



UNIVERSITÀ
DI TRENTO

Department of
Information Engineering and Computer Science

Doctoral Programme in
Information Engineering and Computer Science

QUANTIFYING AND MODELING RISKS AND DYNAMICS OF INFORMATION DISORDERS IN DIGITAL ECOSYSTEMS

Anna Bertani

Advisor
Riccardo Gallotti
Fondazione Bruno Kessler

December 2025

Abstract

Social media platforms, as primary sources of information, have profoundly transformed public communication, reshaping how information is created, disseminated and consumed. While enabling unprecedented connectivity, they also amplify the risks associated with information disorders, ranging from the spread of misleading content to harmful online behaviors. The thesis examines the problem of information disorders by investigating four aspects: the characterization of online conversation during particularly important events, the measurements and validation of indexes able to quantify the risks of infodemic and the modeling of infodemic dynamics and radicalization in online social media. In particular, this dissertation provides a new framework capable of measuring and modeling how unreliable information spreads, how emotional and socio-economic factors influence such diffusion, and how these mechanisms affect the stability and resilience of digital ecosystems. To address this problem, we combined large scale data analysis, network science approaches and modeling to quantify and interpret the dynamics of information disorders. The first part focuses on the emotional responses associated with the online discussion of highly debated topics, such as the COVID-19 vaccination campaign and Long COVID. The second part proposes a new methodological approach to quantify the news media diet and information consumption patterns of users. Building upon existing indexes of infodemic risk, the third part provides a validation of such indexes by using socio-economic indicators, that explain cross country differences in the exposure and susceptibility to misinformation. In addition, a stochastic model is developed to reproduce the empirical volatility of the Infodemic Risk Index, revealing differences in the superspreader dynamics across countries. The thesis also extends its analytical framework to the study of radicalization dynamics on Reddit, showing how users' linguistic and emotional patterns evolve after exposure to misogynistic communities, thus offering predictive insights into harmful online behaviors. This approach advances the state of the art by moving from descriptive analyses of misinformation toward a unified and quantitative modeling framework that connects micro-level user behavior with macro-level systemic instability. This thesis contributes to the development of the computational social science field, proposing a data-driven science of informational risk aimed at supporting policy design and digital governance for healthier and more resilient information ecosystems.

Keywords

[information disorder, misinformation, computational social science, infodemic risk index, network science]

*To the women of my life,
who have inspired me with their example and courage.*

Acknowledgments

My PhD journey has been an intense and rich experience, and it would not have been the same without the people who walked with me, supported me, and inspired me along the way. First of all, I want to thank my supervisor, Riccardo Gallotti, also known as the one who can always be recognized at conferences for being tall and for his dreadlocks. Thank you for being not only a supervisor, but a true mentor: someone I could ask questions about science and about life. Working with you has taught me so much, and thinking that only four years ago I was “just” an intern feels almost unreal. I am deeply grateful to Professor Manlio De Domenico, not only for providing the resources that allowed me to do research on important topics such as the COVID-19 pandemic, but also for being the person who first encouraged me to pursue a PhD. I want to thank all the collaborators I had the pleasure to work with: Pierluigi Sacco, Federico Pilati, Alessandro Cortese, Valeria Mazzeo, Lucila Alvarez-Zuzek, Kaveh Kadkhoda, and Vincent Traag, among others. A special thanks goes to the OSoMe group at Indiana University for hosting me during my visiting period abroad, and especially to Professor Filippo Menczer and Alessandro Flammini for their guidance. I am also grateful to all the amazing and funny people I met there. I also want to thank the reviewers for their comments and my defense committee: Anees Baqir, Floriana Gargiulo, and Elena Pavan. A heartfelt thanks goes to my colleagues and friends who shared parts of this PhD journey with me. Especially, thank you to the entire CHuB group and to Ulysse, Albi, Matteo, and Vero, for being there, listening, helping me in so many ways, and even hosting me when I was homeless. I am also thankful to all the people who taught me to live life with more lightness, to notice the beauty in small details, and to see elephants in the special corners of the world. Throughout these years, I have been lucky enough to grow alongside people who have known me through many phases of my life, from childhood friends to university companions. Rosa, thank you for being my first classmate and still being here after all these years. Cri and Marco, thank you for all the laughs, the shared struggles, and for opening the doors of the Department of Sociology whenever I felt lost. Adua, Marta, Chiara, and Elisa, thank you for always finding ways to stay in touch and for making me feel your support even from far away. Finally, my deepest gratitude goes to my parents, Carla and Paolo. Thank you for trusting me, letting me follow my own path, and always supporting my choices. A special thanks to my sister Laura, who has always cared for me, sometimes even physically, and has been a constant support in so many ways. I owe everything to their love and support and their faith in me is fundamental to make me go on my way.

Contents

1	Introduction	1
1.1	Contributions of the Thesis	2
1.2	List of Publications	5
2	Background and Conceptual Framework	7
2.1	Social Media as an Information Ecosystem	7
2.1.1	Characteristic of different online social platforms	8
2.2	How does online social media work?	9
2.2.1	Algorithm and Virality of online content	9
2.2.2	Information Diffusion and Cascades	10
2.2.3	Emotional Diffusion of Online Content	10
2.3	Information Disorders	12
2.3.1	Definitions and distinctions: misinformation, disinformation, ma- linformation	12
2.3.2	Mitigating the spread of mis- and disinformative content	14
2.3.3	Countering the false information with Artificial Intelligence . . .	15
2.3.4	Related works	16
2.3.5	The Covid-19 Infodemic Platform	17
3	Large-scale digital signatures of emotional response to the COVID-19 Vaccination Campaign	21
3.1	Introduction	21
3.2	Method	24
3.2.1	Overview of the data set	24
3.2.2	Verified Accounts	25
3.2.3	News Reliability and Political Leaning	26
3.2.4	Bot Detection	27
3.2.5	Emotional Detection	28
3.2.6	Emotional Aggregation	29
3.3	Results	29
3.3.1	Emotions distribution among the Covid-19 Vaccination and the Baseline sample	29
3.3.2	Emotional Responses to Different Levels of News Reliability . .	31
3.3.3	Emotional Responses of Verified vs Unverified Users	32

3.3.4	Emotional Responses based on the political orientation	34
3.4	Discussion	36
4	Digital Monitoring of a Rising Global Disease: A Twitter-Based Study of Long Covid Conversations	39
4.1	Introduction	40
4.2	Methods	43
4.2.1	Overview of the data	43
4.2.2	News Reliability and Political Categories	43
4.2.3	Verification Status	44
4.2.4	Emotional Detection and Aggregation	44
4.2.5	Dataset about the prevalence of Long COVID	44
4.2.6	Community and Information Cascades	44
4.2.7	Multilevel Communities	46
4.3	Results	46
4.3.1	Emotional Analysis	48
4.3.2	News Media Characterization	50
4.3.3	Community Analysis	50
4.4	Discussion	55
5	Decoding the News Media Diet of Disinformation Spreaders	59
5.1	Introduction	59
5.2	Materials and Methods	61
5.2.1	Overview of the Dataset	61
5.2.2	News Reliability	61
5.2.3	Building the Information Networks	62
5.2.4	User Categories	63
5.2.5	Entropy	64
5.3	Results	65
5.4	Discussion	73
6	The interplay between socio-economic indicators and infodemic risk indices	77
6.1	Introduction	77
6.2	Methods	81
6.2.1	Twitter dataset	81
6.2.2	Socioeconomic indicators dataset	81
6.2.3	Dimensionality Reduction Techniques	82
6.2.4	k-means	82
6.2.5	News Media Diet	83
6.2.6	Volatility	84
6.3	Results	84
6.4	Discussion	91
6.4.1	Socioeconomic dimensions and country clustering	91
6.4.2	Temporal infodemic patterns and country clusters	92

6.4.3	Evolving correlations between socioeconomic factors and infodemic measures	93
6.4.4	The protective role of media diet diversity	94
6.4.5	Infodemic volatility and institutional stability	95
6.4.6	Theoretical and policy implications	95
6.4.7	Limitations and future directions	98
6.5	Conclusion	99
7	A Stochastic Model for Understanding Volatility in Infodemic Risk	
	Index	101
7.1	Introduction	101
7.2	Methods	102
7.2.1	Overview of the data	102
7.2.2	Infodemic Risk Index	103
7.2.3	Volatility	103
7.2.4	Monte Carlo Simulation	103
7.3	Results	104
7.3.1	Followers Distribution	106
7.3.2	Disinformation Ratio	107
7.3.3	Engagement	108
7.3.4	Superspreaders posting behavior	108
7.4	Discussion	111
8	The Radicalization Pipeline: Characterizing and Predicting online misogyny	115
8.1	Introduction	116
8.2	Methods	117
8.2.1	Classification Subreddits	117
8.2.2	Toxicity and Hate Speech	117
8.2.3	Lexical Feature Analysis	118
8.2.4	Classifier Features	118
8.3	Results	119
8.4	Discussion	126
9	Conclusion	129
	Bibliography	133
A	Supplementary Information Chapter 3	165
A.1	Data	165
A.2	Emotional Analysis	166
A.3	Type of Users Accounts	168
A.4	Statistical Analysis	169
B	Supplementary Information for Chapter 4	173

C	Supplementary Information Chapter 5	185
C.1	Levels of High-Risk users	185
C.2	Verified accounts	185
C.3	Cross-countries Analysis	186
D	Supplementary Information Chapter 6	195
E	Supplementary Information for Chapter 7	201

List of Tables

3.1	Statistics about the datasets. The table shows the number of messages, unique users and the timeframe considered for both the datasets.	25
8.1	Comparison of model performance across evaluation metrics.	125
8.2	Performance metrics for various feature sets.	125
B.1	Scientific sources by category. We show the absolute and percentage values of scientific sources classified into one of the following categories: peer-reviewed journal, science communication, agency, preprint.	182
B.2	Mainstream Source by Category. We show the distribution of unique sources and mentions according to six subcategories.	183
D.1	List of the socioeconomic indicators considered for analysis. Table of the socioeconomic indices, the organization and sources, and a brief description.	195

List of Figures

2.1	Information Disorders. Visualization of the intersection of the three key component of disinformation disorders: Misinformation, Disinformation and Malinformation [1]	13
2.2	Screenshot of the FBK COVID-19 Infodemics Observatory dashboard .	18
2.3	Classification of the news media categories based on their level of reliability based on the work of Gallotti et al. [2].	19
3.1	Emotion Mean Value. For each sample (baseline (left panel) and vaccine (right panel)), we quantified the mean value of each emotion: Anger, Sadness, Surprise, Fear, Anticipation, Trust, Disgust, Joy. The horizontal line is placed at the 0.125 value to better visualize which emotions are over- vs under-represented with respect to our benchmark value. The grey vertical lines show the standard deviation of each distribution. To better understand the statistically significance of each emotions, we performed a Kolmogrov-Smirnov Test with the result that all the distribution are statistically different at the 95 % of confidence level with a p-value lower than 0.01 both across emotions and among the baseline and vaccine samples. In addition, we calculated the standard error of the mean, resulting to be negligible among all the emotion distribution.	30
3.2	Emotional range for different categories of news. For each sample (baseline (left panel) vs vaccine (right panel)), we quantified the distribution of emotional values from -1 to 1, respectively for tweets containing reliable news (top), unreliable news (middle) or just comments without sources (bottom). All the distribution have been tested through a Kolmogrov-Smirnov Test in order to evaluate the difference among the samples and across categories of news considered. The distribution are statistically significant at the 95 % of confidence level with a p-value lower than 0.01 among the two samples and within the same sample, with the only exception of tweets containing unreliable news and reliable news in the Baseline sample, having a p-value equal to 0.22.	32

3.3	Emotional Analysis for different types of user accounts. We quantified, ranging from -1 to +1, across (A) verified and unverified users, (B) both human and bot, over time. The dashed horizontal lines indicate the emotional mean values for each user type, based on a baseline of 10 million randomly sampled tweets. This highlights that the mean emotion of the random distribution is higher than that of our specific interest. We present the distribution of emotional values, aggregated by day, for each category concerning the vaccine sample (C-D). We conducted a Kolmogorov-Smirnov Test across the entire distribution for each category. The results indicate statistical significance at the 95% confidence level, with a p-value below 0.01, not only within categories but also across the various samples considered.	33
3.4	Emotional range for messages posted on Twitter by verified-unverified accounts and with URLs associated to five different political leaning. For each panel, we quantified statistics of tweets across the emotional range (-1, +1) according to the political leaning of the sources and the verified/un-verified status of the users. The left panel reports the 10M random sample of tweets as a baseline, while the central one is about the vaccine tweets. The right panel shows the percentage difference between vaccine and baseline data for each cell. We show how emotional responses can become a sort of marker for different ideological position, especially for the left-center, center and center-right categories. An interesting result is shown by the verified accounts through which the emotional responses vary at different scales, reaching high negative and positive values.	35
4.1	Visualization of the multilevel communities. We show the network visualization for the first five levels identified by a multilevel community detection analysis (see Methods for details). Each node is coloured according to the assigned community.	47
4.2	Reliability and Political Index. Left panel shows the network visualization enriched by the reliability index ranging from -1 (completely untrustworthy) to +1 (completely trustworthy). Right panel shows the network enriched by the political index ranging between -1 (far-right), 0 (least biased leaning) and +1 (far-left).	48
4.3	Emotion Distribution for different types of political leaning. For each sample (Long COVID vs baseline), we quantified the distribution of emotional values from -1 to 1, respectively for tweets with left, left-center, center, right-center and right political leaning.	49

4.4	Emotional range for messages posted on Twitter by verified-unverified accounts and with URLs linking to five different political leaning orientations and eight news media categories. The top panel shows the percentage difference between the baseline and Long COVID data for all the categories considered: news media, verified-unverified accounts, and political leaning. The bottom panel shows the p-value results (in logarithmic scale) from performing the T-Test between the two distributions for each cell.	51
4.5	Degree Distribution and Collective Attention in Information Cascades. For each panel (A-C), we quantified the average degree of the first node for Long COVID and baseline samples, considering respectively 1) all the cascades, cascades that 2) involve 10 or more users, or 3) 50 or more users. For each panel (D-F), we quantified the average degree of the nodes involved in the cascades (but the first node) for the Long COVID and baseline discussions in all three scenarios, as described above. Panel G shows the time needed to reach collective attention for Long COVID and baseline samples.	54
4.6	Long COVID Prevalence vs Science News Regression. Testing the hypothesis that in countries where Long COVID is more prevalent the volume of science-based news is higher. For each panel, we quantified the Spearman Correlation and the linear regression between Prevalence of Long COVID and news classified as Science, for 2020 and 2021 respectively.	55
5.1	The News Media Diet. Network and list of the sequence of web-domains posted by a random user on Twitter. We defined the length of each sequence with the term L . Each sequence is decoded in a list of identified letters. On these lists, we calculate the entropy value to obtain information about the variety of each sequence. In particular, we performed both random entropy S_{rand} and Shannon entropy S_{unc} calculations for each sequence.	63
5.2	Overview of the dataset. (A) Distribution of the length sequence of web-domains, respectively, for the entire dataset, for users having posted more than 2 distinctive web-domains ($N > 1$) and users defined according the categories of reliability. (B) Distribution of the number of messages classified in one of the eight categories of news media types. (C) Distribution of the news media categories used by a particularly active user who has posted reliable and low-risk content, with 6 distinctive categories of news.	66

