

[314] P. Di Maggio, J. Evans, and B. Bryson, “Have americans’ social attitudes become more polarized?” *American Journal of Sociology*, vol. 102, no. 3, 1996.

[315] M. P. Fiorina and S. J. Abrams, "Political polarization in the american public," *Annual Review of Political Science*, vol. 11, no. 1, pp. 563–588, 2008.

[316] S. Iyengar, S. Gaurav, and Y. Lelkes, "Affect, not ideology: A social identity perspective on polarization," *The Public Opinion Quarterly*, vol. 76, no. 3, pp. 405–431, 2012.

[317] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy on social media," *Trans. Soc. Comput.*, vol. 1, no. 1, 2018.

[318] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," *EPJ Data Science*, vol. 8, no. 1, p. 35, Dec 2019. [Online]. Available: https://doi.org/10.1140/epjds/s13688-019-0213-9

[319] I. Himelboim, S. McCreery, and M. Smith, "Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter," *Journal of Computer-Mediated Communication*, vol. 18, no. 2, pp. 40–60, 2013.

[320] J. An, D. Quercia, and J. Crowcroft, "Partisan sharing: Facebook evidence and societal consequences," in *Proceedings of the 5th ACM International Conference on Communities and Technologies (COSN '14)*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 13–24.

[321] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining and Knowledge Discovery 7*, pp. 373–397, 2003.

[322] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics 18 (1)*, pp. 50–60, 1947.

[323] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, pp. 47–97, 2002.

[324] R. Debnath, D. Reiner, B. Sovacool, F. Muller-Hansen, T. Repke, R. Alvarez, and S. Fitzgerald, “Conspiracy spillovers and geoengineering,” *iScience*, vol. 26, pp. 1–25, 2023.

[325] J. Beel, T. Xiang, S. Soni, and D. Yang, “Linguistic characterization of divisive topics online: Case studies on contentiousness in abortion, climate change, and gun control,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, no. 1, pp. 32–42, 2022.

[326] M. Saveski, B. Roy, and D. Roy, “The structure of toxic conversations on twitter,” in *Proceedings of the Web Conference 2021*, ser. WWW ’21. Association for Computing Machinery, 2021, p. 1086–1097.

[327] J. L. Juul and J. Ugander, “Comparing information diffusion mechanisms by matching on cascade size,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 46, p. e2100786118, 2021.

[328] G. Fariello, D. Jemielniak, and A. Sulkowski, “Does godwin’s law (rule of nazi analogies) apply in observable reality? an empirical study of selected words in 199 million reddit posts,” *new media & society*, p. 14614448211062070, 2021.

[329] I. Bedzow, “Godwin’s law and the limits of bioethics and holocaust studies,” in *Bioethics and the Holocaust: A Comprehensive Study in How the Holocaust Continues to Shape the Ethics of Health, Medicine and Human Rights.* Springer International Publishing Cham, 2022, pp. 209–218.

[330] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014.

[331] D. Albrecht, “Vaccination, politics and covid-19 impacts,” *BMC Public Health*, vol. 22, no. 1, pp. 1–12, 2022.

[332] M. Newman, *Networks.* Oxford University Press, 2018.

[333] G. Jasser, J. McSwiney, E. Pertwee, and S. Zannettou, "'welcome to# gabfam': Far-right virtual community on gab," *New Media & Society*, p. 14614448211024546, 2021.

Appendix

Comparing the impact of social media regulations on news consumption

Here we provide further details about the datasets employed for the study, the cumulative and daily evolution of posts and new users posting and the numerical results of the statistical tests performed. The description of the dataset is reported in Section 9, the representation of the time series is reported in Section 9 while the results of the statistical test performed are described in Section 9.

Data breakdown

Here we report the composition of the dataset for Gab and Twitter employed for the study, described in Table 9.1 and Table 9.2 respectively. Each dataset represents the number of unique posts, users and comments, as well as their engagement quantities over the different data collection and processing steps. Due to Twitter API limitations, posts were initially gathered without any information about their number of comments. The collection of this quantity was performed after the categorization of the news outlets through the employment of specific APIs provided from Twitter as part of its Academic Research Program [59].

To provide an overview of the most frequent domains populating the two platforms, in Figure 9.3 and Figure 9.4 we show the top 50 domains by number of posts for Twitter and Gab respectively, computed from the datasets of posts containing both a search hashtag and a link (Column 3 of Table 9.1 and 9.2).

A list of datasets including news outlets and links from Gab and Twitter

Category	Overall	Containing search hashtags and link	Categorized	Questionable	Reliable
Number of Posts	205 458	130 864	83 784	49 772	34 012
Number of users	11 063	8 194	5 681	4 660	3 289
Number of Likes	234 255	117 281	72 435	53 154	19 281
Number of Reblogs	138 793	75 250	48 172	34 960	13 212
Number of Comments	42 993	22 287	14 165	9 489	4 676

Table 9.1. Data breakdown of posts collected on Gab.

Category	Overall	Containing search hashtags and link	Catego
Number of Posts	2 668 286	1 110 030	244 430
Number of users	1 185 541	382 449	118 635
Number of Likes	18 610 555	5 885 562	1 422 62
Number of Reblogs	7 753 971	2 862 098	703 765
Number of Comments	NA	NA	30 262

Table 9.2. Twitter Dataset Breakdown

with their categorization is available on GitHub [1].

Time series evolution

Here we report the evolution of posts and new users collected on Twitter and Gab during the analysis time. Figure 9.1 shows the overall evolution of such quantities, while Figure 9.2 shows the previous evolution after performed the categorization of the news outlets. Figure 9.1 describes a spike on Gab in the number of posts and users in correspondence of June. This was due to the change of the collecting method. Indeed, the lack of chronological order of posts reported on Gab APIs since June required the gathering of all posts from the general stream, which was then filtered by the search hashtags in order to be compliant with the data collection process.

Results of comparison between power law distributions

Here we report the information related to the comparison of the power law fits by means of the Wald test. Tables 9.5 - 9.6 report the estimated coefficients of each power law fit, i.e., $\hat{\alpha}$ and $\hat{x}_{\text{min}}$, depending on the engagement and news

[1] https://github.com/cdcslab/paper-social-media-regulations-policy

outlet category. Furthermore, the results of the Wald test score applied on the previous engagement categories are reported, together with the corresponding p-values.

Validation of the echo chamber phenomenon using community detection

Modularity maximization is an optimization problem that has the modularity function Q as objective function and the network as input [332]. The modularity Q quantifies the quality of a certain community partitioning by means of the following expression:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m})\, \delta(g_i, g_j) \in [-0.5, 1] \tag{9.1}$$

where g_i and g_j are two integers which label the community i and j, respectively, for $i, j = 1, \ldots, N$, $i \neq j$, and with $N \leq n$. The case $N = 1$ means that we have just one community containing all the nodes, while $N = n$ means that we have n communities, each of them with only one node. The Kronecker function $\delta(g_i, g_j)$ is one when $g_i = g_j$ and zero otherwise.

Given the definition of community – i.e., a group of nodes that are densely connected inside their group and sparsely connected to the rest of the network – the optimized value of modularity is considered to provide the most meaningful community partitioning – according to the prefixed criterion driving the construction of the adjacency matrix of the network. Indeed, such a value corresponds to a partitioning of the network in which the number of links among nodes belonging to the same community is substantially higher than the number of links among nodes belonging to different communities. Such an aspect is mathematically represented by the difference in Equation 9.1, that counts the actual number of links among nodes assigned to the same community versus its expected value.

Given a certain assignment of nodes into groups, expressed by the vector $\mathbf{g}$, the modularity value represents the deviation of the number of links among nodes of the same type – which is represented by $\sum_{ij} A_{ij}\delta(g_i, g_j)$ – from the expected number of links among such nodes, given their degree. Indeed, given two nodes with degree k_i and k_j respectively, the expected number of links

Table 9.3. Gab's Power Law fits of Post Consumption Patterns

Likes			**Reblogs**		
	$\hat{\alpha}$	$\hat{x}_{min}$		$\hat{\alpha}$	$\hat{x}_{min}$
Questionable	1.32	1	Questionable	1.31	1
Reliable	1.37	1	Reliable	1.33	1
Wald test score		0.22	Wald's test score		0.034
p-value		0.64	p-value		0.85

between i and j is k_i times $\frac{k_j}{2m}$, that is, k_i times the probability of being connected to j. The modularity function is normalized to range between -0.5 and 1. It assumes low values when there are less links than expected among nodes in the same group, whereas it assumes high values in the opposite case. We exploit a state-of-the-art community detection algorithm called the Louvain algorithm [196]. The algorithm follows an agglomerative greedy approach that optimizes modularity, firstly finding small agglomerates – i.e., communities – of nodes that provide the highest value of modularity; secondly, considering such agglomerates as single nodes in order to re-iterate the first step.

We applied the Louvain algorithm to the Twitter following network, obtaining 12 different communities. We then computed the average leaning of the individuals in each community, observing how the obtained clusters are generally characterized by consumption of reliable comments, providing support for the results in the analysis.

Comparison of US based users on Gab and Twitter

Gab is known as a social media platform whose user base is predominantly US-American [333]. The existence of this demographic structure may introduce some biases in the reflection of users' opinion dynamics between the two platforms. Therefore, we built a sample of US-based users on Twitter by extracting all those with a location containing the name of a US city or county. This sample is composed of 58171 posts from 27513 users. Then, we performed the analysis of the analysis on this reduced dataset, whose results are reported in Figure 9.6, 9.7, 9.8 and 9.9.