- 5.3 **The news media environment.** (A) Weighted networks of the interactions among different news media types (also known as intra-relations) for 25 countries in 2020. (B) Heatmap showing the corresponding value of the inter-relations among the eight categories. (C) Weighted networks normalized and compared with a null model accounting for the proportion of the number of news pieces belonging to each category. The chances of sharing conspiracy and fake news by the same users is much higher than the strong relation observed between mainstream media and political news (A). (D) Heatmap showing the corresponding normalized value of the inter-relations among the eight categories. 68
- 5.4 **Self-loops of web-domains and types of news shared.** (A) The percentage distribution of the number of self-loops, pointing to the same web-domains for all the categories considered. (B) The percentage distribution of the number of self-loops of users pointing to the same category of news, regardless of the different web-domains shared. 69
- 5.5 **News media diet for different types of users.** (A) Random entropy S_{rand} calculated for different types of accounts: users posting reliable or low-risk content and users posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon entropy S_{unc} calculated for the domain and the type of news shared by different types of users accounts: users posting reliable or low-risk content and users posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated for the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content. 71
- 5.6 **Fraction of repeated news media categories and web-domains for users with different N .** In these boxplots, we represent the percentage of news media categories (A,B) or web-domains (C,D) repeated in subsequent tweets with respect to the total length of the sequence associated with an individual user. Each boxplot is aggregated over users with the same N . The dashed line shows the random expectation of repeated news media categories with respect to the increasing number of N unique news media categories. (A) Patterns of news media categories, aggregated for users classified as having posted only reliable and/or low-risk content on Twitter. There are 8 box-plots here since we have 6 distinctive categories and 2 are filtered out. (B) Patterns of news media categories, aggregated for users having also posted high-risk content. There are 8 box-plots here since we have all 8 distinctive categories of news media types. (C) Patterns of web-domains, aggregated for users classified as having posted reliable and/or low-risk content. (D) Patterns of web-domains, aggregated for users classified as having also posted high-risk content in their sequence of messages posted. 72

6.1	UMAP visualization of OECD countries and the socioeconomic indices. UMAP visualization of the OECD* countries based on the selected indices (left) and of the selected socioeconomic indices (right). The employed hyperparameters are $n\text{ neighbors} = 5$ and $\text{min dist} = 0.01$.	85
6.2	UMAP visualization of IRI timeseries. UMAP plot of the IRI time series of the countries featuring at least 500'000 tweets in the relevant time frame. The selected parameters are $n\text{ neighbors} = 5$ and $\text{min dist} = 0.1$. The 9 countries belonging to Cluster 2 are grouped on the left, whereas the remaining ones belong to Cluster 1.	86
6.3	k-means clustering with Euclidean distance and UMAP preprocessing on the IRI time series for the 88 selected countries. The UMAP parameters employed were $n\text{ neighbors} = 5$, $\text{min dist} = 0.1$ and $n\text{ components} = 102$. The cluster average is plotted in red. All the time series are smoothed using a 30-day rolling average.	87
6.4	Evolution of correlation between First Principal Component and Infodemic Indexes The first panel shows the Spearman correlation of the first Principal Component and the monthly average of the infodemic metrics over time for OECD* countries. The second panel displays the p-value of the corresponding correlation with a black dashed line representing the significance level of 5%.	88
6.5	Correlation between the news media diet and the UMAP of the socioeconomic. Panels A and B show the correlation between the socioeconomic indicators with the entropy of all the domains classified as low-risk (satire, clickbait and political), while Panels C and D represent the correlation between the socioeconomic indices and the news classified as political.	89
6.6	Correlation between the second dimension of UMAP and the volatility of Infodemic Risk Index. Panel A shows the correlation between the volatility of the Infodemic Risk Index in 2020, while Panel B shows the same relations in 2021.	90
7.1	Time Series of Infodemic Risk Index. Evolution of the Infodemic Risk Index over time for 4 different countries: Perù, Venezuela, Italy and France.	105
7.2	Correlation between average number of tweets and the Volatility of Infodemic Risk Index. We calculated the spearman correlation between the volatility, defined as standard deviation of IRI, and the average number of tweets, showing that countries with more tweets tend to be more stable in terms of Infodemic Risk Index, due to the effect size of the messages' volume.	106

7.3	Correlation between average number of tweets and the Volatility of Infodemic Risk Index only on our subset of data. We calculated the spearman correlation between the volatility, defined as standard deviation of IRI, and the average number of tweets, showing that countries with more tweets tend to be more stable in terms of Infodemic Risk Index, due to the effect size of the messages' volume.	107
7.4	Distribution of power-law alpha. Distribution of the exponent value of powerlaw distribution fou users sharing reliable content (2.11) and those sharing unreliable content.	108
7.5	Relationship between number of followers and engagement received. We found for all the 4 countries (PER, VEN, ITA, FRA) a strong correlation between number of followers and engagement received (especially with users with more than 10^3 followers.	109
7.6	Unreliability vs Top 5 % of superspreaders Correlation between superspreaders users (defined as the top 5 % of users in terms of followers distribution) and the volatility of the countries. Countries with higher value of superspreaders sharing unreliable content are also the ones with higher instability with respect to the Infodemic Risk Index.	110
7.7	Volatility and Superspreaders posting behavior. Panel A maps the volatility of each country analyzed, while Panel B shows the average unreliable ratio of news posted by superspreaders, showing the relationship between instability of IRI and behavior of the most influential users in the digital ecosystem.	111
7.8	Simulation of the empirical scaling β We simulated a population of different users posting online, characterized by four parameters: disinformation ratio, exponent value of the followers distribution, engagement received and posting behaviour of superspreaders.	112
8.1	Hate Speech and Toxicity Distribution of Hate Speech and Toxicity for comments and posts before and after joining <i>incel</i> communities. . .	120
8.2	Engagement in different communities. Average fraction of posts and comments in different communities before and after joining <i>incel</i> community.	121
8.3	Engagement in different communities normalize by time. Average number of posts (A) and comments (B) normalized by time. Relative Difference between posts (C) and comments (D) over time.	122
8.4	Toxicity across subreddits. Absolute values and Relative Difference of toxicity before and after the engagement with Incel community. . . .	123
8.5	Hate Speech across subreddits. Absolute values and Relative Difference of hatespeech ratio before and after the engagement with Incel community.	124

8.6	Cosine similarity of participation of the same threads for Incels and No Incels users We calculated the cosine similarity of threads commented on by Incels and Non-Incels users to understand whether users had engaged in prior discussions before joining the Incel subreddit.	126
A.1	Terms about vaccine discourse. Terms from the 18 most used language on Twitter, used to filter the Twitter Firehose on Covid-19 for this analysis. The terms are related to vaccination. Additionally, we considered also the following terms related to the vaccine names: astrazeneca, moderna, johnson & johnson, sputnik, pfizer, biontech.	166
A.2	Correlation among Valence-Arousal-Dominance emotional dimension and the eight emotions. For each emotions, we calculated the Spearman correlation, finding that the three VAD emotions are highly correlated among them, while the four positive and four negative emotions are not correlated with respect to each singular emotion.	167
A.3	Heatmap showing the Spearman Correlation among Valence-Arousal-Dominance emotional dimension and the eight emotions. We show the Spearman Correlation against each singular emotions (Joy, Surprise, Trust, Anticipation, Fear, Anger, Disgust, Sadness) and the three VAD emotions. We find that the the VAD emotions are highly correlated among them with values of 0.88 among Dominance and Valence, 0.83 among Dominance and Arousal, 0.67 among Valence and Arousal.	168
A.4	Distribution of positive and negative emotions during the entire phase of Covid-19 Vaccination Campaign. We show the average distribution per day of each emotions. In particular, the emotions with a positive valence (Joy, Trust, Anticipation, Surprise) are represented by different scale of green, whereas the emotions with a negative valence (Fear, Anger, Disgust, Sadness) are drawn by different scale of red. Based on this plot, we decided to normalize the data by summing the value of each emotion and then divided each singular emotion by the total emotional value. Separately, we summed the positive and the negative emotions and then we normalized each new category in a range between -1 to +1 in order to have a more comparable emotional value for each message posted on Twitter.	169

A.5	Evolution of the messages, emotions and events during the Covid-19 Vaccination Campaign.	In the first panel, we quantified the distribution of tweets and retweets posted on Twitter during the entire phase of Covid-19 Vaccination Campaign. The second panel shows the distribution of emotions in a range between -1 (completely negative valence) and +1 (completely positive valence). We calculated the mean emotion for those emotional dimensions that have reach values higher than 0.25 in particular days. The third panel indicates the cumulative logarithmic number of messages with reference to particular vaccines names. Labels refer to manually inspected events about the massive COVID-19 vaccination campaign.	170
A.6	Emotional Analysis for different type of users accounts.	We quantified (A) emotional trends from -1 to +1 for verified/unverified users, (B) for verified and unverified human, verified and unverified bot. We have shown both the mean distribution per day and the rolling mean emotions in order to better visualize the frequent variation of the emotional distribution with respect to each type of users accounts. Especially in the second panel, we observe that the differences between the four categories of users tend to stabilize as the vaccination campaign unfolds, whereas in the very first months the responses are less predictable. . . .	171
B.1	Emotion Mean Value.	For each sample of data (LongCovid and baseline), we quantified the mean value of each emotion: Anger, Sadness, Surprise, Fear, Anticipation, Trust, Disgust, Joy are considered. The horizontal line is placed on the 0.125 value to better visualize the difference between the two samples of data.	174
B.2	Emotion Distribution for different type of news.	For each sample (LongCovid vs baseline), we quantified the distribution of emotional values from -1 to 1, respectively for tweets containing fake news, with no fake news and without a link URL.	174
B.3	Distribution of Sentiment Scores in Information Cascades.	We quantified the sentiment scores both for LongCovid Discussions and Baseline sample.	175
B.4	Distribution of Reliable and Unreliable News for Long Covid Discussion and Baseline Sample.	We quantified the percentage of reliable and unreliable news for LongCovid Discussion and a baseline sample (panel A). Panel B highlights the percentage variation of the news media categories with respect of the baseline. The Panel C displays the cumulative deviation of the different news with respect to the baseline sample.	176

B.5	Distribution of Political Leaning for Long Covid Discussion and Baseline Sample. We quantified the percentage of politicized content for LongCovid Discussion and a baseline sample. Panel B highlights the percentage variation of the politicized content from the baseline. The Panel C displays the cumulative deviation of the different political leaning with respect to the baseline sample.	177
B.6	Communities-level news and political content We identify the largest online communities and, for each community, we show the estimated news distribution and political content.	177
B.7	Communities Centralization For each panel (A-C), we quantified (A) the percentage of social activities, (B) the percentage of news shared, (C) the percentage of political content posting on Twitter concerning discussions about Long Covid.	178
B.8	Distribution of the different news media sources. We quantified the distribution of news media sources. We quantified the variation between the volumes of the eight news media classification posted and the baseline sample.	179
B.9	Distribution of the most frequent web-domains. We quantified the distribution of the most frequent web-domains for news classified as science and mainstream media outlets. We quantified the variation between the volumes of those domains posted and the baseline sample.	180
B.10	Visualization of the Dynamic IRI across countries for each year considered. For each panel(A-C), we quantified the distribution of the dynamic IRI for each year considered across all the world. The three panels show that the dynamic IRI remains consistently low for the entire period.	184
C.1	News Media Diet for different type of users (9 categories of users). (A) Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.	187

C.2	News Media Diet for verified users. (A) Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.	188
C.3	News Media Diet for users classified as unverified. (A) Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.	189
C.4	The News Media Environment (The Netherlands). (A) Weighted networks of the interactions among different news media types (also known as intra-relations) for 25 countries in 2020. (B) Heatmap showing the corresponding value of the inter-relations among the eight categories. (C) Weighted networks normalized comparing with a null model accounting for the proportion of the number of news belonging to each category. The chances of sharing conspiracy and fake news by the same users is much higher than the strong relation observed between mainstream media and political news (Panel A). (D) Heatmap showing the corresponding normalized value of the inter-relations among the eight categories.	190

C.5	News Media Diet for different type of users (The Netherlands). (A) Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.	191
C.6	The News Media Environment (Colombia). (A) Weighted networks of the interactions among different news media types (also known as intra-relations) for 25 countries in 2020. (B) Heatmap showing the corresponding value of the inter-relations among the eight categories. (C) Weighted networks normalized comparing with a null model accounting for the proportion of the number of news belonging to each category. The chances of sharing conspiracy and fake news by the same users is much higher than the strong relation observed between mainstream media and political news (Panel A). (D) Heatmap showing the corresponding normalized value of the inter-relations among the eight categories.	192
C.7	News Media Diet for different type of users (Colombia). (A) Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.	193
E.1	Correlation between average number of tweets and volatility of Infodemic Risk Index We calculated the average number of tweets per country and the volatility, defined as the standard deviation of IRI every two days.	201

Chapter 1

Introduction

The digital age has deeply incorporated technology into our daily lives, producing a huge amount of data that might describe and predict human behavior and social trends in an extraordinary way. The availability of data has caused the rise of the so-called computational social science, an emerging field that exploits the ability to collect and analyze data at scale revealing patterns of individual and group behaviors [3, 4]. Despite the availability of a large amount of data, computational social science needs an accurate methodology to inspect and better understand the emerging macro-phenomena. One key aspect of research in the computational social science has been the study of online social media and its impact on individuals and society. Many scholars have investigated different aspects of the use of social media, including research on the spread of misinformation [5], polarization [6], news consumption [7], and impact on psychology of users and on democracy itself [8].

This thesis focuses specifically on the concept of *information disorders*, as a broad term that indicates the creation and dissemination of false and misleading information. While disinformation is defined as the practice of sharing unreliable content with the intention of misleading, typically used for propaganda to manipulate public opinion, misinformation refers to inaccurate information disseminating unintentionally, which is the objective of this thesis. Even though misinformation and disinformation are not new concept in the literature [9], the digital age has incredibly altered and amplified the risks of misinformation, specifically about how content is created, disseminated and consumed. Misinformation has been defined as one of the most severe global risks of our society by the World Economic Forum, highlighting how this issue is facing several and unique barriers to the resolution of the problem such as data, accountability and communication pathways [10]. These issues are increased by the widespread use of

social media as a prevalent news source and community hub, where content can rapidly reach millions of users without verification or editorial oversight. Indeed, many studies identify that adults choose social media as a primary source for news, 44 % of young adults with an age between 18-25 declare that they use social media and videos network as main source for news consumption [11]. On the other hand, public concerns arises with respect to the diffusion of misinformation: recent surveys show that 59 % of users believe that misinformation is a very big problem for our society [12]. The recent progress in the field of Artificial Intelligence constitutes another emerging risk in the diffusion of false content. Indeed, text, images and videos can be easily created by using generative AI such as Large Language Model (LLMs), enabling the creation and dissemination of realistic but false content in a faster and uncontrollable way [13, 14]. Beyond being fertile ground for the spread of misinformation, online social media are often characterized by *echo-chambers*, a sort of environment where users encounter beliefs that amplify and reinforce their pre-existing beliefs by being exposed to their own existing views without encountering opposing views [15]. This can undoubtedly lead to some cognitive biases such as selection bias [16] and confirmation bias in the way users choose and rely on some specific narratives and news. In this way, echo chambers can increase the social and political polarization, leading to forms of radicalization and extremism. Forms of radicalization can also contribute to an increase of hate and contempt toward specific and marginalized groups of people. Recent research found that nearly 56% of U.S. adults have experienced harassment linked to race, religion, gender, or sexual orientation in digital environments [17]. At a broader level, the United Nations and European Commission also warned that online hate speech and extremist propaganda not only undermine social cohesion but can also serve as catalysts for real-world violence [18, 19], highlighting how these effects not only are threats to individual well-being but also to the democratic institutions and societal resilience [20].

1.1 Contributions of the Thesis

Within this context, this thesis contributes to the study of information disorders not only focusing on the the spread of false content but also on the radicalization of users in online social communities. In particular, I will investigate four aspects: the *characterization* of conversations in online social media platforms during particularly important events, the *measurements* and the *validation* of indexes with respect to the risks of infodemic, and the modeling of infodemic *dynamics* and radicalization in online social media.

The second chapter will serve as an introduction to the relevant literature and previous works in the field of information disorders. Framing from social media as powerful yet pervasive information ecosystems, I will explain the digital ontology of the social media platforms adopted in the following studies: Twitter/ X and Reddit. Then I will present the characteristic features of the online social systems: the networks dynamics as well as the information cascade. A specific attention will be given to the definition of the phenomena that we propose to study in this dissertation, so the concepts of misinformation, disinformation and malinformation, following an extensive discussion about how to fight information disorders and the approaches and data used in the thesis.

The third chapter of this dissertation will investigate how the conversation around the Covid-19 Vaccination is characterized in terms of emotional responses with respect to different features of users involved in the conversation.

- RQ1: How does the online emotional response differ from a baseline when discussing about highly debated topics?

The fourth chapter builds upon the previous work, investigating the conversation around Long COVID on Twitter/ X, by exploring the type of news shared (scientific or more controversial), the emotional response and the information cascades dynamics of the conversation.

- RQ2: How is the conversation around Long COVID characterized with respect to the baseline data? What type of information and news shared prevail in the discussion?

The fifth chapter investigates a new methodology to detect the news media diet of different users, classified as high-risk users when sharing unreliable content and reliable or low-risk users when sharing only reliable content. Using entropic measurements, two indicators are introduced to characterize individual news media consumption patterns. *Variety* captures the diversity of sources a user engages with and is calculated through the entropy of the distribution of consumed outlets, with higher values indicating broader and more evenly distributed exposure. *Regularity* reflects the temporal consistency of source selection and is measured through sequence-based entropy, quantifying the stability of a user's consumption behavior over time.

- RQ3: To what extent do users differ with respect to the way they access and share information online? Is there any different behavioral pattern between users

sharing reliable sources and those sharing unreliable sources? How can we quantify the *variety* and *regularity* of the news media diet of users?

The sixth chapter provides an evaluation and validation of previous infodemic measures [2], which enable us to quantify the risks and exposure to unreliable content, in each country by using online social media data. Few studies have highlighted how infodemic patterns relate to countries' socioeconomic structures, leaving a critical gap in understanding national vulnerabilities. For this reason, we validate these previous infodemic metrics by using socio-economic indexes to answer:

- RQ4: How the infodemic patterns evolve over time? Do countries cluster based on their infodemic and socio-economic values? How do correlations between infodemic metrics and socioeconomic factors evolve over time?

Following the validation of Infodemic Risk Index, in the seventh chapter, we expand the contribution of this indicator, not only focusing on the average value of risks that countries manifest daily, but also on the *volatility* of this index to understand the key features of the online social media that enable the index to fluctuate more. In this context, we propose a stochastic model to reproduce the average and variation values of the empirical findings for each country. We will use statistical properties of the systems such as power law distribution of the followers and characteristics of the systems itself, such as *engagement*, and *exposure* to specific type of reliable or unreliable content.

- RQ5: Why are some countries suffering from a higher instability in the infodemic risk index? What are the key properties of the systems itself that can explain to a certain extent this variability?

The eighth chapter shifts the attention on the second information disorders introduced above: the risks of being immersed in community hub, where people share the same beliefs and knowledge. In this specific case, I introduce a methodology to observe the shift from mainstream to more radicalized communities, investigating the behavioral changes in terms of language, toxicity, hate speech of the users. I will take as a case study the subreddit *Incels* and track the users that joined that community in their behavior before and after being exposed to such community. On the first part, I will focus on the characterization, while the second is about the detection of early signals that might be predictive of such transition.

- RQ6: What are the effects of user migration from traditional online communities to radicalized ones?

- RQ7: What key characteristics distinguish users who become deeply involved in Incels community from those who do not?

Taken together, these research questions address different but complementary dimensions of information disorders. By combining the characterization of online conversations during critical events, the measurement and validation of infodemic risk indicators, and the modeling of dynamic processes underlying both misinformation and radicalization, this thesis provides an integrated perspective on the risks and dynamics of information disorders in digital ecosystems. The empirical analyses are complemented by methodological and modeling contributions, which not only enhance our understanding of how misinformation and radicalization emerge and spread with respect to certain users and contexts, but also offer quantitative tools to anticipate and mitigate their societal impact. In this way, the thesis aims to advance both the theoretical foundations and the practical relevance of computational social science in tackling some of the most pressing challenges of the digital age.

1.2 List of Publications

This dissertation builds upon the works carried out during the last 3 years of my PhD. Some of these works are not published yet but ready to be submitted to scientific journal soon.

- A. Bertani, R. Gallotti, S. Menini, P. Sacco, M. De Domemico, "Large-scale digital signatures of emotional response to the COVID-19 vaccination campaign." *EPJ Data Science* 13.1 (2024): 20.
- A. Bertani, V. Mazzeo, and R. Gallotti, "Decoding the news media diet of disinformation spreaders." *Entropy* 26.3 (2024): 270.
- A. Bertani, P. Sacco, M. De Domenico, "Digital Monitoring of a Rising Global Disease: A Twitter-Based Study of Long Covid Conversations" (Under Review at *Journal of Medical Internet Research*).
- A. Bertani, A. Cortese, F. Pilati, P. Sacco, R. Gallotti, "The interplay between socioeconomic factor and information risk indices" (To be submitted).
- A. Bertani, R. Gallotti "A Stochastic Model for Understanding Volatility in Infodemic Risk Index" (To be submitted).

- A. Bertani, A. Flammini, F. Menczer, "The Radicalization Pipeline: Characterizing and Predicting online misogyny" (Working Paper).

Other publications not covered in this thesis:

- L. G. Alvarez-Zuzek, J. P. Bascur, A. Bertani, R. Gallotti, V. A Traag. (2025). Mapping the interaction between science and misinformation in COVID-19 tweets. arXiv preprint arXiv:2507.01481.

Chapter 2

Background and Conceptual Framework

2.1 Social Media as an Information Ecosystem

In the last 20 years, online social media platforms have emerged as an important digital arena where the information flows and is shared among different users, significantly contributing to an easier access to information. Despite traditional media, which typically features an unidirectional flow of information from producers to consumers, social media enables a more interactive and participatory approach. Indeed, users can create, share and engage in discussions, transforming these platforms into dynamic spaces for public opinion formation [21]. According to Statista, more than 64 % of population use social media, highlighting the role that these platforms play in everyday communication and exchange [22]. In fact, social media are recognized as powerful channels through which communication travels rapidly and extensively. They not only facilitate the dissemination of information, but they also shape how narratives are defined and constructed. The interactivity of such platforms makes them an important resource for agenda-setting [23] as well as collective action and mobilization [24]. Beyond connecting citizens more easily, social media platforms are also ecosystems in which users consume and share news on a daily basis. A recent report by the Reuters Institute of Journalism revealed that the engagement with traditional media such as TV and news websites continues to fall, while dependence on social media, video platforms, and online aggregators grows. This is particularly true for the case of United States where 54 % of users access information through social media and videos, overtaking both TV news (50%) and news websites/apps (48%) for the first time [11]. While all social media

are characterized by speed and user engagement, they differ from one another in their underlying digital ontology. By digital ontology we refer to the specific affordances [25], logic design [26] and algorithmic architectures [27] that shape how communication manifests, what type of content circulates, and which type of interactions are made possible on each platform. In the following section, I will introduce the characteristics and nature of two different online social media platforms, which have been subjects of the research conducted : **X** (formally known as Twitter) and Reddit.

2.1.1 Characteristic of different online social platforms

Twitter is a micro-blogging platform launched in 2006 where users can register a personal account and can share posts, also known as *tweets*. Unlike other social media such as Facebook, the relationship of following and being followed requires no reciprocation. A user can follow any other user, and the user being followed need not follow back. Being a follower on Twitter means that the user can easily see and access posts from those the user follows [28]. Users can decide to have a private or a public account and can also ask for a verification of their own account. Until April 2023, verification was indicated by a blue checkmark placed next to the username and was granted directly by Twitter to accounts considered authentic and notable, such as public figures, companies, or influencers, based on public interest and notability. Following Elon Musk’s acquisition of the platform, however, the verification is now tied to a paid subscription service called Twitter Blue [29]. A key relevant affordance of such platform is the use of hashtags (terms preceded with a hash mark # such as #covid-19 or #1Oct), which allows user to catalyze the attention around a specific topic. Twitter has been recognized as a digital platform attracting different public figures such as activists [30] [31], political candidates during election campaigns [32], scientists during health emergencies , and influencers as well as the generic audience. For this reason, many scientific researchers have been focusing on this platform, taking advantage both of the huge amount of users active but also of the relatively easy access to the data through Twitter API (Application Programming Interface), that used to retrieve up to approximately 10 million of tweets per month.

On the contrary, Reddit is a forum-based social media platform where users can create anonymous accounts and choose to subscribe to specific subreddits, which are communities centered around particular topics such as movies and politics among others. An examples is also gaming-related subreddits, where enthusiasts discuss different types of games and provide advice on new releases. In this context, users participate to

the online communities typically by posting questions or sharing experiences to receive comments and advices from others. Each submission can be upvoted or downvoted by the members’s community, which influences its visibility, given by the aggregation of these votes. Reddit is characterized by a strong participatory culture, where discussions are shaped by collective evaluation and feedback. Indeed, many subreddits benefit by having a group of users called moderators who participate in a variety of tasks that include creating subreddit rules, removing content that violates rules, and responding to user inquiries and complaints [33].

2.2 How does online social media work?

2.2.1 Algorithm and Virality of online content

People spend an incredible amount of time online, specifically on social media platforms. On average users spend about 141 minutes on social networks on a daily basis [22]. This is due to several factors, such as the gratification that users receive in being online and the engagement that the platforms cause. According to the uses and gratification theory [34, 35], users are intentionally active online because they can engage with forms of content most likely to meet their needs (e.g. social interactions, information seeking and entertainment, self-expression and surveillance of others [36, 37]). On the other hand, social and communication theories have defined the concept of *engagement* as a form of active interactions such as reposting other user content, responding to others, posting new pieces of content such as videos, texts, and images [38]. Social media platforms are deliberately designed to exploit these motivations through algorithmic recommendation systems that govern the visibility of posts in users’ feeds, prioritizing content based on predicted interest, engagement, and relevance. Algorithms rank the content according to the past digital behaviour of the user, based on the interactions with similar topics or accounts. This increases the probability that users might be exposed to similar narratives and views, reinforcing their beliefs and increasing their probability of re-sharing that content.

Another characteristic aspect of modern platforms such as Twitter/ $\mathbb{X}$, TikTok and Youtube is the reliance on personalized feeds and “for you” recommendations, where trending or viral content is strategically used to maximize the attention. This contributes to a self-reinforcing cycle of visibility, in which highly engaging posts, often those that are emotionally charged or sensational, achieve disproportionate reach through virality [39, 40]. The algorithmic mechanisms are also combined with the low-

attention behaviors of users, which interact with these platforms in a face-paced and distracted way. This could have an effect on the way they share content, highlighting how time constrain and distraction can decrease the attention through which users are able to discern a true from a false news [41]. All these features highlight how the algorithms and virality mechanisms jointly transform social media into ecosystems where information disorders can rapidly emerge and spread.

2.2.2 Information Diffusion and Cascades

Beyond recommendation algorithms and engagement-driven design, another central feature of online social media is the way information propagates across networks. Information does not diffuse randomly: it follows structured pathways shaped by the network topology and user interactions. These processes, often referred to as *information cascades*, capture how messages, news, or narratives can rapidly spread from a small number of initial sharers to potentially millions of users. While the quantitative study of diffusion has long been a topic of interest, the large-scale analysis of the structure of diffusion processes has its origin in the study of Web and particularly on the spread of blog-links [42]. Since then and after the advent of social media, the scale of data has drastically increased, allowing researchers to investigate the spreading dynamics of a wide range of topics and content. This includes research on how videos, images, news and petitions spread on Twitter [43] as well as how false and true news spreading online [39]. By quantifying the statistical properties of information cascades, we can observe different mechanisms driving the diffusion of such content. More recent work has emphasized the importance of comparing cascades of equal size before drawing conclusions on content-specific differences, in order to disentangle structural effects from topical ones [44].

2.2.3 Emotional Diffusion of Online Content

Beside the structural features of information cascades, another relevant aspect is the role played by emotions in online conversation and news media. Emotions such as fear, anger, or joy are powerful drivers of online engagement, making the study of emotional diffusion a fundamental step in understanding the risks of information disorders. Prior research shows that news consumers tend to prefer and share emotionally arousing content, and that news coverage has increasingly shifted toward more sensationalist and negative tones [45]. The arousal created by news does not merely expand audience

size but it also exacerbates motivated reasoning, reinforcing prior beliefs and contributing to political polarization and social fragmentation [46]. The analysis of emotions has become a central focus in both computational social science and communication studies, as scholars seek to explain not only how information spreads, but also why certain narratives attract particularly a certain audience. There is a rapidly growing literature that is exploring various methodological approaches to the measurement of emotions and of their socio-behavioral effects in online interactions, with special emphasis on contentious and polarizing topics. The work presented in [47] introduces a method using change point detection and topic modeling to measure online emotional reactions to offline polarizing events such as the COVID-19 pandemic and racial justice protests, effectively dis-aggregating different topics to measure the emotional and moral reactions of the public opinion. Guo et al.[48] use natural language processing methods to highlight a surprising rise in positive affect during the early stages of the COVID-19 pandemic, likely reflecting the role of social media as a tool for emotion regulation and reducing fear of the unknown during the pandemic, while also revealing a partisan divide in emotional and moral reactions. Vemprala et al. [49] also develop a natural language processing approach to the study of public negative emotional responses during the COVID-19 pandemic to discern how emotional patterns react to the prevailing focus of the public discussion being placed upon health vs. economic concerns, with fear dominating in health-related conversations, and a more complex mix of emotions, mainly anger and fear, emerging in economics-related ones. Wang et al. [50] also develop a natural language processing approach based upon a BERT transformer model for sentiment classification to show that the COVID-19 pandemic caused a significant drop in expressed sentiment globally, followed by an asymmetric, slower recovery, with minor effects of sentiment expression related to lockdown policies, although with significant variation across countries. Zhang et al. [51] use instead a machine learning classification model based upon deep learning language models to identify positive and negative emotions in online conversations about COVID-19, and including an additional, so far not analyzed, ambivalent emotional expression (joking), finding a rapid burst and a slower decline in the online conversations in all of the 6 languages they analyze. In this thesis, I will analyze the study of emotions with respect to two important online discussions, the COVID-19 vaccination campaign in Chapter 3 and the Long COVID discussion in Chapter 4.

2.3 Information Disorders

2.3.1 Definitions and distinctions: misinformation, disinformation, malinformation

The terminology associated with the concept of *information disorders* is often confusing and lacks a shared interpretation among experts. Despite this difficulty, establishing conceptual clarity is essential to avoid oversimplification and to frame the complexity of the problem. It follows a review of the definition and distinction of terms most used in the field of communication.

The most used term in both academic and public discourse is *fake news*, defined as "false news that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke"¹. Despite this definition, according to Wardle and Derakhshan [52], fake news is an inadequate term to describe a complex issue like that. In particular, Wardle stated that this idiom is too vague, as it has been employed to describe a variety of issues, from intentionally misleading stories to biased reporting. This vagueness makes it difficult to understand the specific problem being addressed [1].

In fact, Tandoc Jr., Lim, and Ling [53] scrutinized 34 academic articles spanning from 2003 to 2017, shedding light on the diverse interpretations and contexts associated with the term "fake news". Their analysis uncovered instances such as satire, inadvertent or hasty sharing of misleading content, and purposeful creation of false information with the intent to harm others. Additionally, the authors highlighted that the term extends beyond mere falsehoods to encompass inaccurate correlations lacking substantial evidence. Wardle further emphasizes that the term "fake news" has acquired a politically charged connotation, frequently wielded by various political figures to undermine journalists and news outlets with critical or opposing perspectives, thereby fostering a political rift on the subject. Furthermore, critics argue that the term "fake news" oversimplifies a complex reality, implying a stark dichotomy between "real" and "fake" news, whereas the actual landscape often involves nuanced shades in between.

Wardle stated that the term "fake news" implies a passive stance, attributing the issue solely to the media or other information sources. It is essential to underline that individuals also play a role in propagating false and misleading information through the act of sharing and interacting with such content. Due to these considerations, Wardle

¹Source: Cambridge Advanced Learner's Dictionary & Thesaurus (<https://dictionary.cambridge.org/dictionary/english/fake-news>)

advocates for the adoption of the term "information disorder" as a more unbiased and informative expression [1].