Table 9.4. Twitter's Power Law fits of Post Consumption Patterns

Likes			**Reblogs**		
	$\hat{\alpha}$	$\hat{x}_{min}$		$\hat{\alpha}$	$\hat{x}_{min}$
Questionable	1.52	1	Questionable	1.57	1
Reliable	1.55	1	Reliable	1.55	1
Wald test score		0.15	Wald's test score		0.05
p-value		0.70	p-value		0.83

Likes			**Reblogs**		
	$\hat{\alpha}$	$\hat{x}_{min}$		$\hat{\alpha}$	$\hat{x}_{min}$
Questionable	1.81	1	Questionable	1.83	1
Reliable	1.83	1	Reliable	1.83	1
Wald test score		0.015	Wald's test score		0.0004
p-value		0.90	p-value		0.98

Table 9.5. Gab's Power Law fits of User Consumption Patterns

Table 9.6. Twitter's Power Law fits of User Consumption Patterns

Likes			**Reblogs**		
	$\hat{\alpha}$	$\hat{x}_{min}$		$\hat{\alpha}$	$\hat{x}_{min}$
Questionable	3.33	1	Questionable	1.75	1
Reliable	2.21	1	Reliable	1.67	1
Wald test score		1286.34	Wald's test score		1.06
p-value		< 0.001	p-value		0.30

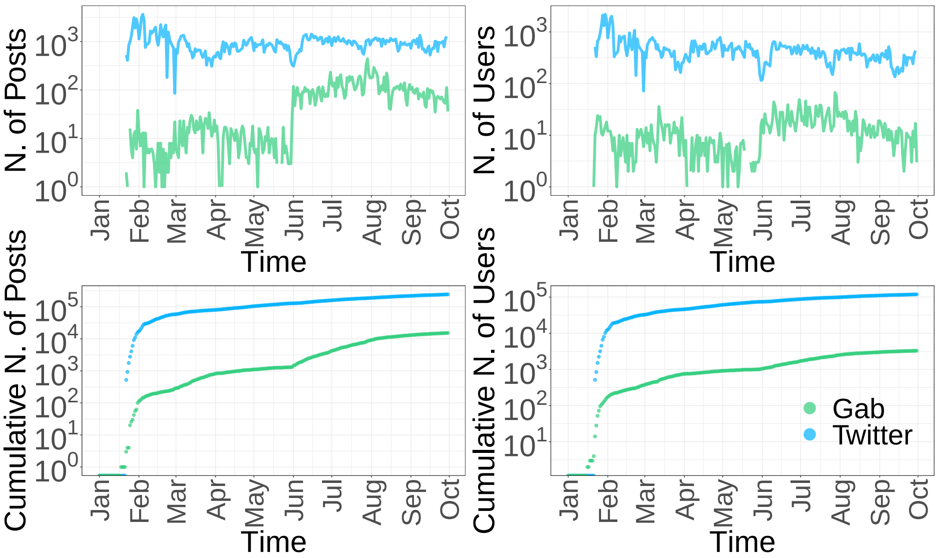


Figure 9.1. *Upper panel*: Time series evolution of new posts (left) and users posting for the first time (right) on Gab and Twitter. *Lower panel*: cumulative evolution of new posts (left) and users posting of the first time (right) on Gab and Twitter.

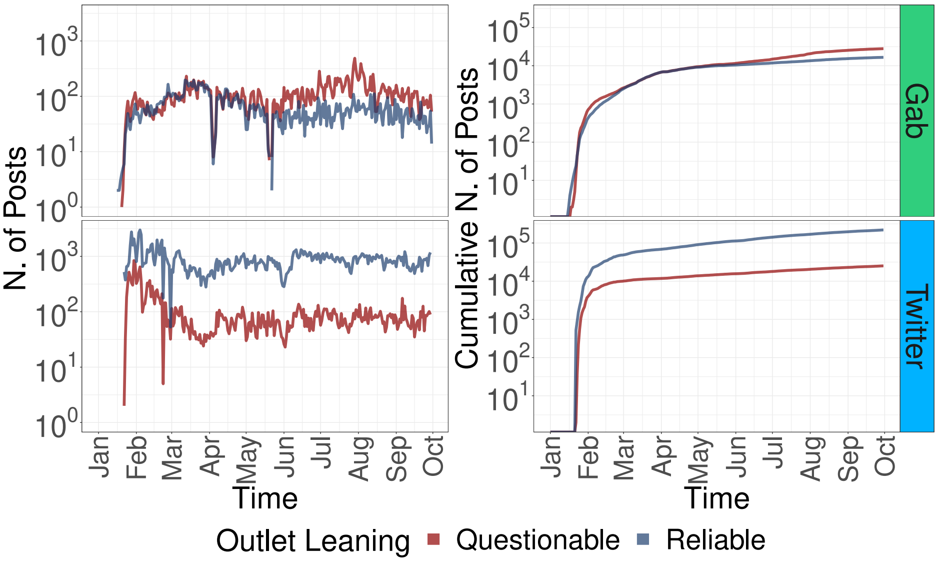


Figure 9.2. Evolution of posts based on their outlet leaning, categorized as Questionable or Reliable. *Upper panel*: Time series evolution of new posts (left) together with its cumulative representation (right) on Gab. *Lower panel*: Time series evolution of new posts (left) together with its cumulative representation (right) on Twitter

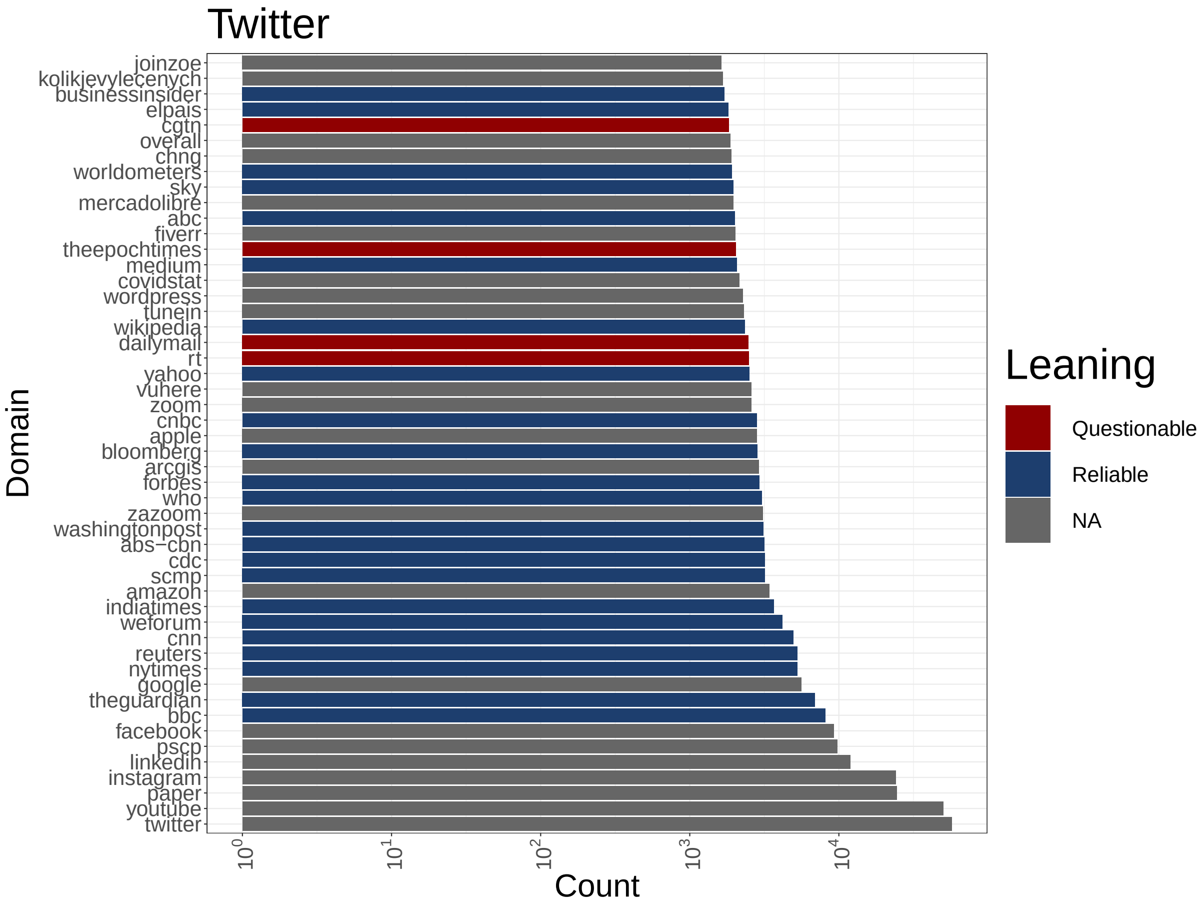


Figure 9.3. Frequency distribution of posts from the most 50 news sources with their corresponding leaning, if existing, on Twitter. We observe a remarkable presence of posts from Twitter and Youtube, whose leaning cannot be inferred since they are platform. From a composition perspective, we observe how Gab posts mainly come from news outlets classified as questionable, describing a specific news diet from users in the platform.

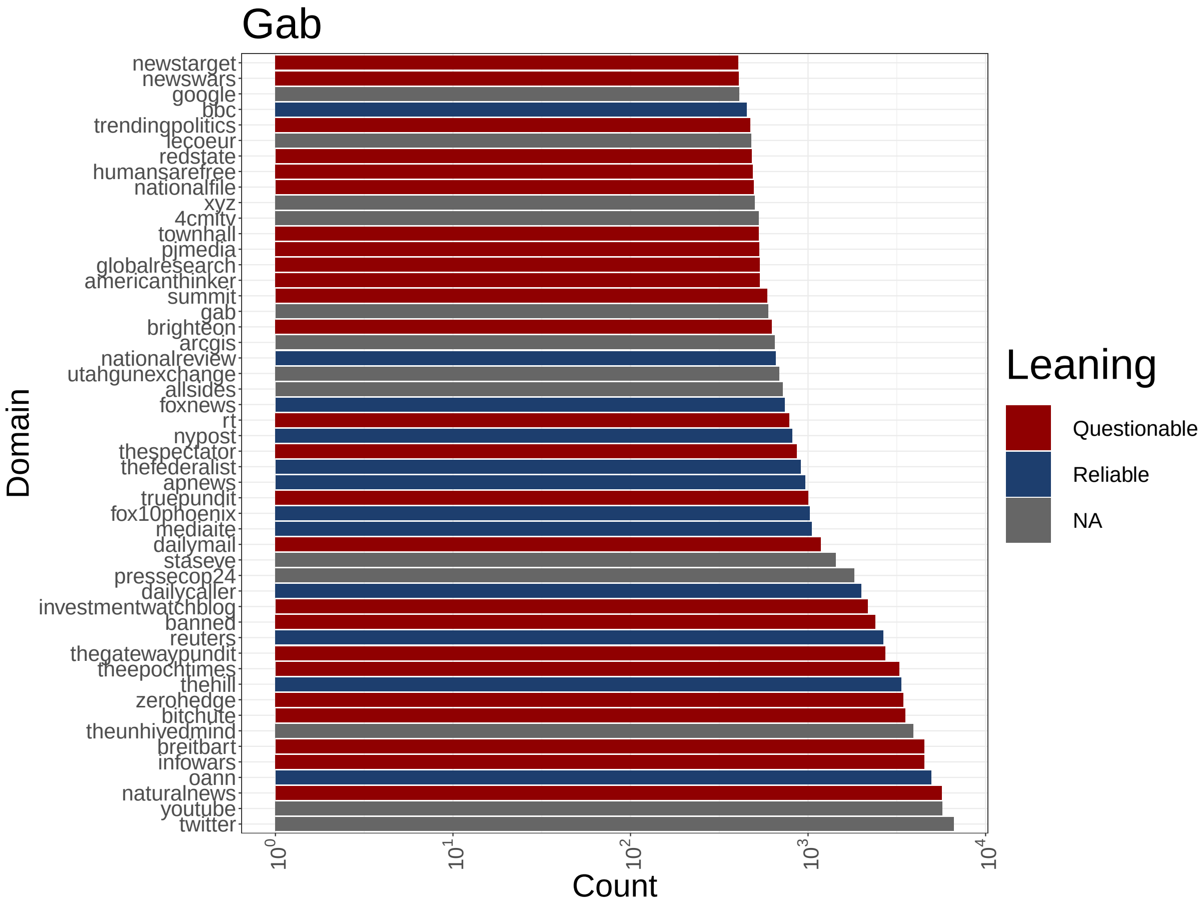


Figure 9.4. Frequency distribution of posts from the most 50 news sources with their corresponding leaning, if existing, on Gab. We observe a remarkable presence of posts from social media platforms like Instagram, Facebook, Twitter, Linkedin, and Youtube, whose leaning cannot be inferred since they are platform. From a composition perspective, we observe how Twitter posts mainly come from news outlets classified as Reliable.