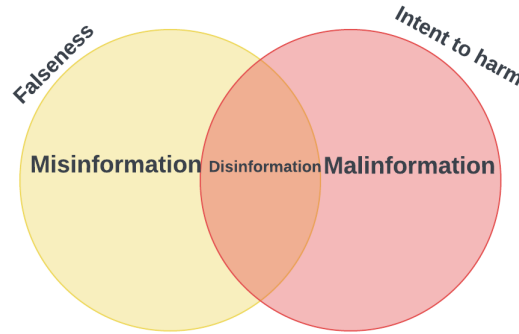


Figure 2.1: **Information Disorders.** Visualization of the intersection of the three key component of disinformation disorders: Misinformation, Disinformation and Malinformation [1]

In turn, this expression contains three different and more precise notions:

- **Misinformation:** This refers to false and inaccurate information disseminating unintentionally, often stemming from a lack of knowledge, misunderstanding, accidental oversight, or casual and thoughtless actions. An example could consist in sharing a rumor about a celebrity's demise before realizing it was false.
- **Disinformation:** This involves the deliberate spread of false or deceptive information with the intention to mislead. Disinformation is commonly employed as a tool for propaganda or manipulating public opinion. Those behind the creation and dissemination of disinformation often pursue political, financial, or psychological gains [54]. Coordinated disinformation efforts can extend across numerous social media accounts and may utilize automated tools like social bots, software mimicking human behavior on social networks, to post and share content online, presenting themselves as authentic disinformation campaigns. Periodically, organizations such as the Internet Research Agency, a well-known Russian troll factory, orchestrate these campaigns with the goal of manipulating public opinion and fostering discord by generating and spreading inflammatory or misleading content on social media platforms [55].
- **Malinformation:** This denotes the intentional communication of accurate information with the aim of causing harm or distress to an individual. Malinformation encompasses actions like divulging private or confidential details, leaking

documents, or sharing embarrassing or detrimental content. An instance of misinformation is exemplified by doxing, where private or personally identifiable information (such as full names, addresses, phone numbers, photos, etc.) about an individual is published online without their consent.

2.3.2 Mitigating the spread of mis- and disinformative content

Several approaches have been identified to reduce the spread of mis- and disinformative contents. One of the most common and well-established methods to fight the spread of false information is *fact-checking*, defined as the process of verifying the accuracy and the truthfulness of claims, statements and narratives circulating online.

Fact-checking is a fundamental practice of journalism and the online media industry and can be done either before the content is disseminated (*ante hoc*) or after it has been spread (*post hoc*). The fact-checking procedure typically involves a complete investigation of the claim or the information diffused, including the identification of the original source, reviewing expert opinions and cross-referencing the information with other reliable sources. The experts then evaluate the gathered evidence and reach a shared consensus about whether the claim offers a faithful representation of reality or misleading point of view and help to provide a more precise context or necessary clarifications in case of partially true statements [56]. Despite the rigor of this process, fact-checking analysis turns out to be expensive, time-consuming and an intensive activity which has not been always effective [56]. There is also the belief that ex-post corrections are often ineffective because, despite the corrections, people tend to keep relying on the information they know to be false, also known as the *continued-influence effect* [57]. Another widely recognized approach is debunking, a method that involves actively challenging or refuting false, exaggerated, and anti-scientific assertions using critical thinking, logic, and factual data. Debunking typically requires identifying and revealing the shortcomings or inaccuracies in the initial claim, presenting evidence, or proposing alternative explanations that contradict it. The ultimate goal is to rectify the perspectives of individuals exposed to misinformation and potentially hinder its further dissemination [58]. However, individuals often continue to resist factual information, maintaining inaccurate beliefs even when presented with evidence to the contrary [59]. In many cases, corrective interventions fail to reduce these misperceptions [60] and may even produce a backfire effect, reinforcing the very inaccuracies they aim to correct [61]. Another possible *ex-ante* correction is referred as prebunking, involving the proactive education of individuals about potentially misleading information before its dissemina-

tion [62]. This concept is also referred to as "inoculation theory", suggesting that when individuals are preemptively exposed to a properly diluted form of misinformation or disinformation, they begin to develop "mental antibodies" through the elicited cognitive response. These mental antibodies enhance individuals' immunity to subsequent attempts at manipulation.

2.3.3 Countering the false information with Artificial Intelligence

While Artificial Intelligence can contribute to the dissemination of misinformation [63], it also plays a pivotal role in its prevention. Machine and Deep Learning algorithms aim to identify, flag false information, predict viral potential, and offer fact-checking services [64]. Various methods, including the five feature extraction approaches listed by Shu et al.[65], have been employed for automatic fake news detection on social media platforms.

- *linguistic-based* approach relies on linguistic features that capture distinct writing styles and sensational headlines commonly associated with fake news. This approach operates under the assumption that certain linguistic behaviors, such as punctuation usage, word choice, part-of-speech tags, and emotional valence, are largely involuntary and beyond the author's conscious control. As stated by Pérez-Rosas et al. [66], these features offer valuable insights into the nature of the text. In this context, Natural Language Processing (NLP) algorithms can analyze extensive text volumes, revealing patterns indicative of false information.
- the *visual-based* approach extracts information about visual elements, such as images or videos.
- the *used-based* approach identifies the features of the users interacting with the news on online social media, like the account creation date, their number of followers/followees and their verification status. All these features can be really relevant in the detection of automated accounts, which are often used to spread propaganda and disinformative content [67].
- the *post-based* approach focuses on the identification of information exploring the reactions and comments that users shared with regard to a specific post.
- the *network-based* approach where features are extracted by constructing specific networks among users who published related posts, or among users who have a

follower/followee or friendship relationship. Several social network analysis techniques can be used to describe these networks, which can provide extremely useful insight, given the tendency of social media to form echo chambers.

2.3.4 Related works

In the recent years, the literature about the spread of mis- and disinformative content has been in rapid expansion. More specifically, the 2016 election marked a crucial moment in the proliferation of "fake news" through social media, and numerous studies have delved into scrutinizing this phenomenon in relation to that specific event. In a study conducted by Grinberg et al. [32], the dissemination of false news articles on Twitter during the 2016 U.S. presidential election was examined. The findings indicated that fake news comprised nearly 6% of overall news consumption, but its distribution was highly concentrated. Specifically, only 1% of users were exposed to 80% of fake news, and a mere 0.1% of users were responsible for sharing 80% of fake news. Significantly, the concentration of fake news was particularly pronounced among conservative voters.

Another investigation by Bovet and Makse [68] analyzed over 30 million tweets containing links to news outlets shared in the five months leading up to the same election. The authors determined that approximately a quarter of these links directed to websites categorized as disseminating fake news or exhibiting extreme partisan bias. Additionally, a causal analysis unveiled that the primary disseminators of news from center and left-leaning outlets significantly influenced the Twitter activity of Clinton supporters. In contrast, the top spreaders of "fake news", influenced by the Twitter activity of Trump supporters, had a notable impact in the opposite direction. Other research has investigated the propagation of misinformation and disinformation within the domains of health and science. Wang et al. [69] documented the efforts of many scholars who examined online dialogues surrounding topics such as vaccinations, the Zika virus, and Ebola. In the last years, Covid-19 Pandemic has been one of the most pressing case studies, analyzed by several scholars, using computational social science frameworks. In particular, one of the main works from which we take inspiration is the one of [2], where they proposed three measures to evaluate the infodemic of each country by using features of the users (followers count) and characteristics of the network communication such as retweets, replies, mentions.

2.3.5 The Covid-19 Infodemic Platform

In the context of Covid-19, the concept of *infodemic* has been gained special attention and linked to this health crisis [70, 2, 71]. This term encompasses other two words, information and epidemic and was originally conceived in 2003 by David Rothkopf to describe the virus-like spread of mis- and disinformation in the context of the 2003 SARS outbreak [72]. Rothkopf highlighted the meaning of this word as "a few facts, mixed, with fear, speculation and rumor, amplified and relayed swiftly worldwide by modern information technologies [that] have affected national and international economies, politics and even security in ways that are utterly disproportionate with the root realities". On the other hand, the WHO has defined the term infodemic as "an over-abundance of information - some accurate and some not — that makes it hard for people to find trustworthy sources and reliable guidance when they need it" [73].

With respect to the Covid-19 health crisis, one of the most rich dataset used in this thesis has been collected by using the COVID-19 Infodemics Observatory, a project developed by the Fondazione Bruno Kessler (FBK) in collaboration with the World Health Organization (WHO). The aim of this website was to collect, track and extract information about the infodemic patterns around the Covid-19 Pandemic by extracting daily tweets across countries. The result of this collection ended up in an interactive dashboard, showing the estimation of the infodemic risks in each country on a daily basis, as shown in Figure 2.2. Beside being an easy and accessible tool, it also provides with time series analysis with information about the infodemic itself as well as enlarged by information about the disease such as number of daily Covid-19 cases, number of daily deaths per country.

The platform has collected via Twitter’s Public Filter API over 1.3 billion tweets from 127 countries in the time span ranging from 22 January 2020 to 13 March 2023. To be included in the dataset, tweets have to contain at least one of the following keywords: *coronavirus*, *ncov*, *#Wuhan*, *covid19*, *covid-19*, *sarscov2*, *covid* [2].

The retrieved data also include retweets and replies, which are respectively repostings of an original tweet used to quickly share that tweet with one’s own followers and social interactions that generate and sustain conversations among multiple users. The collected tweets are then geocoded at country-level granularity using either the included geographic information or the self-reported user locations.

To assess the reliability of each message, each has been evaluated for reliability through comparison with various publicly accessible databases covering scientific and journalistic sources. We specifically employed reliability classifications from Media-

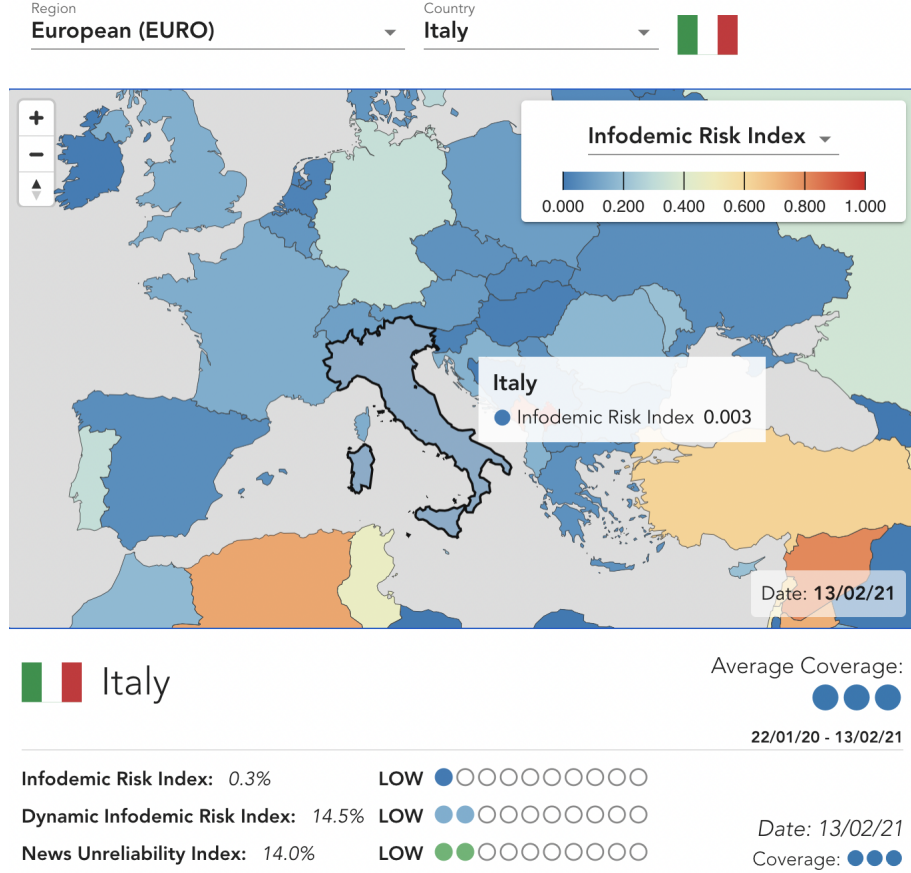


Figure 2.2: Screenshot of the FBK COVID-19 Infodemics Observatory dashboard. Source: <https://covid19obs.fbk.eu/#/press>

BiasFactCheck [74], an organization that maintains an extensive and regularly updated database evaluating the ideological leanings and factual accuracy of media outlets through both quantitative metrics and qualitative assessments.

The categorization can be simplified into *reliable* sources, encompassing science and mainstream media, *unreliable*, differentiated between *low-risk unreliability*, including satire, clickbait, political and *high-risk unreliability*, defined as fake or hoax, conspiracy and junk science.

Building on this classification system, we employed three metrics developed by [2] to quantify different dimensions of infodemic risk.

1. **Infodemic Risk Index (IRI)**: this index quantifies the rate at which a generic user is exposed to unreliable news. It is derived indirectly by the number of followers of users who have tweeted, retweeted or replied to messages containing links belonging to the *unreliable* macro-group as a measure of *exposure* to unreliable news. More formally, if E_U is the sum of the number of followers reached by each

News Category	Level	Definition
Science	Reliable	Domains providing content validated via scientific scrutiny.
Mainstream Media (MSM)	Reliable	Domains providing content that is generally subjected to professional fact checking and abides by the rules of media accountability.
Satire	Low-Risk	Domains providing content that is intentionally and explicitly aiming at providing a distorted representation of events as a form of humour and/or social critique.
Political	Low-Risk	Domains providing content that presents a partisan representation and interpretation of facts to support a political position over rival ones.
Clickbait	Low-Risk	Domains providing content that generally distorts or intentionally misrepresents information to capture attention.
Other	Low-Risk	Domains pointing to general content that cannot be easily classified, such as videos on YouTube.
Fake/Hoax	High-Risk	Domains providing manipulative and fabricated content with the purpose of misleading public opinion on socially relevant issues and provoking inflammatory responses.
Conspiracy/Junk science	High-Risk	Domains providing systematically manipulative and fabricated content with the purpose of legitimizing implausible conceptualizations of facts and knowledge through argumentative methods that coarsely mimic those of scientific reasoning but without any sound logical or factual basis, targeting individuals or social groups as covert instigators or perpetrators of harmful actions.

Figure 2.3: Classification of the news media categories based on their level of reliability based on the work of Gallotti et al. [2].

tweet containing unreliable content and E_R is similarly the sum of the number of followers reached by each tweet containing reliable content, the IRI is simply computed as

$$IRI = \frac{E_U}{E_U + E_R}$$

It is clear to see how a prominent figure with a large following sharing misleading information can significantly increase the IRI.

2. **Dynamic Infodemic Risk Index** (DynIRI): this index is slightly different in the sense that, in this case, *exposure* is calculated by directly analyzing user engagement with unreliable online messages, namely recording the number of retweets, replies and replies that each tweet has. The index is calculated as above, but E_U and E_R represent the total number of interactions (retweets, replies, quotes) recorded in tweets containing respectively unreliable and reliable news. In this case, if a certain number of tweets generate a significant amount of interactions, that could cause an increase in the DynIRI.
3. **Fraction of unreliable news** (Unreliability): this index represents the proportion of unreliable content in relation to the total volume of tweets containing links pointing to any information sources in the domains list. The metric provides valuable insight into the reliability of COVID-19 news being generated and spread in a given country at any point in time. Furthermore, the platform provides more granularity, distinguishing also in each of the nine categories described above. This can allow users to be informed about which specific types of news (mainstream media, clickbait, ...) are consumed in a certain period in their area.

These metrics allow us to study the risks of infodemic in each country by focusing on two different fundamental aspects of Twitter: the role of followers, tracked by IRI, and the role of the connectivity of the systems, defined through the Dynamic Risk Index. In Chapters 6 and 7, I will discuss and validate how these indexes are correlated with socio-economic factors as well as the digital characterization of different countries based on the instability of such indicator.

Chapter 3

Large-scale digital signatures of emotional response to the COVID-19 Vaccination Campaign

The previous chapter outlined the theoretical and conceptual framework for understanding information disorders as a complex and emergent phenomena within social media ecosystems. This chapter introduces the first empirical study of this thesis based on the journal paper: Anna Bertani, Riccardo Gallotti, Stefano Menini, Pierluigi Sacco, Manlio De Domenico. Large-scale digital signatures of emotional response to the COVID-19 vaccination campaign. *EPJ Data Science* 13.1 (2024): 20 [75]. The main focus here is to analyze how emotional tone, users features, and news reliability interact in a major global health online conversations. The main contributions of this chapter are:

- we develop a methodology combining content reliability, political leaning, and emotional valence to study online vaccination discussions.
- we reveal patterns in emotional responses of vaccination conversations across political orientations and reliability of information .
- we provide insights for public health communication, showing how social media can be used to monitor public attitudes and counter misinformation.