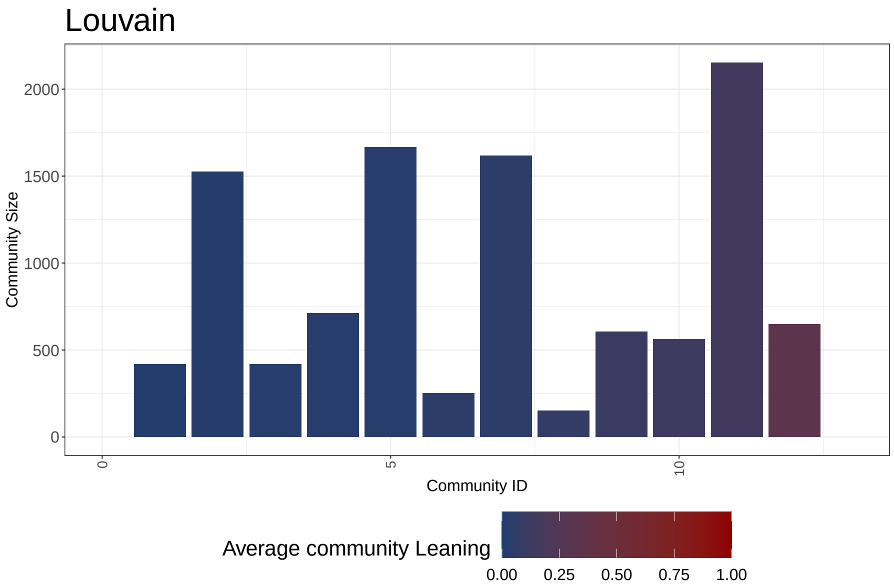


Figure 9.5. Distribution of size and average leaning of community obtained with Louvain community detection algorithm.

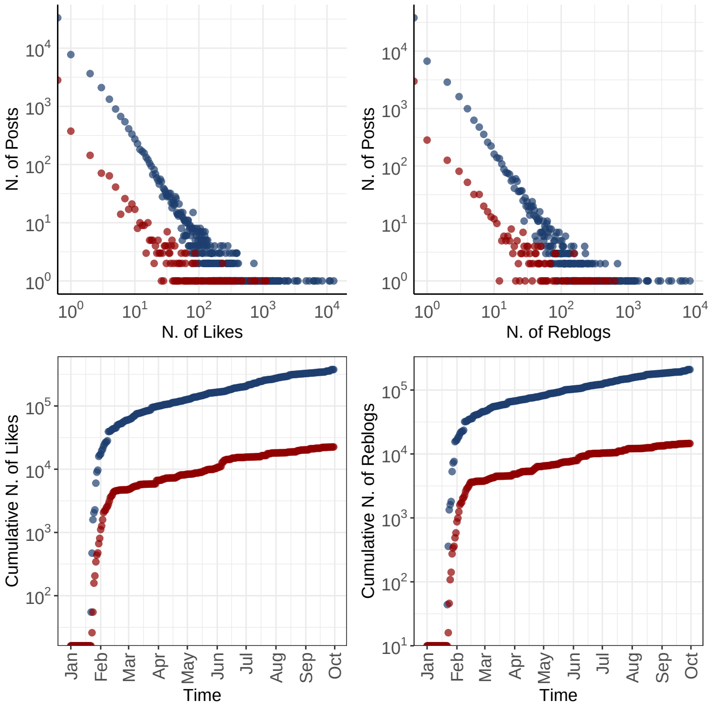


Figure 9.6. Distribution of the number of Twitter posts from US users against the number of Likes (left column) and Reblogs (right column) that posts received, based on their news outlet category.

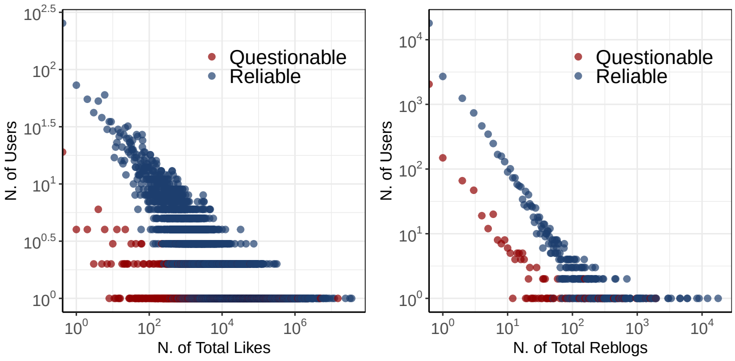


Figure 9.7. Distribution of Likes (left column) and Reblogs (right column) received by US users posting Questionable or Reliable contents on Twitter.

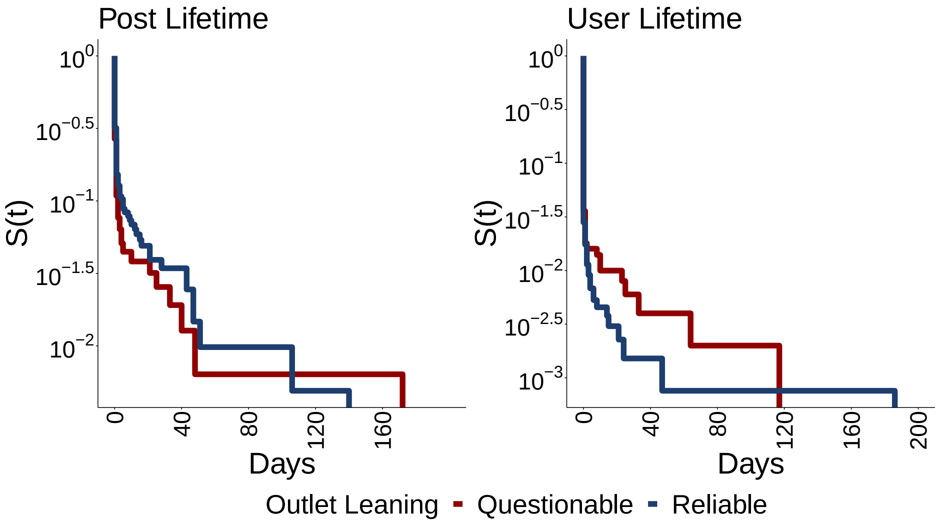


Figure 9.8. Kaplan-Meier estimates based on Twitter posts from US users and, grouped by outlet category. Left column: estimates obtained through the computation of post lifetime, i.e., the period between the first and last comment a post received. Right column: estimates obtained through the computation of post lifetime, i.e., the period between the user's first and last comment

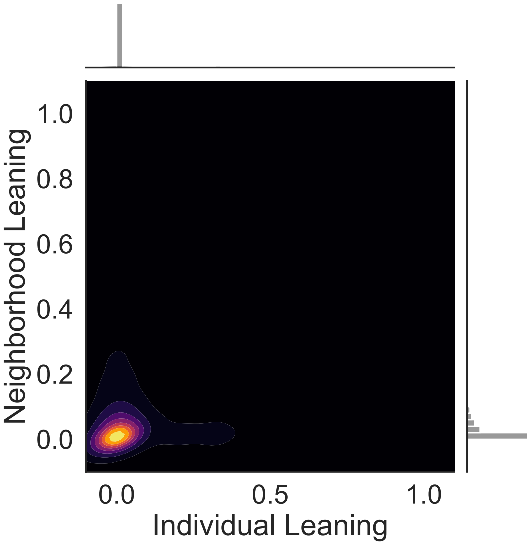


Figure 9.9. Joint distribution between individual and average neighborhood leaning of US users posting classifiable contents at least three times on Twitter.

A Topology-Based Approach for Predicting Toxic Outcomes on Twitter and YouTube

Here we report the supporting information for the analysis.

Keywords list employed in the dataset

In table 9.7 we report the list of search terms employed in the obtainment or original posts from Twitter and Youtube.

Topic	Keywords
Football	SerieA, SerieATim, VAR, Napoli, ForzaNapoliSempre, RangersNapoli, Atalanta, GoAtalantaGo, ForzaAtalanta, Milan, ACMilan, SempreMilan, Udinese, ForzaUdinese, AlèUdin, Inter, IMInter, ForzaInter, Forza Lazio, La Lazio, SSLazio, CMonEagles, ASRoma, Juventus, JuventusFC, Juve, ForzaJuve, Il Torino, Forza Torino, SFT, Salernitana, forzagranata, Fiorentina, forzaviola, ACFFiorentina, Il Bologna, Forza Bologna, ForzaBFC, WeAreOne, Il Sassuolo, Forza Sassuolo, ForzaSasol, "LEmpoli" Forza Empoli, EmpoliFC, EmpoliFootballChannel, HellasVerona, Hellas, DaiVerona, HVFC, Lo Spezia, Forza Spezia, SpeziaCalcio, Il Lecce, Forza Lecce, avantilecce,Cremonese, SolAmAi, ForzaGrigiorossi, DaiCremo, Sampdoria, FORZADORIA, Il Monza, Forza Monza, ACMonza, Monza, InsiemealMonza
Elections	Calenda, Salvini, Meloni, Letta, Bonino, Fratoianni, Conte, DiMaio, Renzi, Paragone, Berlusconi, Civati, Draghi, ElezioniPolitiche2022, Elezioni2022, ElezioniPolitiche, CampagnaElettorale, CampagnaElettorale2022, Azione, LegaNord, Lega, fratelliditalia, fdi, pd, pdnetwork, partitodemocratico, PiuEuropa, sinistraitaliana, M5S, movimento5stelle, impegnocivico, italiaviva, italexit, forzaitalia, possibileit

Table 9.7. List of keywords employed in the data collection process. In the case of Football, keywords include general terms related to the game and to specific football themes. In the case of Elections, keywords include general terms related to 2022 elections, political parties and main candidates.

Results from OLS Linear Regression

In table 9.8 we report the results from OLS linear regressions on the daily average toxicity values.

Social	Topic	Intercept	Date	R^2
Twitter	Elections	-8×10^{-1} . (4.27×10^{-1})	5.06×10^{-5} * (2.21×10^{-5})	0.03
Twitter	Football	-2.98 *** (7.57×10^{-1})	1.59×10^{-4} *** (3.92×10^{-5})	0.11
YouTube	Elections	5.22 *** (6.96×10^{-1})	-2.59×10^{-4} *** (3.60×10^{-5})	0.30
YouTube	Football	9.78 *** (8.22×10^{-1})	-5×10^{-4} *** (4.26×10^{-5})	0.53

Table 9.8. Results for the OLS linear regression model over the days with respect to the average toxicity score by topic and social media. The standard errors of the coefficients are reported in parenthesis, whilst the asterisks refer to the significance of their p-values in the following way: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, . $P < 0.1$, ‘ ’ $P < 1$

Dataset Creation Procedure for Toxicity Prediction

In table 9.9 we report the results from the dataset creation procedure employed in the prediction of the following toxic comment in a conversation.

Prediction Results

Results concerning model prediction, validation on further toxicity thresholds, cross-topic benchmarking and feature importance analysis are available at the following link.

Social	Topic	Interval	N. of pairs	Train Size	Test Size
Twitter	Football	(1, 10]	5460	8735	2185
Twitter	Football	(10, 100]	10359	16574	4144
Twitter	Football	(1000, 10000]	8001	12801	3201
Twitter	Football	(100, 1000]	1330	2128	532
Twitter	Elections	(1, 10]	14738	23580	5896
Twitter	Elections	(10, 100]	35296	56473	14119
Twitter	Elections	(1000, 10000]	7737	12379	3095
Twitter	Elections	(100, 1000]	42147	67435	16859
YouTube	Football	(1, 10]	3132	5011	1253
YouTube	Football	(10, 100]	12224	19558	4890
YouTube	Football	(1000, 10000]	346	553	139
YouTube	Football	(100, 1000]	10787	17258	4316
YouTube	Elections	(1, 10]	13378	21404	5352
YouTube	Elections	(10, 100]	33990	54383	13597
YouTube	Elections	(1000, 10000]	14147	22634	5660
YouTube	Elections	(100, 1000]	55632	89011	22253

Table 9.9. Results of the dataset creation pipeline for each topic and social. The number of pairs represents the number of each toxic/non-toxic pair obtained. The train and test size represent the number of comments used to fit and test the model respectively.

Toxicity in Online Conversations

Dataset	T_o	AC_o $\cdot 10^{-3}$	p	$\langle AC_r \rangle$ $\cdot 10^{-3}$	$\sigma(AC_r)$ $\cdot 10^{-3}$	% ↑	%?	$T_o(16)$	$T_o(26)$
Facebook brexit	↑	46.516	< 0.001	-0.052	6.120	0.050	0.850	↑	↑
Facebook snews	↑	69.323	< 0.001	0.008	1.011	0.050	0.930	↑	↑
Facebook vaccines	↑	42.222	< 0.001	-0.063	0.770	0.030	0.920	↑	↑
Gab other	↑	96.026	< 0.001	0.054	0.554	0.060	0.920	↑	↑
Reddit climatechange	↑	33.960	< 0.001	-0.149	3.967	0.010	0.990	↑	↑
Reddit conspiracy	↑	63.069	< 0.001	0.190	2.015	0.060	0.900	↑	↑
Reddit news	?	6.040	0.112	-0.461	3.599	0.010	0.960	?	?
Reddit science	↑	7.797	0.007	-0.014	0.718	0.010	0.990	↑	↑
Reddit vaccines	?	13.003	0.050	-0.477	4.251	0.020	0.930	↑	?
Telegram conspiracy	↑	116.329	< 0.001	0.025	4.320	0.050	0.900	↑	↑
Telegram news	↑	11.832	< 0.001	-0.022	1.077	0.080	0.910	↑	↑
Telegram politics	↑	16.923	< 0.001	0.101	2.041	0.050	0.890	↑	↑
Twitter climatechange	↑	78.234	< 0.001	0.042	0.549	0.020	0.950	↑	↑
Twitter news	↑	53.778	< 0.001	-0.068	1.221	0.040	0.920	↑	↑
Twitter vaccines	↑	59.489	< 0.001	-0.027	0.867	0.070	0.860	↑	↑
Usenet alt.politics	↓	-48.673	< 0.001	0.021	0.757	0.010	0.960	↓	↓
Usenet conspiracy	↑	14.791	0.014	-0.283	1.648	0.010	0.960	?	?
Usenet news	↑	38.492	0.010	0.027	1.277	0.040	0.940	↑	↑
Usenet talk	?	14.070	0.538	0.027	0.564	0.010	0.970	?	?
Voat conspiracy	↑	33.812	< 0.001	-0.188	1.890	0.030	0.870	↑	↑
Voat news	↑	79.822	< 0.001	0.052	1.787	0.020	0.920	↑	↑
Voat politics	↑	87.597	< 0.001	0.144	1.727	0.030	0.940	↑	↑
YouTube climatechange	↑	35.607	< 0.001	-0.369	3.360	0.040	0.890	↑	↑
YouTube news	↑	27.338	< 0.001	0.009	0.920	0.050	0.870	↑	↑
YouTube vaccines	↑	21.964	< 0.001	0.068	3.055	0.130	0.810	↑	↑

Table 9.10. Trend in toxicity versus conversation size T_o as resulting from a Mann-Kendall test, its linear regression angular coefficient AC_o, its $p-$value, the mean angular coefficient from 200 randomizations of the binary toxicity label $\langle AC_r \rangle$, the standard deviation of their resulting distributions $\sigma(AC_r)$, the percentage of randomizations resulting in an increasing trend % ↑, the percentage of randomizations resulting in an ambiguous trend %?, trend in toxicity for 16 and 26 size intervals $T_o(16)$, $T_o(26)$. For randomizations and other size intervals, a random subset of the Facebook news dataset containing $\sim$ 6.5M comments was used.

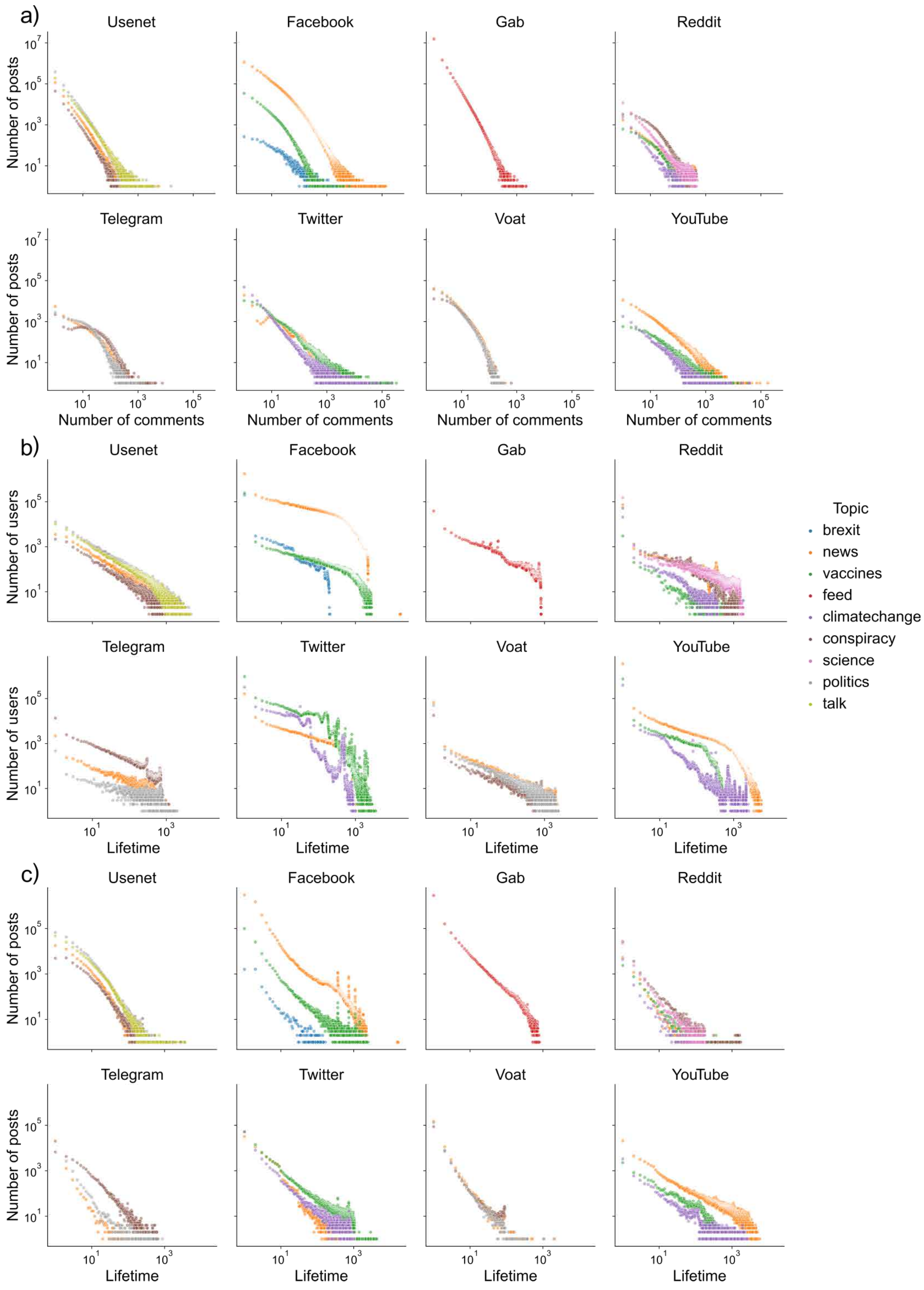


Figure 9.10. General characteristics of online conversations a. Distributions of conversation sizes (number of comments in a thread). **b.** Distributions of the time duration (days) of user activity on a platform for each platform and each topic. **c.** Time duration distributions of threads. Color-coded legend on the side.