3.1 Introduction

Face-to-face interaction is notoriously important to facilitate civilized exchange and social cooperation [76, 77]. Through their nonverbal language [78], interacting humans

send complex arrays of signals (of dominance, trust, composure etc.) [79] that favor mutual alignment [80] and even elicit behavioral mimicry [81]. For these reasons, face-to-face interactions are normally expected not to escalate into violent and confrontational behaviors [82], and even function as a driver of social cohesion [83] with distinctive psycho-physiological signatures [84]. With the advent of online interactions, however, this carefully evolved package of socio-cognitive skills has been put to a hard test. In digital interaction, the moderating role of nonverbal cues is largely lost [85], and subjects must learn how the affordances of digital communication enable alternative ways to signal pro-sociality and affection [86]. However, developing alternative systems that work on a large social scale is challenging, and as a result it is widely observed that digitally based exchanges have higher chances to become vitriolic and prone to escalation than face-to-face ones [87].

In this context, a particularly crucial role is played by emotions. The once widely held conviction that human emotions were biologically programmed, universal ‘natural kinds’, spanned by a core group of six basic emotions and characterized by specific, inter-culturally readable sets of bodily cues, has gradually given way to an alternative paradigm that identifies a whole spectrum of emotional states mapped by a high-dimensional set of verbal and nonverbal signals [88] whose socio-cognitive indexing depends on social learning and cultural frames [89]. In view of the central role of emotions in social cognition [90], the use and interpretation of such emotionally-related signals is pivotal in human interaction. In particular, the expression and decoding of emotions is likely to be one of the most critical aspects related to the shift from face-to-face to online interactions [91].

If emotions are not ‘natural kinds’, a radical change in the socio-cognitive environment such as that brought about by the digital shift may not only affect how emotions are interpreted, but also how they are defined [92], expressed [93], and socially transmitted [94]. Despite that the affordances of the digital environment facilitate emotional over-reaction [95], it is also true that, unlike face-to-face interaction where nonverbal signals are observed at a millisecond scale and are largely automatic and not controlled [96], in online interaction the emotional response may be distilled and even constructed on the scale of seconds and even minutes or hours [97]. On the other hand, nonverbal cues are largely substituted by verbal equivalents that intentionally amplify the affective content of communication to compensate for the lower-dimensional nature of the signal [98]. Moreover, in online interactions some of the participants may be nonhuman (bots) which are explicitly designed to implement a certain strategy of affective communication to elicit types of emotional responses from users [99].

There is then a basic difference between emotions expressed in face-to-face interaction and ‘emotions’ as constructed in online exchanges [100]. And such difference is likely to be more substantial and relevant the more the topic of the exchange concerns socially sensitive, polarizing issues [101, 102]. In particular, fake or misleading content that embeds certain emotional references (which can be positive such as anticipation or trust, or negative such as anger, according to cases) is more likely to go viral and has a longer expected lifetime [103].

Moving from this premise, the aim of this research is to investigate how online ‘emotional’ responses change when comparing two different kinds of exchanges: one regarding a baseline, non-polarizing topic and one regarding one of the most polarizing topics of today: vaccination [104]. It is intuitive to conjecture that more socially controversial issues incite more emotional reaction (and vice versa) online than less controversial ones [105], not unlike what happens offline, although specific features of the online interaction environment may make a significant difference [106]. However, it is of interest to understand what kind of emotional reaction is incited in what circumstances, depending on the characteristics of the users involved. The emotional response patterns that are found give us additional insight on the nature of the social contentiousness of the issue, and at the same time help us understand better how online ‘emotions’ are constructed to pursue specific social goals. In this regard, vaccination is clearly a relevant test bed [107], and we are in particular interested in studying social exchanges on Twitter with special attention to the verified/un-verified user and human/bot dyads. To what extent the emotional communication of verified users will distinguish itself from that of un-verified ones? Will there be major differences between human and artificial ‘emotional’ responses when the topic is potentially more inflammatory, and in which sense?

In this rapidly evolving field of research, the specific contribution of this chapter is making use again of a natural language processing approach and of the same lexicon used by [49], NRC Word-Emotion Association Lexicon, to analyze specifically the differences in emotional expression between a baseline non-contentious topic and a highly polarizing one such as COVID-19 vaccination, and testing specifically for differences related to whether online participants are humans or bots.

The NRC Lexicon considers eight ‘emotions’, four of which with positive valence and four with negative valence. With this choice, we intentionally move from the six basic emotions of Ekman [108] not because we agree with the ‘natural kinds’ framework (within which there is a significant lack of consensus about the specific list of what emotions are basic; see [109]), but because this may be a simple benchmark for the

analysis of their online counterparts. Specifically, the four ‘basic’ negative emotions (fear, anger, sadness, and disgust) and one of the positive ones (surprise) are kept in the NRC Lexicon, whereas the other (happiness) is unpacked into two positive emotions (anticipation and joy). Finally, trust has been added as a final positive emotion. Despite that trust is technically not considered an emotion [110], it can be interesting to consider a specific emotional signal related to trust in the relatively unfavorable conditions given by social signaling through a digital medium. This set of ‘emotions’ as encapsulated in the NRC Lexicon should therefore be seen as a useful first benchmark rather than as an invitation to consider them as more basic or foundational than others.

Within this framework, we find a significant difference in emotional reactions between no-contentious and polarizing conversations, and moreover we find that verified users, no matter whether humans or bots, exhibit more positively valenced emotional responses than unverified ones. In this regard, this research provides fresh insights on specific aspects that have not been covered in the previous literature, while at the same time connecting to a solid and growing stream of studies both in methodological and thematic terms.

Getting deeper insights into these issues is crucial for the future design of environments that favor more civilized online interaction. As argued by [111], civilized need not amount to ‘polite’ according to pre-digital standards of social etiquette. Online discussion may be more emotionally charged than offline ones for reasons that mostly concern the different affordances of the social environments [112]. However, as the experience of the pandemic has taught us, the social implications of massive uncivilized exchange on issues of primary public interest may be devastating [113] and may offer ample opportunities for manipulation by malevolent parties [114]. Therefore, a deeper understanding of the socio-emotional grammar of online interactions is a key issue for both computational social science and public policy.

3.2 Method

3.2.1 Overview of the data set

We collected social media data through the special Twitter’s endpoint dedicated to COVID-19 research ¹, which allowed researchers to study the comprehensive, public conversation about COVID-19 in real-time. We focused our attention on data captured

¹https://blog.twitter.com/developer/en_us/topics/tools/2020/covid19_public_conversation_data

by the filter of the Twitter Firehose on COVID-19 in 18 among the most represented languages on Twitter. More specifically, we focused on terms related to the vaccination, to anti-vax campaigns but also to the most known vaccine brands, such as Pfizer, Astrazeneca, Johnson&Johnson, Moderna, Sputnik V (see further details in the SI Appendix).

In addition, we considered only a small fraction of the overall data, the about 1% of tweets that are geotagged, to guarantee an accurate information also as to their location, as signalled by the user’s device. The data collected covers the period between August 31, 2020 and July 15, 2021, that is from the announcement of the availability of the first COVID-19 vaccines up to the peak of the vaccine campaign in Europe and in the United States. We compare our sample of tweets related to the vaccination topics with a 10 million of messages posted on Twitter as baseline sample in which no specific contentious topic dominates. The baseline sample was obtained using the stream API without specific filtering or keyword searches. As a result, the dataset encompasses a diverse range of languages, including but not limited to English, Japanese, Spanish, Korean, Portuguese, Thai, Indian languages, French, German, and Italian, among others.

Dataset	Messages	Unique Users	Timeframe
Vaccine	9.6 million	3 million	31/08/2020-15/07/2021
Baseline	10 million	6 million	21/04/2021-24/04/2021

Table 3.1: Statistics about the datasets. The table shows the number of messages, unique users and the timeframe considered for both the datasets.

3.2.2 Verified Accounts

Verified users are those having the blue verified badge, a blue check mark, that defines those accounts that are of public interest because they are considered authentic, notable and active on Twitter. The verification process of users is given by the blue check mark that can be found next to the username, while unverified accounts do not have this distinctive signal. However, this definition pertains to the period before Musk’s takeover, during which accounts were required to undergo request verification. Starting from April 1st, 2023, there has been a change in the rule. For users to acquire the verification badge now, they are required to subscribe to Twitter Blue [29]. Regarding this research, we adhere to the initial definition since the new rule was not in effect during the considered time frame.

3.2.3 News Reliability and Political Leaning

In this analysis, we also considered some metadata associated to the textual content of the messages.

We collected manually checked web domains from various publicly accessible databases, encompassing scientific and journalistic sources. In particular, we examined data provided by the MediaBiasFactCheck [74]. This is an organization that provide a huge database continuously updated whose methodology is to systematically evaluate the ideological leanings and factual accuracy of media and information outlets through a multi-faceted approach that incorporates both quantitative metrics and qualitative assessments. We found a total of 4,988 domains, reduced to 4,417 after removing hard duplicates across databases. Given the nature of our multilingual and multicultural analysis, we evaluated the language coverage of the web-domains classified, taking into account the English centric nature of the web. Building upon [2], we gathered statistics from Amazon Alexa (www.alexa.com/topsites/countries) about web traffic (the top 50 most visited websites) for all countries across the globe, matching these lists with the list of domains used in our analysis. For 127 countries we found at least one domain in the reliable top-50 news source, and for 21 (iso2 codes: AE, AR, BB, BE, CA, DK, FR, KE, MX, NG, PA, PE, PH, PR, PT, QA, SD, SE, TT, US and VE) they have at least one domain in the top-50 websites labelled as unreliable. In fact, this is a lower bound, because Alexa provided only major domains, disregarding subdomains that we instead classified as well. This large presence among the very top tier of websites suggests that the results are robust for multilanguage/multicultural analysis.

We considered the URLs contained in messages, and labeled each URL according to the political leaning (left, left-center, neutral, right-center and right) of its media source and the type of source (political, satire, mainstream media, science, conspiracy/junk science, clickbait, fake/hoax) as manually classified by external experts. In particular, building on [2], we have classified our news sources as *reliable* (when belonging to the Science, Mainstream Media categories) and *unreliable* (when belonging to the Satire, Clickbait, Political, Fake or Hoax, Conspiracy and Junk science and Shadow categories). Finally, as a third category we consider tweets not containing any url to be classified (i.e. mere opinions without a source). We excluded all the web-domains classified as 'Other' as they point to general content that cannot be easily classified, such as videos on YouTube or post on Instagram.

3.2.4 Bot Detection

Social bots are automated accounts that mimic human behavior, create posts, comments and likes on social networks. Analyzing the role of social bots in the emotional response to controversial topics such as those studied in this research is important, given that they are being systematically used for the manipulation of the public opinion through social media [115]. In particular, bots have proven successful in spreading low-credibility content by strategically targeting influential human users [116]. Their typical mode of operation also includes amplification of inflammatory contents by human users [117], whereas actual instigation of emotions is rarer. However, cases of successful bot-to-human transmission of negative emotions like anger have been documented in online COVID-19 related conversations [118]. Bots can therefore be considered a significant threat to public health [119], whose mode of operation also includes emotional manipulation. It is therefore important to investigate the potential role of bots in the social dynamics of emotional responses on controversial topics.

To distinguish a bot from a human, we chose some criteria associated to some forms of unusual social behavior. In particular, based on [114], automated accounts tend to show an important productivity on social media by posting excessive number of content and, especially, in their concentration in particular moments during the day (e.g. overnight or all day long).

We identified automated accounts using a machine learning algorithm designed to classify Twitter accounts as humans or bots and used in previous research [120, 117]. The classification of users into "human" or "bot" is based on ten features that yield the best classification accuracy according to several authors [117, 121]. The features are (1) statuses count; (2) followers count; (3) friends count; (4) favorites count; (5) listed count; (6) default profile; (7) geo enabled; (8) profile use background image; (9) protected; and (10) verified, following the same prescriptions of previous studies [120, 122, 117, 121]. The models undergo training using 80% of the data and are subsequently validated on the remaining 20%. The division between these two sets is performed while ensuring a balance between bots and humans at the level of each individual original dataset. The models gave us the highest accuracy ($> 90\%$) and precision in identifying bots ($> 95\%$) [121].

In a previous work [120], we tested the ability of the algorithm to generalize the classification out of the data sample used for training and validation by applying the model on an independent data set [123], consisting of labeled information about 8,092 humans and 42,446 bots. The results of the classifier are satisfactory, with an accuracy

of 60 %, an F1-score higher than 70 %, and a recall of 58 %. In this specific work, we did not manually inspect the users but we assume that our results would be consistent with the ones found in the above mentioned work.

3.2.5 Emotional Detection

The NLP pipeline used to process the tweets allows processing of text and emotions in multiple languages. For the emotions we rely on the NRC Word-Emotion Association Lexicon [124], containing 14,182 English words associated with one or more of eight basic emotions (*anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy*, and *disgust*) and two sentiments (*negative* and *positive*). These words are then translated into several languages, including the 18 used in this work. A second resource used in the pipeline is the NRC-VAD Lexicon [125], containing 19,971 English words associated with three scores representing respectively valence, arousal, and dominance. As in the previous resource the NRC-VAD is also including the translation of each word in all the 18 needed languages. Taking as input the text of the tweet and its language, the pipeline returns the list of words associated with emotions, the amount of each emotion contained in the message and the total values of valence, arousal, and dominance. Being based on a lexicon, the emotions extraction is preceded by two preprocessing steps in order to increase the amount of emotions retrieved from the text. The first preprocessing step involves elements that are relevant in social media as hashtags and emoji, often carrying an important part of the content of a tweet. To process the hashtags we expand the Ekphrasis library [126], originally developed for English, to cover additional languages. This allows us to identify hashtags in tweets and to split them in the words composing it, e.g. #saysafe into ‘stay safe’. To detect emotions related to emojis we adopted the strategy of replacing them with their textual descriptions in the language of the tweet. The description can then be used to search for matches in the NRC lexicons , e.g. “worried face”. Since the NRC lexicons doesn’t contain all the inflected forms of annotated words, we preprocess the tweets lemmatizing them (including the text extracted from hashtags and emoji), to increase the number of matches with the words in the lexicon. The lemmatization is done using the Spacy library [127]. The pipeline also has the possibility of using a list of words that need to be excluded from the emotions extraction, for instance when working on tweets about Covid, we can decide to exclude words as ‘virus’ being present in data, as topic, regardless of the emotions expressed by the messages.

3.2.6 Emotional Aggregation

Various emotional algorithms have been already tested on different kind of data. One of them is surely the Valence-Arousal-Dominance model [128]. We found that the three sentiments are highly correlated among them in our dataset, as shown in the Appendix A. Based on the purpose of this research, we decide to adopt another algorithm with eight emotions, four of which with positive valence (Trust, Anticipation, Joy, Surprise) and four with negative valence (Anger, Sadness, Fear, Disgust) in order to better capture each single emotion. Before starting the analysis, a preprocessing phase has been fundamental to understand how to normalize the data, since the emotional range of each emotion does not follow the same scale. In particular, we found the total emotional value of the messages posted on Twitter by summing each singular emotion. Then, we divided each emotion by the total emotional value found in order to obtain that the sum of each emotion should be equal to 1. We grouped the positive emotions (Anticipation, Surprise, Joy and Trust) and the negative emotions (Sadness, Anger, Fear, Disgust) into two different categories (positive vs negative emotions). After this procedure, we normalized the valence of each tweet across the emotional range of -1 (completely negative valence) and +1 (completely positive valence). In the emotional analysis of the three reliability-ranked categories of news Fig. 3.1 and Fig. 3.2, we also decided to delete the emotional values equal to 0 because our interest was to observe the distribution of emotional response and the 0 value might indicate the absence of any emotion in the message posted.