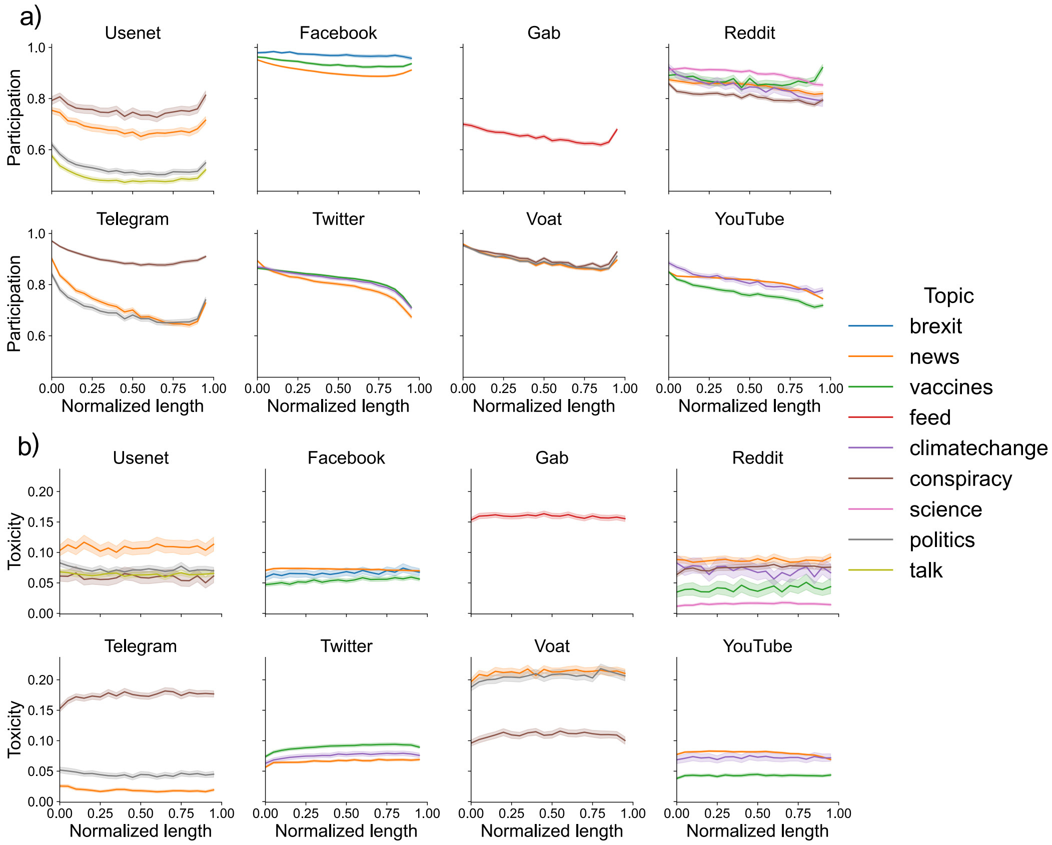


Figure 9.11. User participation and toxicity as conversations evolve. In both figures, the x-axis represents the normalized position of comment intervals in the threads. For each dataset, participation and toxicity are computed in the thread size interval $[0.7-1]$ (see main text and Table 9.12). Trends are reported with their 95% confidence interval. Color-coded legend on the side. **a.** Mean participation of users along threads. **b.** Mean fraction of toxic comments as conversations progress.

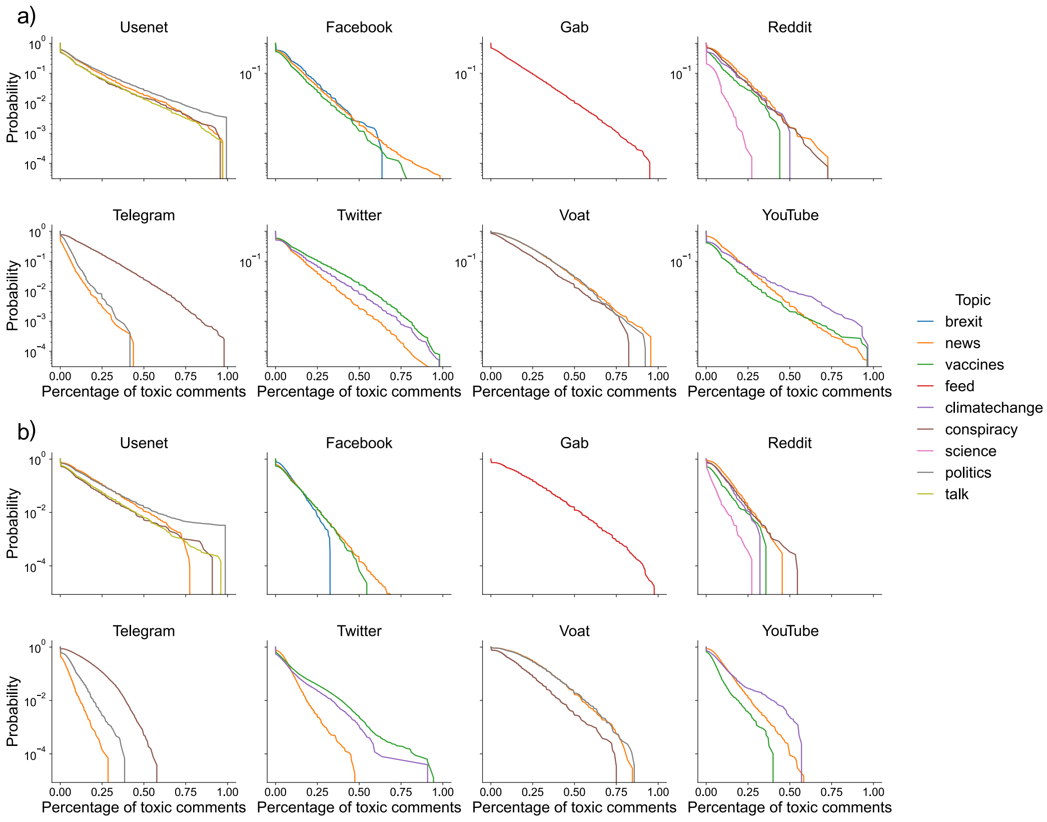


Figure 9.12. Extremely toxic authors and conversations are rare. a. Complementary cumulative distribution functions (CCDFs) of the toxicity of authors who posted more than 10 comments. Toxicity is defined as the fraction of toxic comments over the total of comments posted by a user. Color-coded legend on the side. **b.** CCDFs of the toxicity of conversations containing more than 10 comments.

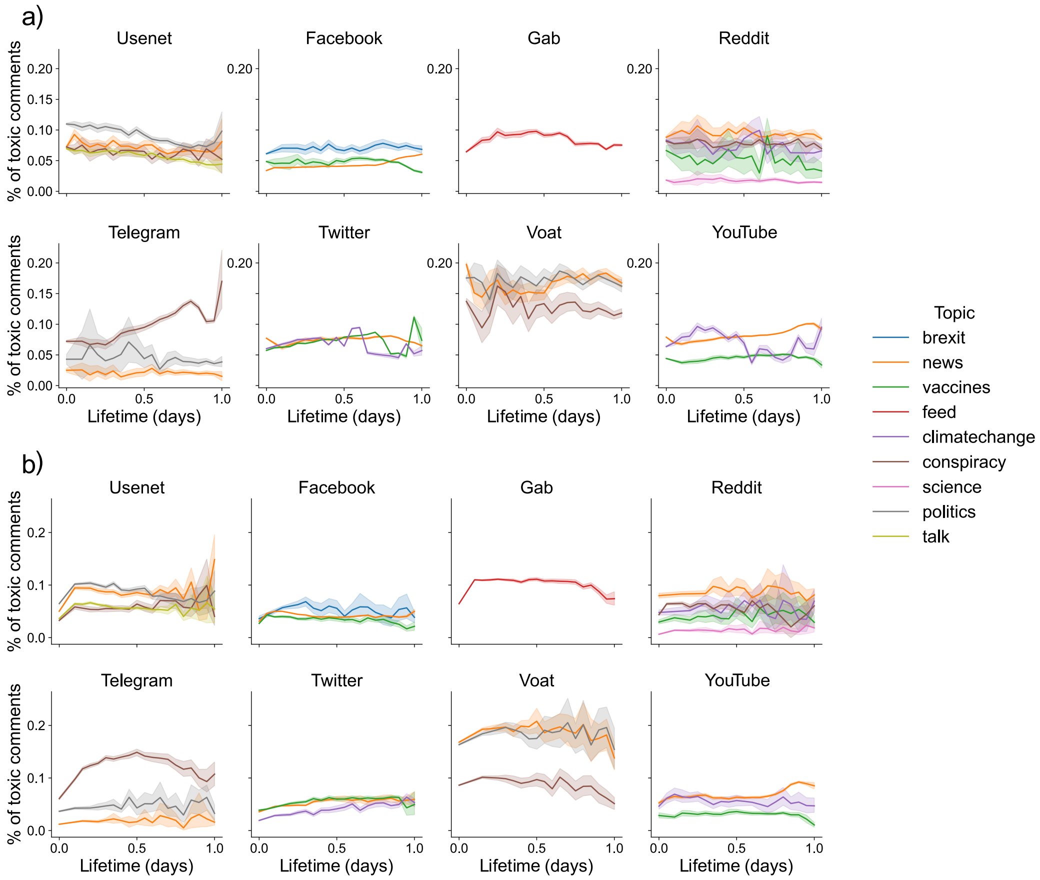


Figure 9.13. Toxicity is not associated with conversation lifetime. Toxicity of **a.** users versus their time of permanence in the dataset and **b.** of threads versus their time duration (in days; log-binned and normalized). Trends are reported with their 95% confidence interval. Color-coded legend on the side.

Dataset	S_o	$AC_o \cdot 10^{-2}$	p	$\langle AC_r \rangle \cdot 10^{-2}$	$\sigma(AC_r) \cdot 10^{-2}$	% ↑	%?
Facebook brexit	?	-0.878	0.239	0.630	0.252	0.590	0.410
Facebook news	↑	2.369	< 0.001	0.477	0.049	1.000	0.000
Facebook vaccines	?	-1.061	0.174	0.586	0.096	0.990	0.010
Gab other	↑	5.151	< 0.001	0.570	0.070	1.000	0.000
Reddit climatechange	↑	2.238	0.002	0.572	0.560	0.160	0.840
Reddit conspiracy	↑	3.919	< 0.001	0.609	0.128	0.990	0.010
Reddit news	↑	4.412	< 0.001	0.591	0.237	0.550	0.450
Reddit science	↑	3.656	< 0.001	0.601	0.216	0.750	0.250
Reddit vaccines	↑	2.033	0.010	0.547	0.479	0.200	0.790
Telegram conspiracy	?	-1.543	0.695	0.606	0.116	0.990	0.010
Telegram news	↑	4.657	< 0.001	1.554	0.114	1.000	0.000
Telegram politics	↑	2.721	< 0.001	1.559	0.125	1.000	0.000
Twitter climatechange	↑	4.408	< 0.001	0.609	0.060	1.000	0.000
Twitter news	↑	4.923	< 0.001	0.495	0.049	1.000	0.000
Twitter vaccines	↑	4.172	< 0.001	0.543	0.037	1.000	0.000
Usenet alt.politics	↑	4.253	< 0.001	0.419	0.086	0.960	0.040
Usenet conspiracy	↑	2.522	0.003	0.467	0.270	0.320	0.680
Usenet news	↑	1.387	0.004	0.471	0.154	0.730	0.270
Usenet talk	↑	4.496	< 0.001	0.459	0.078	0.950	0.050
Voat conspiracy	↑	2.587	< 0.001	0.671	0.152	0.980	0.020
Voat news	↑	2.623	< 0.001	0.612	0.143	1.000	0.000
Voat politics	↑	2.791	< 0.001	0.643	0.166	0.960	0.040
YouTube climatechange	↑	2.478	0.050	0.652	0.234	0.790	0.210
YouTube news	↑	2.134	< 0.001	0.434	0.039	1.000	0.000
YouTube vaccines	↑	7.004	< 0.001	0.516	0.111	0.940	0.060

Table 9.11. Trends in sentiment standard deviation versus conversation size S_o as resulting from a Kendall-Mann test and relative p-value p, its linear regression angular coefficient AC_o, the mean angular coefficient from 100 randomizations of the sentiment score label $\langle AC_r \rangle$, the standard deviation of their resulting distributions $\sigma(AC_r)$, the percentage of randomizations resulting in an increasing trend % ↑, and the percentage of randomizations resulting in an ambiguous trend %?. For randomizations and other size intervals, a random subset of the Facebook news dataset containing ~ 6.5M comments was used.

Dataset	Threads	Min. size	Max. size
Facebook brexit	711	140	8986
Facebook news	10000	1831	131558
Facebook vaccines	3892	82	6552
Gab other	6028	104	2184
Reddit climatechange	320	52	481
Reddit conspiracy	2499	62	481
Reddit news	1520	59	445
Reddit science	2058	63	491
Reddit vaccines	730	26	341
Telegram conspiracy	3804	115	7580
Telegram news	2717	77	1383
Telegram politics	1801	62	1534
Twitter climatechange	2065	837	233855
Twitter news	3440	474	69417
Twitter vaccines	6291	1973	329653
Usenet alt.politics	1864	115	15561
Usenet conspiracy	805	42	1766
Usenet news	984	63	5623
Usenet talk	1967	132	5093
Voat conspiracy	5619	34	636
Voat news	6064	40	386
Voat politics	5134	37	358
YouTube climatechange	800	182	38793
YouTube news	6265	732	178889
YouTube vaccines	1772	324	34106

Table 9.12. Dataset subsets for the analysis of participation and toxicity along threads. For each dataset, it is reported the number of conversations, along with their minimum and maximum size, which correspond to the 0.7 and 1 values of the normalized log-binning. Facebook news threads were limited to 10000 for computational reasons.

Dataset	Threads	Profiled users	$\langle PC \rangle$	r	ρ	τ
Facebook news	$1,395,899$	$6,818,569$	0.723	0.86	0.961	0.886
Gab feed	$69,202$	$18,031$	0.825	0.842	0.909	0.75
Twitter news	$23,505$	$54,295$	0.554	0.723	0.764	0.581
Twitter vaccines	$1,214$	$14,405$	0.171	0.908	0.929	0.817

Table 9.13. The datasets used in the analysis of controversy. For each dataset, the number of conversations (Threads), the number of users to which a political leaning could be assigned (Profiled users), the mean percentage of comments from a profiled user in the conversations ($\langle PC \rangle$), the Pearson's r, Spearman's ρ and Kendall's τ correlations between the trends in toxicity and controversy.

Dataset	N. of Threads	Peak >Pre	Peak >Post	Post >Pre
Facebook brexit	1408	0.000	0.658	0.000
Facebook news	5000	0.000	0.141	0.000
Facebook vaccines	1655	0.000	0.019	0.000
Gab feed	4176	0.000	0.000	0.028
Reddit climatechange	329	0.062	0.314	0.471
Reddit conspiracy	3240	0.000	0.000	0.000
Reddit news	1658	0.001	0.013	0.411
Reddit science	2341	0.000	0.000	0.000
Reddit vaccines	258	0.000	0.105	0.033
Telegram conspiracy	11904	0.000	0.000	0.000
Telegram news	4657	0.000	0.000	0.000
Telegram politics	2589	0.000	0.000	0.320
Twitter climatechange	5000	0.000	1.000	0.000
Twitter news	5000	0.000	0.002	0.000
Twitter vaccines	5000	0.000	0.004	0.000
Voat conspiracy	2407	0.000	0.000	0.000
Voat news	3844	0.000	0.000	0.000
Voat politics	2807	0.000	0.001	0.000
YouTube climatechange	637	0.000	0.042	0.001
YouTube news	5000	0.000	0.014	0.000
YouTube vaccines	2132	0.000	0.033	0.000

Table 9.14. Conversations are more toxic at the peak of activity. Burst analysis of activity in conversations. For each dataset, the number of threads considered in the analysis, along with the p-values for the hypothesis (H1) that the distributions in toxicity are more skewed towards higher toxicity content at the peak of activity w.r.t. previous and subsequent activity levels (Peak >Pre and Peak >Post, respectively), and after the peak compared to before the peak (Post >Pre). H1 is considered accepted if $p < 0.01$.

Platform and Topic	Time range	Comments	Threads	Users	TP	TD	TI
Facebook film	2015/01/01-2017/08/10	493825	13385	305415	0.03	0.09	0.13
Facebook sports	2015/01/01-2017/08/19	494275	19632	271575	0.04	0.11	0.18
Reddit askreddit	2021/02/01-2021/02/15	2544457	170691	512917	0.06	0.11	0.16
Reddit iama	2020/01/01-2022/12/31	529584	27487	206413	0.03	0.06	0.10
Reddit movies	2020/01/01-2023/01/01	12169720	587603	1357318	0.05	0.09	0.17
Telegram crypto	2021/07/23-2023/06/22	152428	4818	19375	0.05	0.08	0.20
Twitter got	2019/01/01-2022/12/24	441274	62359	222158	0.02	0.04	0.05
Twitter nasa	2019/01/01-2022/12/29	342654	22736	185922	0.02	0.03	0.06
Voat askvoat	2014/06/19-2020/12/25	612668	50389	55251	0.12	0.19	0.28
Voat whatever	2015/05/31-2020/12/25	1366671	177278	76940	0.20	0.27	0.40
YouTube carbonara	2018/01/05-2023/07/02	699670	3615	545174	0.03	0.04	0.04
YouTube football	2022/08/02-2023/04/07	948918	16887	165706	0.02	0.08	0.13

Table 9.15. Validation dataset breakdown. TP, TD and TI indicate, respectively, the percentage of toxic comments in a dataset labelled by Perspective, Detoxify and IMSYPP.

Platform and Topic	Threads	Min. size	Max. size
Facebook film	919	114	7894
Facebook sports	1083	91	10689
Reddit askreddit	365	396	33507
Reddit iama	567	174	12298
Reddit movies	1939	920	90978
Telegram crypto	1280	40	575
Twitter got	572	80	20233
Twitter nasa	607	89	6418
Voat askvoat	3072	37	429
Voat whatever	6233	37	387
YouTube carbonara	694	199	20249
YouTube football	2283	115	2024

Table 9.16. The subset used for the analysis of participation.

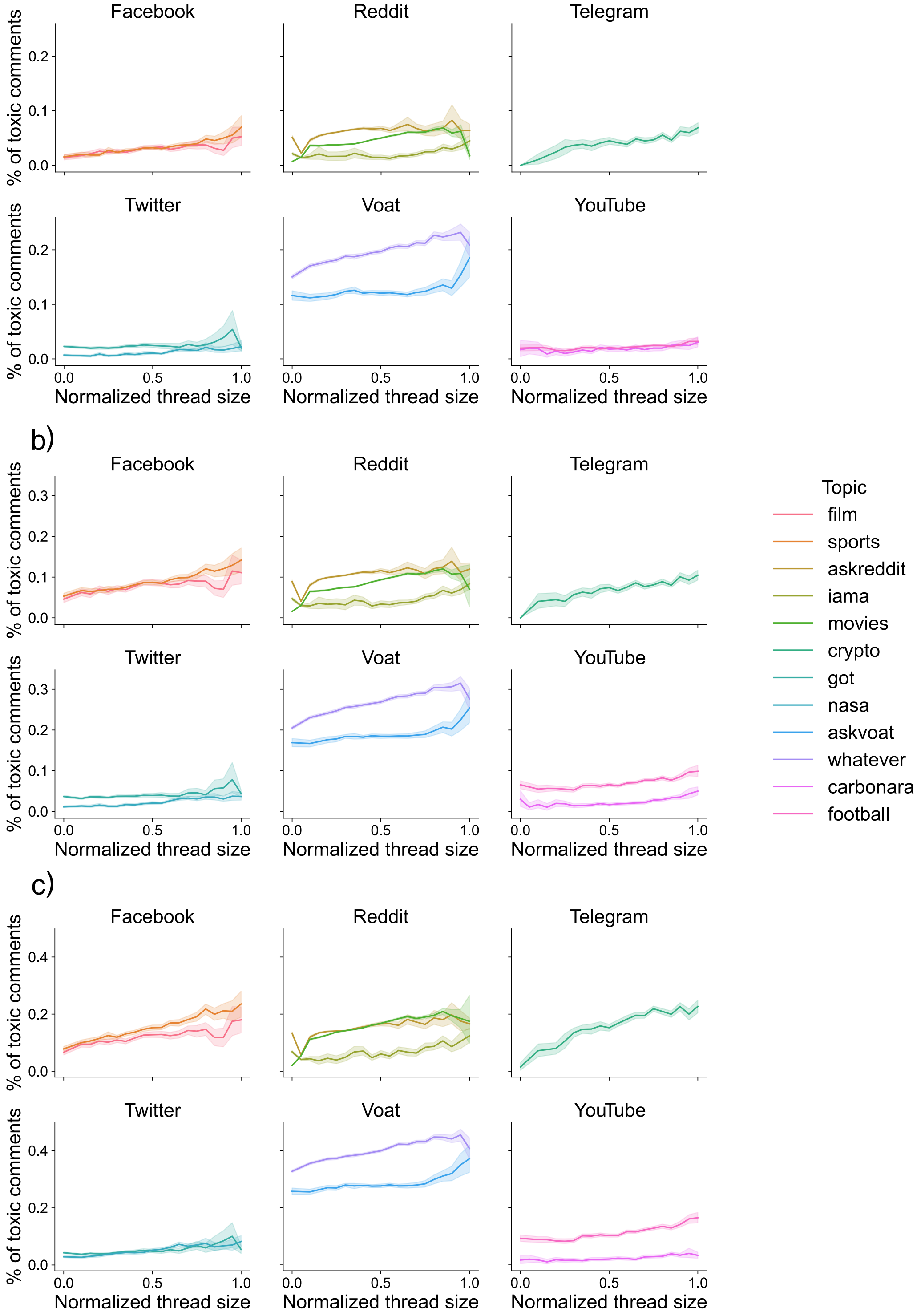


Figure 9.14. Validation dataset: toxicity increases with conversation size.s Fraction of toxic comments in conversations versus conversation size (cfr. Figure 8.2), using Perspective API **a)**, Detoxify **b)** and IMSYPP **c)**.

.