3.3 Results

3.3.1 Emotions distribution among the Covid-19 Vaccination and the Baseline sample

We consider online discussions in which no specific contentious topic dominates consisting of more than 10 millions messages posted to a popular microblogging platform, Twitter, between 21/04/2021 and 24/04/2021 from 6 millions of unique users. We also consider the online discussions concerning the massive COVID-19 vaccination campaign between August 31, 2020 and July 15, 2021, that is from the announcement of the availability of the first COVID-19 vaccines up to the peak of the vaccine campaign in Europe and in the United States, consisting of 9.6 million of posts from 3 million of unique users. We find that the conversation related to a contentious topic such as vaccination

3.3. Results

generate a spectrum of emotional response that differs from that of the baseline, as illustrated in Fig. 3.1.

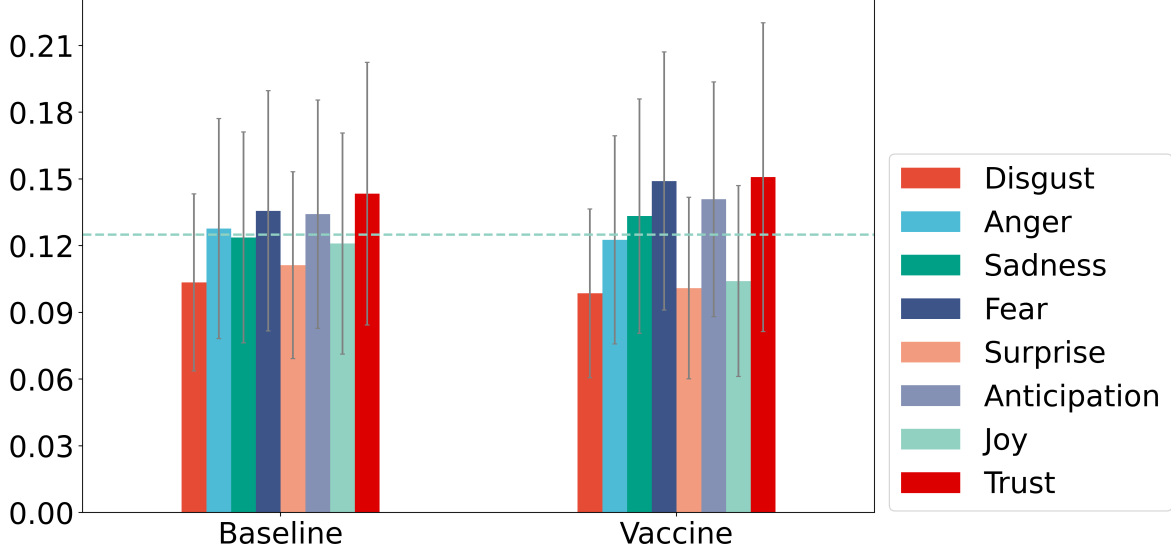


Figure 3.1: **Emotion Mean Value.** For each sample (baseline (left panel) and vaccine (right panel)), we quantified the mean value of each emotion: Anger, Sadness, Surprise, Fear, Anticipation, Trust, Disgust, Joy. The horizontal line is placed at the 0.125 value to better visualize which emotions are over- vs under-represented with respect to our benchmark value. The grey vertical lines show the standard deviation of each distribution. To better understand the statistical significance of each emotions, we performed a Kolmogorov-Smirnov Test with the result that all the distributions are statistically different at the 95 % of confidence level with a p-value lower than 0.01 both across emotions and among the baseline and vaccine samples. In addition, we calculated the standard error of the mean, resulting to be negligible among all the emotion distributions.

In particular, the vaccine sample shows an over representation of three emotions with respect to the baseline: Fear, Anticipation and Trust. Indeed, these three emotions have a mean value greater than the threshold value. If we imagine an ideal benchmark in which each of the eight emotions is equally represented, each of them should be by a frequency of 0.125. We therefore draw this level as a dotted line in the figure to make it more readable which emotions are actually over- vs. under-represented with respect to the benchmark. We can notice that the baseline distribution is more evenly distributed than the vaccination one. In particular, for the baseline we see that Joy, Sadness and Anger are very close to the benchmark value. On the contrary, the distribution is considerably less uniform for the vaccination sample, having an over-representation of Fear, Anticipation Trust and Sadness. This suggests that a contentious topic may lead

to the selective amplification of certain emotions with respect to others. A Kolmogorov-Smirnov test has been performed to evaluate the statistical significance of the two samples. We find that the distributions are actually statistically different between the two samples.

3.3.2 Emotional Responses to Different Levels of News Reliability

At this point, however, it may be reasonable to conjecture that this selective amplification of certain emotions in the case of contentious events may be in turn further modulated by specific features. First of all, we can ask whether getting in contact with unreliable news related to the contentious topic is different than getting in contact with reliable ones as far as the emotional response is concerned.

To test this, Figure 3.2 shows the distribution of emotions for the three reliability-ranked categories of news: reliable, unreliable and opinions. The eight emotions have been normalized in a range between -1 (the most negative emotional valence) and +1 (the most positive emotional valence) to improve readability.

Interestingly, we find that in the baseline sample there are no systematic differences in terms of emotional response to the three categories of news. In particular, the median is 0 for all the news categories, showing that users not only do not discriminate across news categories in terms of emotional response, but also present a balanced overall response to the news, since there is not a skew towards positive or negative responses, indicating a more balanced emotional response, specifically with respect to the vaccination sample.

Also in the case of the differences between reliable and unreliable news we performed a non-parametric Kolmogorov-Smirnov Test, finding that the two distributions are not statistically significantly different, with a p-value equal to 0.22.

In the case of the vaccine sample, the pattern is completely different, and the distributions of emotions are statistically different for each type of news considered.

Reliable news are characterized by being skewed toward the positive valence side of the spectrum when compared to unreliable ones. In other words, reliable news elicit in users more positive emotions in the case of a contentious topic with respect to unreliable ones.

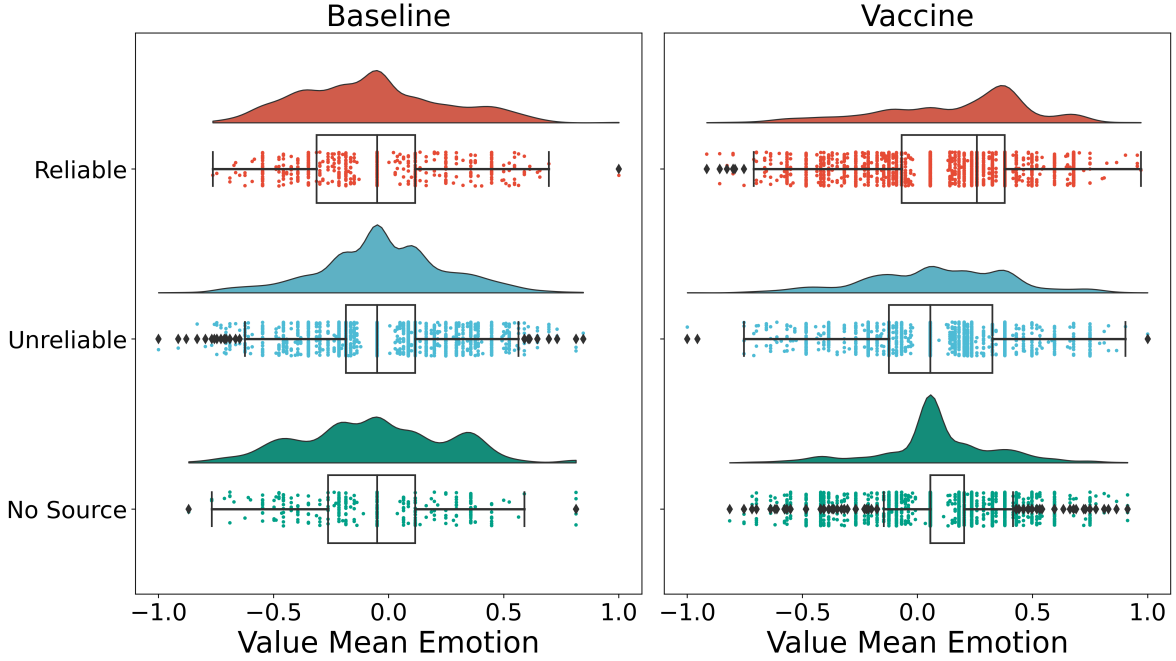


Figure 3.2: **Emotional range for different categories of news.** For each sample (baseline (left panel) vs vaccine (right panel)), we quantified the distribution of emotional values from -1 to 1, respectively for tweets containing reliable news (top), unreliable news (middle) or just comments without sources (bottom). All the distribution have been tested through a Kolmogorov-Smirnov Test in order to evaluate the difference among the samples and across categories of news considered. The distributions are statistically significant at the 95 % of confidence level with a p-value lower than 0.01 among the two samples and within the same sample, with the only exception of tweets containing unreliable news and reliable news in the Baseline sample, having a p-value equal to 0.22.

3.3.3 Emotional Responses of Verified vs Unverified Users

In the Figure 3.3 panel (A), we compare the emotional responses of verified vs. unverified users. We find that verified users are characterized by a more positive emotional response. Once again, the difference between the two distributions is statistically significant as shown by the Kolmogorov-Smirnov Test. Moreover, also for the baseline sample we find more positive valenced responses for verified users (an average level of 0.31) in contrast with the unverified users (0.27).

Interestingly, in the case of the baseline the positive emotional response of the verified users is stronger than in the case of the vaccine sample. Even if verified users, as a consequence of their incentives to reputation management and accountability, tend to favor positively valenced emotional reactions, it turns out that the contentious na-

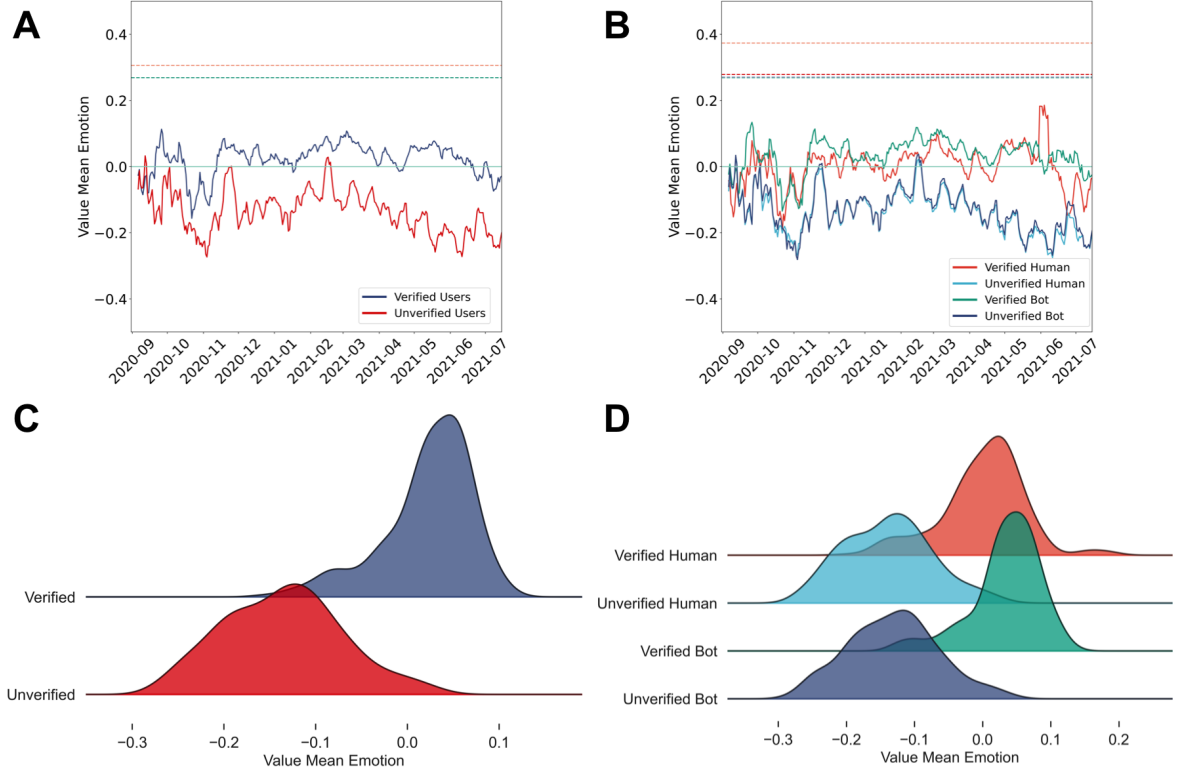


Figure 3.3: Emotional Analysis for different types of user accounts. We quantified, ranging from -1 to +1, across (A) verified and unverified users, (B) both human and bot, over time. The dashed horizontal lines indicate the emotional mean values for each user type, based on a baseline of 10 million randomly sampled tweets. This highlights that the mean emotion of the random distribution is higher than that of our specific interest. We present the distribution of emotional values, aggregated by day, for each category concerning the vaccine sample (C-D). We conducted a Kolmogorov-Smirnov Test across the entire distribution for each category. The results indicate statistical significance at the 95% confidence level, with a p-value below 0.01, not only within categories but also across the various samples considered.

ture of the vaccine topic influences their own mode of response and determines a less positively valenced response, although maintaining an overall positive emotional tone. Interestingly, for un-verified users the emotional response is not only less positive, but particularly so in the first months of the vaccination campaign, when the level of contentiousness was particularly high.

In the second panel (B), we then find the emotional responses of the verified and unverified users as distinguished between bots and humans. Discerning bots from humans is usually carried out through the observation of the respective behaviors on social media platforms. More specifically, automated accounts are usually characterized by some

forms of unusual behavior: very large volume of content created, or very high frequency of creation, for instance (see Method section for further details). In our dataset, we found 22,116 verified and 1,318,935 unverified automated accounts. Interestingly, the most significant distinction in terms of emotional expression in our sample is not that between bots and humans, but rather that between verified and un-verified users within each category. It turns out that when we compare the entire distribution of emotional responses between any pair of these four categories of users, each couple of distributions is significantly different in all cases. In particular, the C panel shows the emotional distribution across time respectively for verified and unverified users, while the D panel shows the same analysis discerning between bots and human users to better visualize the statistical distribution of each category.

When an user is verified, be it a human or bot, their emotional response is more positively valenced over time. Likewise, responses are less positively valenced for un-verified users, be them humans or bots. Comparing our results with the baseline sample, we notice an important difference. Whereas verified humans tend to present a more positive response in the baseline, for verified bots the response is more positive in the vaccine sample. The differences between the four categories of users tend to stabilize as the vaccination campaign unfolds, whereas in the very first months the responses are less predictable (see further details in the SI Appendix).

3.3.4 Emotional Responses based on the political orientation

Another important aspect to consider is the role of news media sources according to their political orientation. In Figure 3.4, we analyzed the political orientation of the sources in relation to the verified/un-verified status of the users and their emotional response. As always we have normalized the valence of the tweets across the emotional range (-1, +1) for each of the categories mentioned. In particular, each URL appearing in the messages has been manually inspected by experts (see Method section), classifying the various sources in one of the following categories: left, left-center, center, right-center, right.