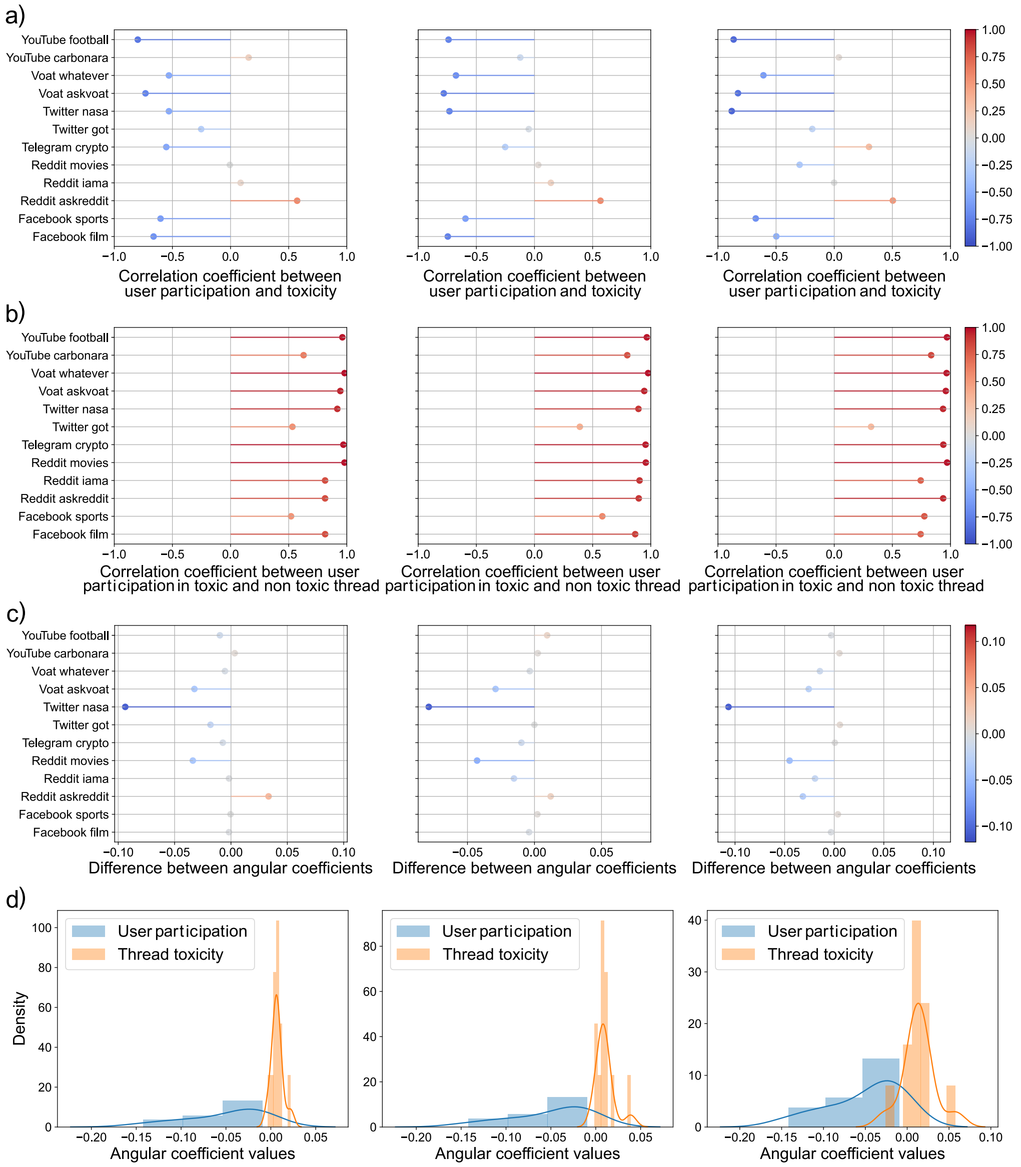


Figure 9.15. Participation analysis for the validation dataset. (From left to right, results obtained with: Perspective API, Detoxify, IMSYPP) **a.** Pearson's correlation coefficients between user participation and toxicity trends for each dataset. **b.** Pearson's correlation coefficients between users participation in toxic and non-toxic thread sets, for each dataset. **c.** Difference between toxic and non-toxic thread sets participation slopes resulting from linear regression. **d.** Density distribution of toxicity and participation trend slopes, as resulting from linear regression.

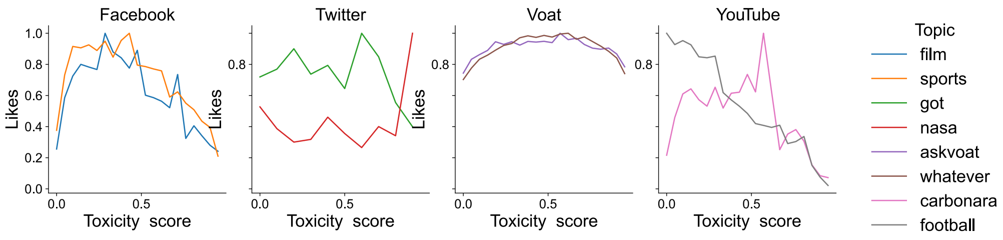


Figure 9.16. Burst analysis of activity in conversations (see Table 9.14) for the validation dataset, using the three toxicity detectors. Lower: Likes/upvoats versus binned toxicity (as detected by Perspective API)

Dataset	N. of Threads	Peak >Pre	Peak >Post	Post >Pre	Detector
reddit askreddit	17879	0.000	0.000	0.000	Perspective
reddit iama	1403	0.000	1.000	0.000	Perspective
reddit movies	5000	0.000	0.000	0.000	Perspective
voat askvoat	8597	0.000	0.000	0.002	Perspective
voat whatever	16537	0.000	0.000	0.000	Perspective
telegram crypto	2167	0.000	0.043	0.000	Perspective
twitter got	2912	0.000	0.032	0.000	Perspective
twitter nasa	2580	0.000	1.000	0.000	Perspective
facebook film	3990	0.000	0.860	0.000	Perspective
facebook sports	4588	0.000	0.147	0.000	Perspective
YouTube carbonara	1864	0.000	1.000	0.000	Perspective
YouTube football	6824	0.000	0.002	0.000	Perspective
reddit askreddit	17879	0.000	0.000	0.000	Detoxify
reddit iama	1403	0.000	1.000	0.000	Detoxify
reddit movies	5000	0.000	0.000	0.000	Detoxify
voat askvoat	8597	0.000	0.005	0.049	Detoxify
voat whatever	16537	0.000	0.022	0.000	Detoxify
telegram crypto	2167	0.000	0.004	0.000	Detoxify
twitter got	2912	0.000	0.117	0.000	Detoxify
twitter nasa	2580	0.000	1.000	0.000	Detoxify
facebook film	3990	0.000	0.896	0.000	Detoxify
facebook sports	4588	0.000	0.267	0.000	Detoxify
YouTube carbonara	1864	0.000	0.713	0.000	Detoxify
YouTube football	6824	0.000	0.000	0.000	Detoxify
reddit askreddit	17879	0.000	0.012	0.000	IMSYPP
reddit iama	1403	0.000	1.000	0.000	IMSYPP
reddit movies	5000	0.025	0.471	0.590	IMSYPP
voat askvoat	8597	0.004	0.488	0.110	IMSYPP
voat whatever	16537	0.002	0.764	0.019	IMSYPP
telegram crypto	2167	0.000	0.003	0.000	IMSYPP
twitter got	2912	0.000	0.331	0.000	IMSYPP
twitter nasa	2580	0.000	0.772	0.000	IMSYPP
facebook film	3990	0.000	1.000	0.000	IMSYPP
facebook sports	4588	0.000	1.000	0.000	IMSYPP
YouTube carbonara	1864	0.000	1.000	0.000	IMSYPP
YouTube football	6824	0.000	0.020	0.000	IMSYPP

Table 9.17. Burst analysis of activity in conversations (see Table 9.14) for the validation dataset, using the three toxicity detectors.

Dataset	Perspective		Detoxify		IMSYPP	
	T_o	AC_o $\cdot 10^{-3}$	T_o	AC_o $\cdot 10^{-3}$	T_o	AC_o $\cdot 10^{-3}$
Facebook film	$\uparrow$	26.514	$\uparrow$	43.801	$\uparrow$	76.734
Facebook sports	$\uparrow$	43.609	$\uparrow$	79.326	$\uparrow$	145.053
Reddit askreddit*	$\uparrow$	25.876	$\uparrow$	49.323	$\uparrow$	77.223
Reddit iama**	$\uparrow$	18.571	$\uparrow$	37.081	$\uparrow$	65.016
Reddit movies	$\uparrow$	36.789	$\uparrow$	70.762	$\uparrow$	132.068
Telegram crypto	$\uparrow$	52.078	$\uparrow$	78.797	$\uparrow$	185.316
Twitter got***	$\uparrow$	15.263	$\uparrow$	25.209	$\uparrow$	43.190
Twitter nasa	$\uparrow$	16.958	$\uparrow$	30.139	$\uparrow$	55.154
Voat askvoat	$\uparrow$	38.608	$\uparrow$	56.176	$\uparrow$	83.373
Voat whatever	$\uparrow$	68.315	$\uparrow$	89.382	$\uparrow$	108.309
YouTube carbonara*	$\uparrow$	11.882	$\uparrow$	24.988	$\uparrow$	22.925
YouTube football	$\uparrow$	11.946	$\uparrow$	39.510	$\uparrow$	76.700

Table 9.18. Results of Mann-Kendall test on the toxicity trends T_o and slope from linear regression AC_o of the validation dataset for Perspective Detoxify and IMSYPP. All tests p-values are < 0.001, except *$= 0.001$, **$= 0.002$, ***$= 0.003$.

		Detoxify		IMSYPP				
Dataset	Perspective	NT	T	acc.	off.	viol.	inapp.	T^*
All combined	NT	93.10	6.90	73.01	25.72	0.53	0.75	26.99
	T	12.57	87.43	13.53	75.87	2.07	8.52	86.47
Facebook	NT	94.00	6.00	78.68	20.28	0.60	0.45	21.32
	T	4.59	95,41	7.75	83.25	3.40	5.60	92.25
Gab	NT	94.07	5.93	74.18	24.69	0.65	0.47	25.82
	T	18.30	81.70	17.30	72.53	2.90	7.26	82.70
Reddit	NT	93.93	6.07	78.05	20.21	0.40	1.34	21.95
	T	4.01	95.99	4.80	73.58	1.22	20.40	95.20
Telegram	NT	93.96	6.34	81.21	15.83	0.84	2.12	18.79
	T	19.86	80.14	16.17	59.41	4.18	20.23	83.83
Twitter	NT	94.15	5.85	80.62	18.38	0.16	0.84	19.38
	T	26.87	73.13	30.94	63.79	0.45	4.82	69.06
Usenet	NT	91.20	8.80	56.78	42.71	0.32	0.19	43.22
	T	6.74	93.26	6.63	89.98	1.07	2.32	93.37
Voat	NT	90.88	9.12	67.33	30.84	1.02	0.82	32.67
	T	4.12	95.88	6.08	80.89	3.45	9.59	93.92
YouTube	NT	93.94	5.51	75.93	23.29	0.41	0.37	24.07
	T	18.50	81.50	18.28	74.60	1.29	5.82	81.72

Table 9.19. Agreement table between the classification given by Perspective API, and those of Detoxify and IMSYPP. Classification labels abbreviations: NT: non-toxic; T: toxic; acc.: acceptable; off.: offensive; viol.: violent; inapp.: inappropriate; T^*: toxic, considered as the sum of offensive, violent and inappropriate. Numbers represent percentages.