We quantified the distribution of the political orientation of the sources and the status of the users both for the baseline and the vaccination samples. We obtain an interesting result: the variation between the two samples is apparent, and the difference is statistically significant among all the political orientations, whether the users are verified or not. The first panel of the Figure indicates the statistics for tweets from the baseline sample, showing that a higher number of tweets has a left-center orientation

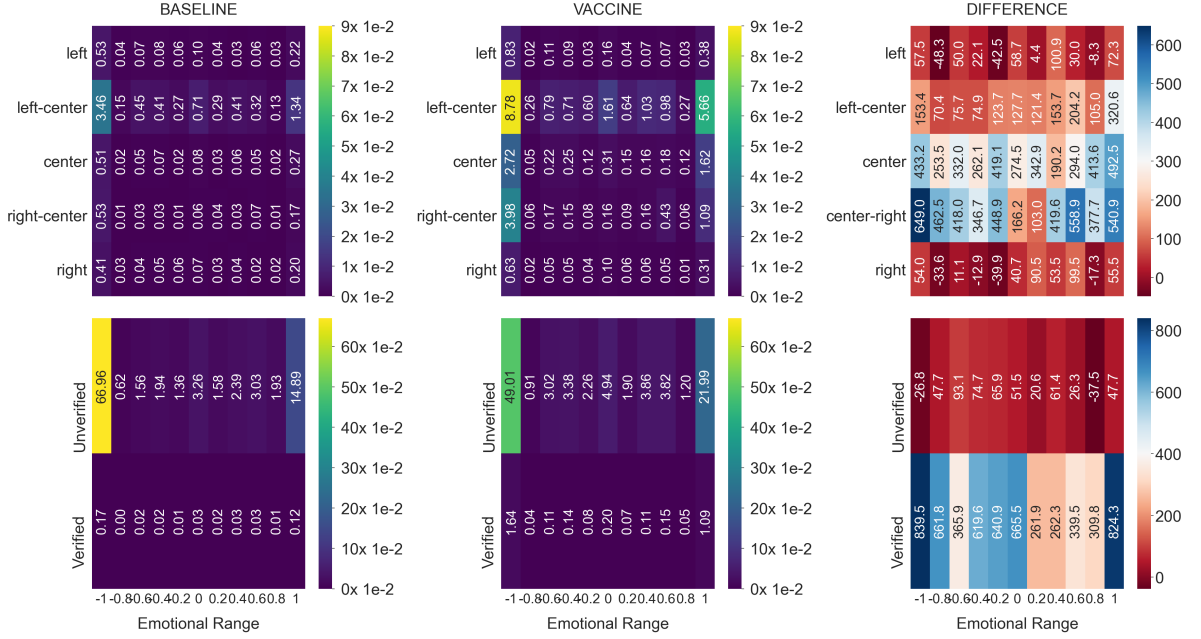


Figure 3.4: **Emotional range for messages posted on Twitter by verified-unverified accounts and with URLs associated to five different political leaning.** For each panel, we quantified statistics of tweets across the emotional range (-1, +1) according to the political leaning of the sources and the verified/un-verified status of the users. The left panel reports the 10M random sample of tweets as a baseline, while the central one is about the vaccine tweets. The right panel shows the percentage difference between vaccine and baseline data for each cell. We show how emotional responses can become a sort of marker for different ideological position, especially for the left-center, center and center-right categories. An interesting result is shown by the verified accounts through which the emotional responses vary at different scales, reaching high negative and positive values.

compared to the others. However, in the vaccine sample the incidence of political orientations changes, and also the emotional responses are pretty different, especially for the left-center, center and center-right categories. The difference is substantial in the third panel where we can observe significant variations in emotional responses especially for the left-center, center and center-right categories. This clearly shows how emotional responses become markers of a differentiation between different ideological positions in the case of a highly contentious topic such as vaccination. Interestingly, the positions where we observe the most heterogeneity in response are the relatively moderate ones, and not the most extreme, as the former are the ones that need to differentiate more from their relatively closer analogs. Extreme positions are already well differentiated and their emotional response patterns are less characteristic.

In the bottom panel, where we further differentiate between verified and un-verified

users, we find that again the emotional responses are different between the two categories. For unverified users, there is scarce differentiation in terms of dominant emotions. However, in the case of verified users, the emotional response changes significantly, with the emotional valence that tends to radicalize toward highly positive or highly negative levels, respectively (see Figure 3.4), similarly to what we observed in 3.3.

3.4 Discussion

Are online ‘emotions’ an important source of social cognition, which is modulated by the status and by other specific characteristics of the subjects? Our analysis shows that this is the case, especially when the topic of the discussion is contentious. As we have seen, a contentious topic such as vaccination elicits stronger and more nuanced emotional responses than baseline conversations. Moreover, users who are characterized by a specific status, that of verified users, have an emotional response whose valence is more positive than that of unverified users. The higher social status related to the verification, and the consequent incentive to reputation management and higher accountability, implies that such users feel more pressured toward providing constructive, inspiring emotional responses rather than dismissive and confrontational ones. Interestingly, this is the case whether the users are humans or bots. However, given the recent changes in the conditions of access to a verified status on Twitter, it is possible that such social pressure effects on emotional responses would not be found under the new regime, as verified accounts now do not signal social status any longer but only mark the purchase of a specific service.

On the other hand, also political orientations of the sources cited have an important implication for emotional response, and in the case of contentious topics there is a general tendency to differentiate the emotional response so that to better differentiate political identity accordingly. However, unlike what could be expected, it is especially in the case of the differentiation between relatively moderate positions that emotional responses are used as a differentiating factor, whereas in the case of more extreme positions this effect is less marked, as the positioning at an extreme side of the political spectrum already suffices to ensure differentiation. This confirms that emotional responses may be strategically inflated in online interactions to construct a specific, and optimally differentiated, social identity. [129]. This may be especially the case for contentious topics, where emotional arousal is pursued by the parties involved also to enhance the salience of intentional signals of false information that function

as demonstrations of commitment to the group’s cause, aiding in the strengthening of group solidarity against opponents in polarizing discussions [130].

This implies, in particular, that users with different political orientations might develop their own specific ‘emotion playbook’ which is characteristic of their political position and that might allow very complex forms of social coordination and even synchronization with their base through a suitable, skillful use of emotional signals. A political base that is particularly sensitive to fear or disgust could therefore be activated by any kind of content where such emotions are balanced with others in a certain, characteristic proportion. It is possible that this kind of strategy has been already experimented by populist leaders worldwide, where the emphasis on the emotional component often overrides that of the specific content [131]. This is certainly a topic that deserves further analysis and that could prove of great importance in understanding future social responses to major contentious topics and events.

Chapter 4

Digital Monitoring of a Rising Global Disease: A Twitter-Based Study of Long Covid Conversations

Building upon the work presented in 3, this present chapter moves from the study of online conversations around vaccinations against Covid-19 to another major health related conversation. This work is part of an article under submission at *Journal of Medical Internet Research* by Anna Bertani, Pierluigi Sacco and Manlio De Domenico with the title Digital Monitoring of a Rising Global Disease: A Twitter-Based Study of Long Covid Conversations. The main contributions of this work are:

- we propose a methodology integrating social network analysis, news reliability, political leaning, and emotional valence to study Long COVID discussions.
- we show that online conversations are dominated by reliable, science-based information with generally positive emotional valence and limited political polarization.
- we reconstruct information cascades and community structures, revealing how attention spreads quickly and is driven by key actors.
- we provide initial evidence that social media activity may reflect real-world disease prevalence, suggesting a tool for digital health monitoring.

4.1 Introduction

Long COVID, also known as Post-Acute Sequelae of SARS-CoV-2 infection (PASC), is an emerging condition characterized by persistent symptoms and health complications that extend beyond the initial phase of COVID-19 infection [132]. As the global pandemic continues to unfold, understanding the long-term effects of COVID-19 has become a critical area of research [133, 134, 135]. Defining Long COVID has been challenging due to the wide range of symptoms and the lack of a universally accepted case definition [136]. The World Health Organization (WHO) has proposed a clinical case definition of the Post COVID-19 condition, which includes symptoms that last for at least two months and cannot be explained by an alternative diagnosis [137]. These symptoms may include fatigue, cognitive dysfunction (or "brain fog"), respiratory difficulties, and a variety of other systemic manifestations [138].

The prevalence of Long COVID is still being investigated, but current estimates suggest that a significant proportion of individuals who have recovered from COVID-19 may experience prolonged symptoms [139]. An early systematic review [140] found that the median proportion of individuals experiencing at least one persistent symptom was 54% at one month, 55% at two to five months, and 54% at six or more months. These findings highlight how Long COVID may cause a substantial burden on both individuals and healthcare systems. Research on the pathophysiology of Long COVID is ongoing, with several hypotheses being explored. [141] consider a number of potential, concurrent pathogenetic mechanisms at work, including injury to one or multiple organs; persistent reservoirs of SARS-CoV-2 in certain tissues; re-activation of neurotrophic pathogens such as, e.g., herpesviruses under conditions of COVID-related immune dysregulation; SARS-CoV-2 interactions with host microbiome/virome communities; clotting and coagulation issues; dysfunctional brainstem/vagus nerve signaling; ongoing activity of primed immune cells, and autoimmunity due to molecular mimicry between pathogen and host proteins. In addition, it is conjectured that certain human genome variants and/or human endogenous retrovirus (HERV) activity might predispose individuals to developing Long COVID. In the past few years, a considerable amount of research has been devoted to developing a deeper understanding of Long COVID in relation to, among others, early predictors [142], COVID-19 vaccination [143], inflammosome [144], and incomplete viral clearance [145]. However, we are still very far from a satisfactory understanding of the disease, which, to our current state of knowledge, manifests in over 200 symptoms affecting multiple organ systems, including cardiovascular, neurological, respiratory, and gastrointestinal systems. Currently, at least 10% of current

SARS-CoV-2 infections lead to Long COVID conditions. As of 2023, at least 65 million patients are estimated to have the disease, with cases steadily increasing. Moreover, current diagnostic tools are insufficient and there is an urgent need for robust clinical trials to test potential treatments [133]. Therefore, diagnosis of Long COVID remains primarily based on clinical presentation and the exclusion of other potential causes, and management of Long COVID is still currently focused on symptom relief and supportive care [146, 147]. Long COVID then represents a significant challenge for both patients and healthcare professionals, and a major public health issue [148].

On the other hand, Long COVID shares many sociological features with other so-called contested illnesses [149, 150], such as chronic fatigue syndrome, chronic Lyme disease, or mast cell activation syndrome. With the term contested illness, we refer to those chronic conditions with limited symptoms and controversial physical signs. These illnesses are considered contested because conflicting views persist among patients and doctors about their nature, cures and treatment [151]. These conditions emerge at the intersection of medical ambiguity and patient activism, where individuals must “fight to get” their illness to be recognized [149]. They are characterized by online communities that mobilize lay expertise, personal experience, and collective identity to advocate for recognition, care, and research. In these digital spaces, patients, journalists, and scientists collaboratively produce and contest knowledge, making the online sphere a crucial arena of epistemic negotiation [152].

From this perspective, Long COVID is not only a biomedical challenge but also a socio-political one, reflecting tensions between institutional medicine and real experience, between public health authority and digital grassroots knowledge. Understanding Long COVID thus requires an interdisciplinary approach that integrates the study of online digital communication as a solid empirical ground for examining how social dynamics, information reliability, and affective engagement evolve in real time.

Building on this theoretical framework, we investigate whether social media activity can serve as a tool for monitoring not only epidemiological signals but also the evolving social construction of disease.

For this reason, developing advanced methods for the tracking of the diffusion of the disease [153] and for monitoring how specific policy and professional communities as well as the public opinion as a whole understand the disease as a clear public health threat is of fundamental importance [154, 155]. Whereas telemedicine and e-health hold great promise as transformative approaches in healthcare [156], also in view of the lessons learnt from the global pandemic [157], accessibility to large-scale databases populated by information at regional and sub-regional level, allowing to perform ecological studies

on key population health issues is not always easy, and especially so for emergent and poorly diagnosed diseases such as Long COVID.

In this study, we ask whether available digital resources such as online social networks can be used to monitor, identify and characterize in nearly real-time the emergence of a novel disease characterized by a global (i.e., non-localized) incidence and concerning prevalence [158, 159] such as Long COVID.

Specifically, our research examines how digital conversations reflect (a) the reliability and emotional valence of information shared about Long COVID, (b) the role of verified and authoritative users in disseminating scientific knowledge, and (c) the relationship between online information dynamics and epidemiological indicators of disease prevalence.

We work on a database extracted from a popular micro-blogging platform (Twitter, now $\mathbb{X}$) related to Long COVID, across almost three years. We have enriched such data with measures for the reliability of the shared content and natural language analysis, to study the behavioral response of the online community over time and analyze the digital fingerprints of this rising global disease. We find relevant differences between online users conversations about Long COVID and a baseline of conversations about non-contentious topics. Collective attention towards Long COVID is increasingly high, with peculiar information cascades involving distinct types of actors. Importantly, we find a predominance of mainstream media and scientific coverage of Long COVID in shared content, although other types of news, such as conspiracy and junk science, also play a role in the online conversation.

This means that, for a poorly understood condition with lack of valid clinical approaches such as Long COVID, online media can be an important source of scientifically sound information. In particular, verified users (i.e. highly engaged, reputable professionals as defined by the criteria for verification on Twitter before March 2023) turn out to be reliable sources of information in this regard. Moreover, the emotional content of such online conversations is generally positively valenced (with a main emphasis on anticipation), and weakly correlated to political leaning. We consequently use the volume of messages categorized as scientifically sound to check whether it can be used as a proxy for the prevalence of Long COVID at a global level. Fusing our data with the estimated prevalence in 2020 and 2021 for several countries worldwide, we find a weak correlation that is however significant ($p\text{-value} < 0.05$). These findings therefore suggest a promising line of research on online social networks as a potential global laboratory for the identification and monitoring of emergent global diseases – a new and possibly transformational tool for decision- and policy-making for public health.

The unique value of this study is that, to our knowledge, this is the first large-scale socio-psychological tracking in real time of the emergence of an illness with global implications from online conversations. In this perspective, and even more so in view of the prevalence of scientifically reliable sources in such conversations, we can think of tracking of collective online conversations as a complementary resource with respect to prediction polling in the context of a collective intelligence approach to public health surveillance and monitoring [160, 161], also in view of the fact that viral agents themselves seem to possess sophisticated forms of collective intelligence that need to be countervailed [162].

4.2 Methods

4.2.1 Overview of the data

We collected data through Twitter API by retrieving all the messages that contained at least one relevant keyword related to the Long COVID Discussion (see Appendix B for further details), consisting in total of 12 million of tweets and 1.7 million of unique users. Our dataset covers two years and a half of online conversations about one of the most discussed diseases globally. We started collecting data from the early stage of the Covid-19 pandemic (February 2020) until the end of November 2022. We compare our case study with another sample of randomly collected posts dealing with non contentious topics, consisting of 10 million tweets taken between April 21, 2021 and April 24, 2021 from 6 millions of unique users. We only considered geo-tagged tweets, to be able to accurately monitor users' locations, as reported by their device.

4.2.2 News Reliability and Political Categories

As we built our database, we also collected several metadata about type of users, type of news shared and the political leaning of the shared content, extracted from the url attached to each tweet. We used the same methodology presented in 3 with respect to the 5-item spectrum of political leaning (right, right-center, center, center-left, left) and to a 7-item typology of media sources (political, satire, mainstream media, science, conspiracy/junk science, clickbait, fake/hoax, as manually inspected by experts). In particular, building on [2], we have classified our news sources as *reliable* (when belonging to the Science, Mainstream Media categories) and *unreliable* (when belonging to the Satire, Clickbait, Political, Fake or Hoax, Conspiracy and Junk science categories).

Finally, as a third option we consider tweets not containing any url to be classified (i.e. mere opinions without a source).

4.2.3 Verification Status

In this study, we refer to the verification status definition used in Section 3.2.2.

4.2.4 Emotional Detection and Aggregation

The NLP pipeline used to analyze the tweets allows processing of text and emotions in multiple languages. In this case, we employed the same methodology used in chapter 3.

4.2.5 Dataset about the prevalence of Long COVID

In addition to the above described data, we also analyzed a dataset containing information about the prevalence and incidence of Long COVID for each country in 2020 and 2021. The data estimates by country the number of patients affected by long COVID in 2020 and 2021, regardless of the severity of their symptoms, and is obtained from [163]. The data is based on information from 1,906 community infections and 10,526 hospitalized patients from ten collaborating cohorts, three of which included children, as well as ICD-coded medical record data concerning 1.3 million infections. Such evidence has allowed us to explore how the prevalence of Long COVID has affected the online discourse by analyzing different message features, such as political leaning and the type of news media content linked to each tweet.