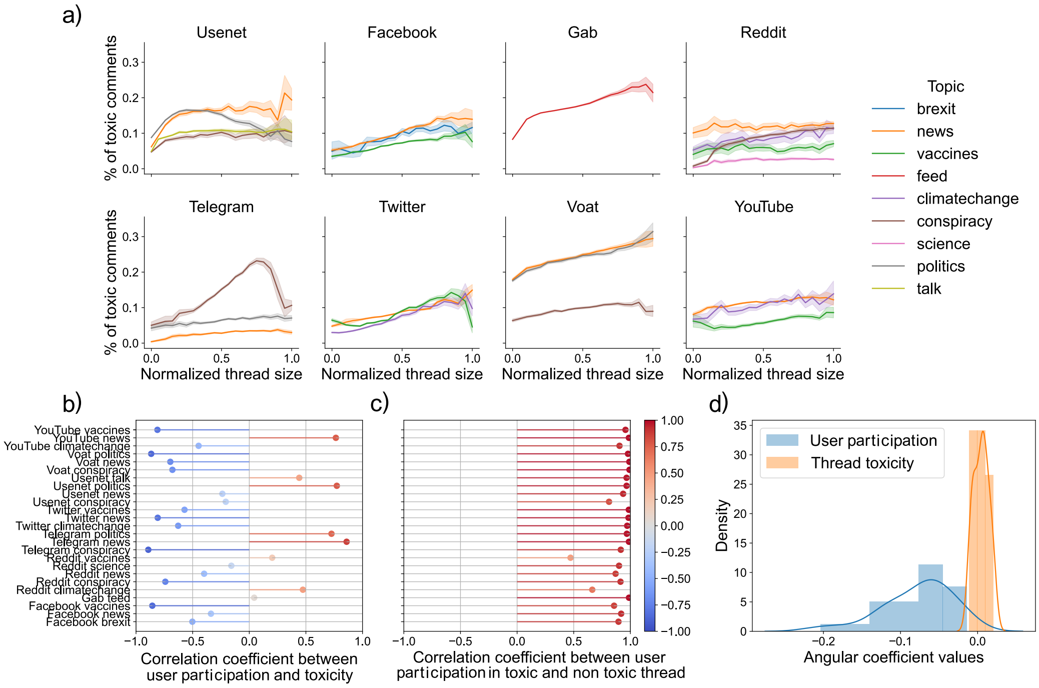


Figure 9.17. Results hold for a different toxicity threshold. Core analyses presented in the analysis repeated employing a lower (0.5) toxicity binary classification threshold. **a.** Fraction of toxic comments in conversations versus conversation size, for each dataset (see Figure 8.2). **b.** Pearson's correlation coefficients between user participation and toxicity trends for each dataset.**c.** Pearson's correlation coefficients between users' participation in toxic and non-toxic thread sets, for each dataset. **d.** Density distribution of toxicity and participation trend slopes, as resulting from linear regression. The results of the relative Mann-Kendall tests for trend assessment are shown in Table 9.20.

Table 9.20. Results of Mann-Kendall tests applied to the toxicity vs conversation size trends (T_o) and slopes from linear regression (AC_o) using 0.5 as threshold for toxicity.

Dataset	T_o $\cdot 10^{-3}$	AC_o	p
Facebook brexit	↑	74.636	< 0.001
Facebook news	↑	103.339	< 0.001
Facebook vaccines	↑	60.415	< 0.001
Gab other	↑	118.554	< 0.001
Reddit climatechange	↑	48.237	< 0.001
Reddit conspiracy	↑	91.641	< 0.001
Reddit news	↑	11.655	0.018
Reddit science	↑	15.760	0.001
Reddit vaccines	?	10.462	0.093
Telegram conspiracy	↑	131.799	< 0.001
Telegram news	↑	24.736	< 0.001
Telegram politics	↑	30.354	< 0.001
Twitter climatechange	↑	105.319	< 0.001
Twitter news	↑	85.595	< 0.001
Twitter vaccines	↑	77.410	< 0.001
Usenet alt.politics	↓	-60.915	< 0.001
Usenet conspiracy	↑	32.323	< 0.001
Usenet news	↑	68.091	< 0.001
Usenet talk	↑	27.043	0.021
Voat conspiracy	↑	42.022	< 0.001
Voat news	↑	99.098	< 0.001
Voat politics	↑	108.226	< 0.001
YouTube climatechange	↑	60.412	< 0.001
YouTube news	↑	39.834	< 0.001
YouTube vaccines	↑	37.279	< 0.001

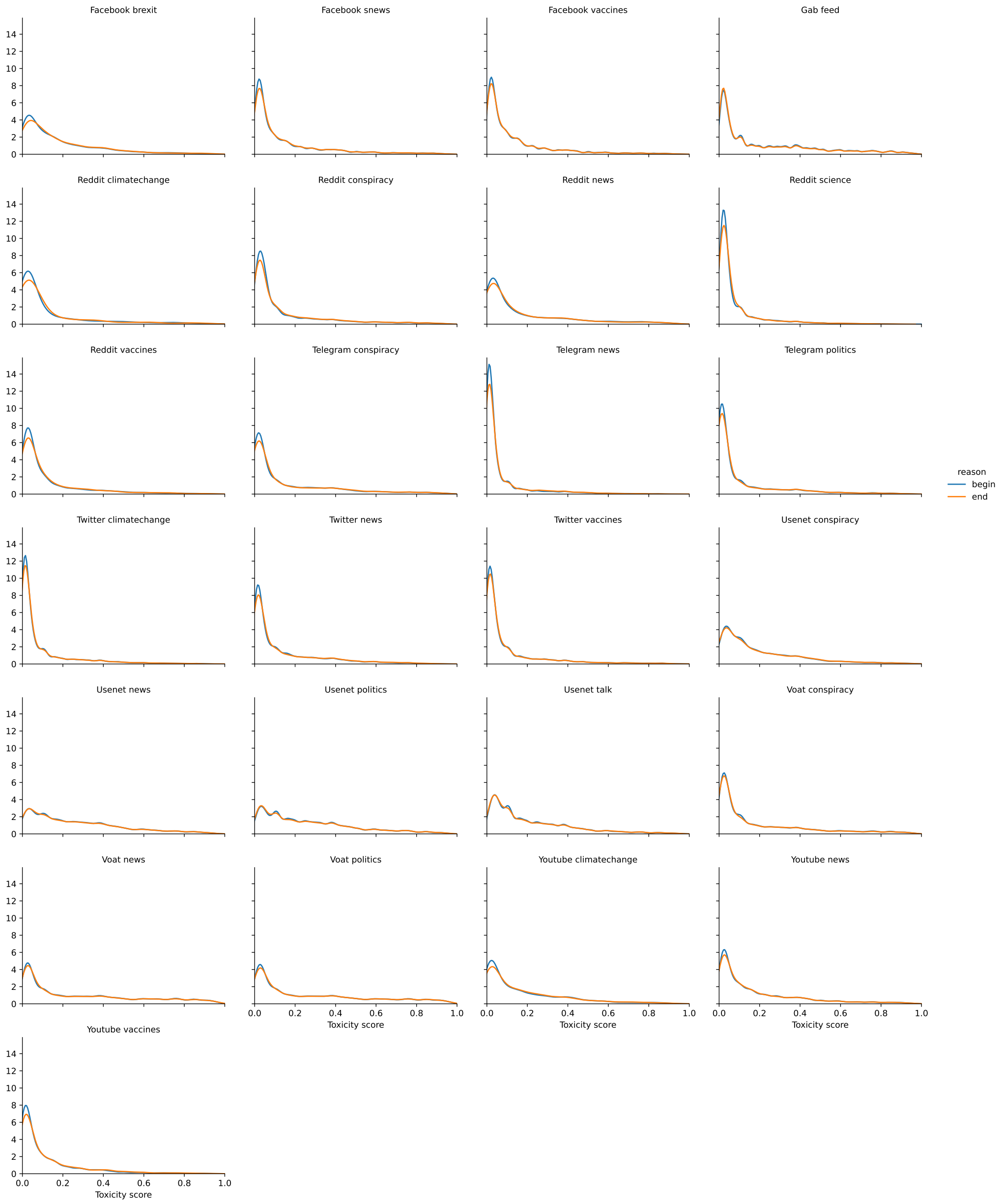


Figure 9.18. Short conversations are not short because of toxicity (1). Each plot shows the density of the toxicity distribution for all comments but the last three ('begin' label) and for the last three comments ('end' label) for conversations composed of $6 - 20$ comments. No significant differences appear between any of the density pairs. In general, the last comments are not more toxic than those preceding them – see also Figure 9.19. For Facebook news, a subsample containing ~ 6.5M comments was used.

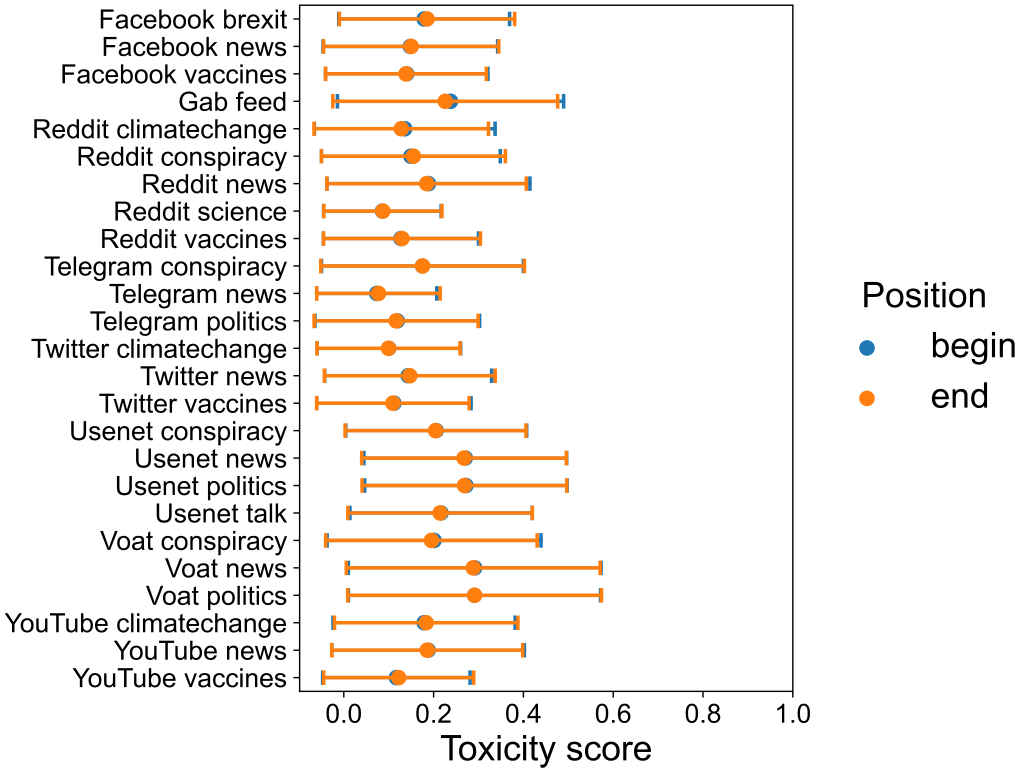


Figure 9.19. Short conversations are not short because of toxicity (2). Average toxicity of last three comments ('end') and remaining ones ('begin') in short conversations (see Figure 9.18). Bars represent one standard deviation.