4.2.6 Community and Information Cascades

The Twitter-based global communication network hosting Long COVID conversations has been analyzed by re-constructing the pairwise interactions between different users. Pairwise interactions (retweets, replies and mentions) are used to build an un-directed weighted communication network where nodes are the user's accounts and weighted links represent the number of active (who endorses whom, who replies to whom) or passive (who is endorsed by whom, who receives a reply by whom; etc) interactions between a couple of users in the time frame considered. We decided to exclusively keep links with at least 10 or more interactions in order to consider only users who were particularly active on the topic, whereas occasional interactions have been discarded. We

used one of the most well known algorithms for the detection of communities in online complex networks, the *Louvain Method* [164], a widely adopted modularity-optimization algorithm that efficiently identifies cohesive groups of nodes in large networks. This approach was chosen over alternatives, such as Girvan-Newman, due to its scalability and ability to capture robust multilevel community structure. This approach allowed for a first, computationally efficient exploration of the network’s modular structure and the identification of prominent communities of users actively engaged in Long COVID conversations. In contrast, the stochastic block model (SBM), explained in the next section, was applied to further characterize the network at multiple hierarchical levels, providing a generative perspective that accounts for the probability of connections between communities and allows for degree correction. The SBM analysis thus complements the Louvain partitioning by revealing nested community structures and offering a more nuanced understanding of the relationships between groups of users. However, both methods operate on the aggregated, static network and do not explicitly account for the temporal evolution of interactions. Given that Twitter activity unfolds over time, temporal community detection methods could in principle capture the dynamic formation, merging, or dissolution of communities. In the present work, the use of static methods was motivated by the need to obtain a tractable representation of the overall structural patterns of communication and information diffusion, while acknowledging that some temporal dynamics may not be fully captured.

We also reconstructed the information diffusion process by tracking so-called information cascades. In network theory, information flows in the form of cascades along the links of the network. The first node is the one sparking the information cascade: the first user who posts a message. The links reflect the actions undertaken by other users as a consequence, such as retweeting, replying to, or mentioning the original message (node). Users could as well re-share the content of another node, not necessarily the originating one, etc. The process unfolds in time, through a series of cascading informational waves whose structure depends on the interaction mechanism. However, it has to be noted that in the case of Twitter the reconstruction of cascades is inherently limited by the data made publicly available by the platform. In particular, intermediate steps in the diffusion process are often not traceable: when a user retweets a message, the platform records only the connection to the original tweet, even if the user saw it through another retweet by someone they follow. As a result, the reconstructed cascades represent a simplified approximation of the true underlying diffusion paths and may not capture all intermediate interactions that contributed to the spread of information. Tracking information cascades allows us to move beyond static network connections and quantify

how content spreads and which users play key roles in amplifying information, following foundational approaches in social network research [165, 40].

4.2.7 Multilevel Communities

In this study, we also employed a stochastic block model (SBM) using the *graph-tool* library to analyze the community structure within the network at multiple hierarchical levels [166, 167]. The stochastic block model is a generative model for random graphs that partitions nodes into communities or blocks, with the probability of an edge existing between any two nodes depending solely on their block memberships. To enhance the accuracy of our community detection, we applied degree correction to the model, allowing it to account for the varying degrees of nodes within the same community. This method enabled us to identify and characterize the different communities and their multiscale relationships within the network. Furthermore, we enriched our analysis by creating two distinct indices to characterize the nodes: a political index and a reliability index. The political index was designed to capture the political orientation of each user, based on the characterization of news media sources shared by each user, normalized between -1 (far right) and +1 (far-left). The reliability index, on the other hand, was based on the proportion of reliable versus unreliable content posted by each user, thereby offering a measure of the trustworthiness of the information disseminated within the communities, ranging from -1 (completely untrustworthy) to +1 (completely trustworthy user).

4.3 Results

In the following, we analyze the social bulk [122] (see Methods) of the online activity unfolding on the popular microblogging platform Twitter (now X). The social bulk is a representation of the online activity, which is built on the number of interactions (such as retweets, replies, mentions) among users, where nodes are the users and the weight is given by the number of interactions. The social bulk represents the core of meaningful social interactions within the platform, a filtered, weighted network that captures only the most relevant and sustained exchanges among users. Throughout this study, we will focus on the analysis of the structure of the social bulk, while enriching the network with information extracted from the content of users' posts. Specifically, we focus on the URLs they shared, to infer the reliability and the political leaning of the corresponding news, and on the content of each post, to infer the emotional status

by means of machine learning techniques for natural language processing. We refer to Methods for details about these quantities.

From a structural perspective, we unravel a complex mesoscale organization of the social bulk across multiple topological scales (see Methods). Figure 4.1 shows the evolution of online communities discussing about Long COVID for different levels, as captured by an inferential procedure based on degree-corrected stochastic block model (see Methods). The presence of multiple levels reveals a non-trivial hierarchical modular organization of online social interactions, which is typical of communication systems that are highly efficient in exchanging communication [168]. A more detailed analysis, based on the enrichment of this network with metadata, is reported later (see Community Analysis).

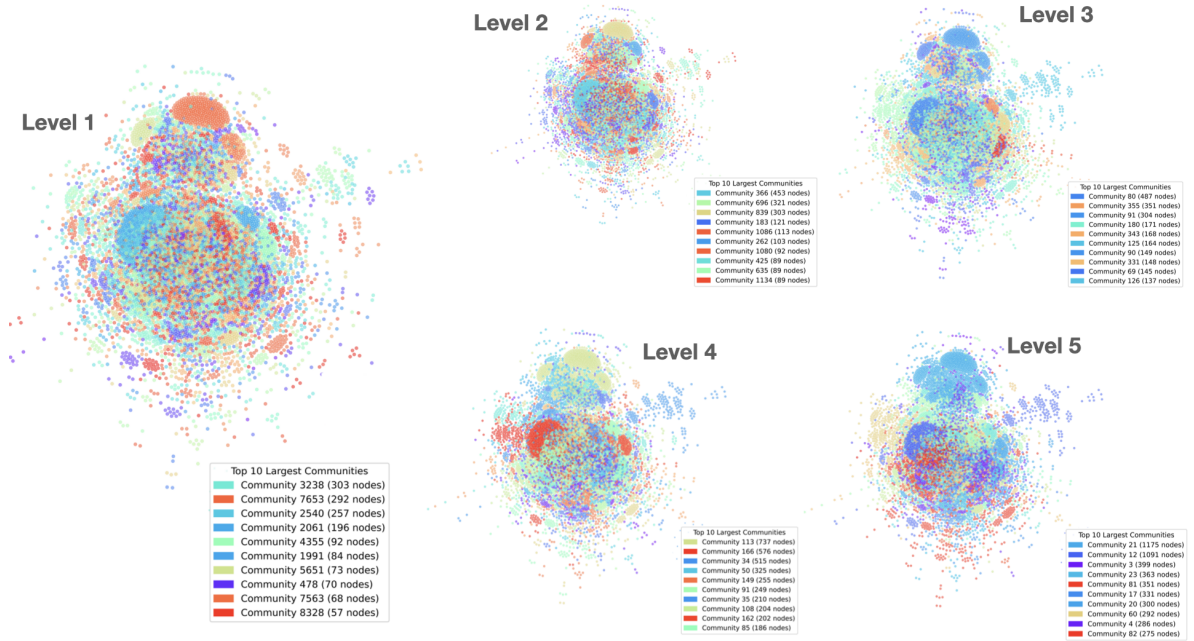


Figure 4.1: **Visualization of the multilevel communities.** We show the network visualization for the first five levels identified by a multilevel community detection analysis (see Methods for details). Each node is coloured according to the assigned community.

Figure 4.2 shows the reliability and political index of each user according to the proportion of trustworthy news shared, demonstrating that the majority of accounts share reliable content and has a political leaning classified as left wing. However, in both examples, at a finer granularity we have cases of users with either a lower reliability score or a political leaning more aligned with the right side of the political spectrum.

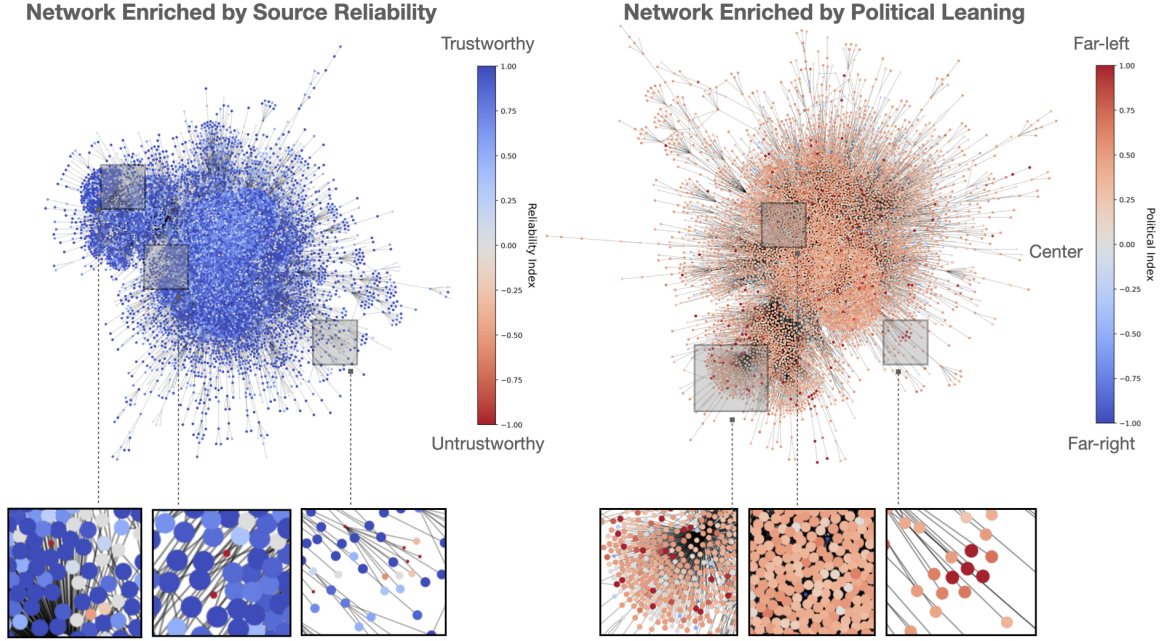


Figure 4.2: **Reliability and Political Index.** Left panel shows the network visualization enriched by the reliability index ranging from -1 (completely untrustworthy) to +1 (completely trustworthy). Right panel shows the network enriched by the political index ranging between -1 (far-right), 0 (least biased leaning) and +1 (far-left).

4.3.1 Emotional Analysis

We consider a dataset consisting of more than 10 million online messages posted on Twitter with a non contentious topical focus. We also consider the online conversations about the Long COVID disease occurring from the beginning of the pandemic (February 2020) until November 2022, consisting of 12 million of online messages and 1.7 million of unique users. We found that the Long COVID dataset substantially differs from the baseline sample in terms of emotional responses. In particular, the Long COVID sample shows an over-representation of one particular emotion with respect to the others and, even more importantly, with respect to the baseline sample. Anticipation is the most common emotion in the Long COVID dataset, reaching a value of 0.18, whereas in the benchmark sample the corresponding value is 0.125 (see Appendix B). The prevalence of this particular emotion highlights how the Long COVID discussions have been characterized by a significant emotional arousal of a specific kind, which signals the relevance acquired by the topic in the online conversational space as well as a typical and somewhat unexpected emotional imprint.

Moving from our results on the distribution of each emotion, we now want to in-

investigate how online messages are characterized in terms of emotional responses with respect to different features. In particular, we are interested in the emotional reactions elicited by the political leaning of the shared content. With respect to this feature 4.1, we find an interesting result: online messages linking news classified as right-center leaning are characterized by a more positive valence compared to left-center leaning ones. In addition, such difference is statistically significant even when the distribution is compared with the baseline sample, where right and right-center posts feature a more pronounced negative valence.

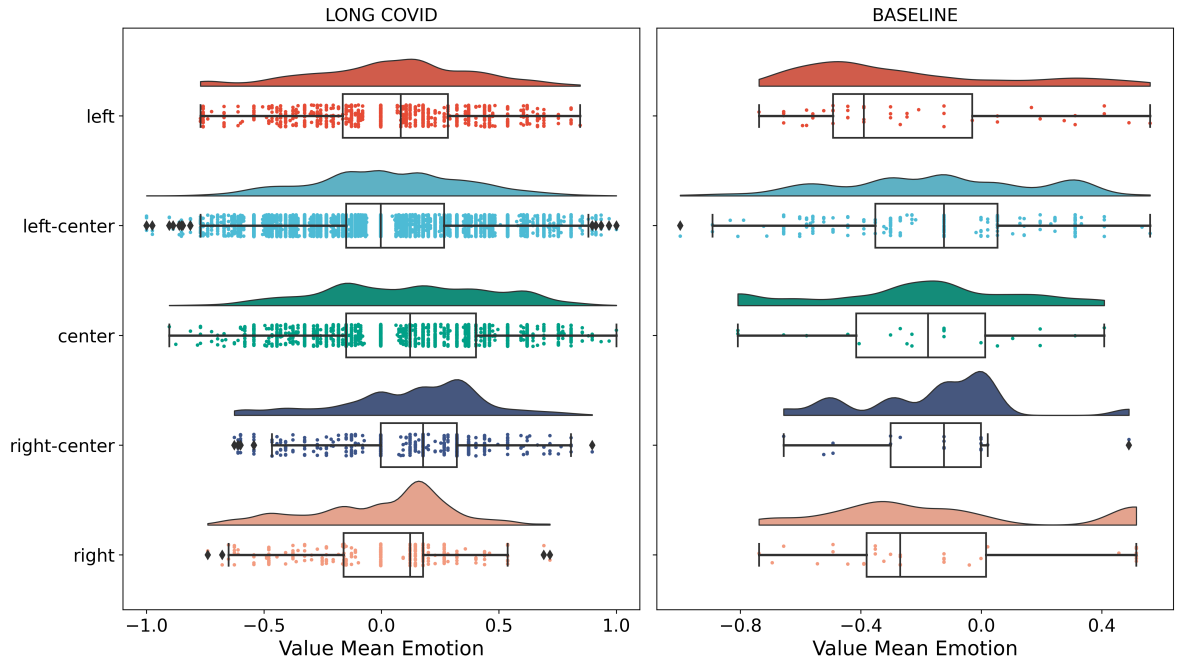


Figure 4.3: **Emotion Distribution for different types of political leaning.** For each sample (Long COVID vs baseline), we quantified the distribution of emotional values from -1 to 1, respectively for tweets with left, left-center, center, right-center and right political leaning.

In addition to political leaning, we also consider whether a verified user status and the news media category of the shared content have any significant effect on emotional reactions.

Figure 4.3 shows the percentage difference between the Long COVID dataset and the baseline sample with respect to the three features under analysis: news media type, verification status and political leaning. More specifically, we quantified the distribution of categorized tweets in the -1 to +1 emotional range. In addition, we computed the T-test on the difference between the Long COVID and baseline sample distributions, to test the statistical significance of our results.

With respect to the news media categories, we found that Science features a higher percentage difference for the entire emotional range, and therefore scientific content prevails in the Long COVID conversations with respect to all other kinds of news. As to verification status, it turns out that verified accounts are associated to a higher percentage difference between the Long COVID and the baseline sample. Finally, concerning political leaning, the percentage difference between Long COVID and Baseline samples highlights huge differences that reach values up to 3000, especially for the center and center-right categories. Long COVID-related conversations seem therefore to be characterized by a stronger focus on reliable information, more attention toward reliable sources, and a more positive emotional valence with respect to the non-contentious topical baseline.

4.3.2 News Media Characterization

Online Long COVID-related discussions have been analyzed with a focus on the news media categories linked to the content shared by the users, also as compared to the baseline sample (more information in the Appendix B). Although the baseline sample is limited to only 4 days, it gives us a proxy of the overall trend, providing a valuable benchmark to compare Long COVID data. The analysis shows the percentage distribution of news shared according to news categories over time (panel A), whereas panel B reports the variation in the distribution of news categories for the Long COVID vs. baseline samples. Finally, Panel C reports the cumulative deviation from the baseline, showing the extent of the over-representation of the Science category with respect to the baseline. More generally, analyzing all news categories as distinguished between reliable (Mainstream Media, Science) and unreliable sources (Fake/Hoax, Conspiracy/Junk Science, Political, Clickbait, Satire), we found that the incidence of reliable sources in the Long COVID sample is much higher with respect to the baseline, where unreliable sources are more prevalent (see more details in the Appendix B).

In addition, we also evaluated political leaning compared to the baseline, finding that center leaning is the most prevalent one in the Long COVID dataset. Interestingly, both right and right-center leanings present spikes in specific moments of the conversation, especially in the first months of the observation period (as shown in the Appendix B).

4.3.3 Community Analysis

The communication network of Long COVID conversations has also been enriched with the information obtained from the analysis of the content shared during the obser-

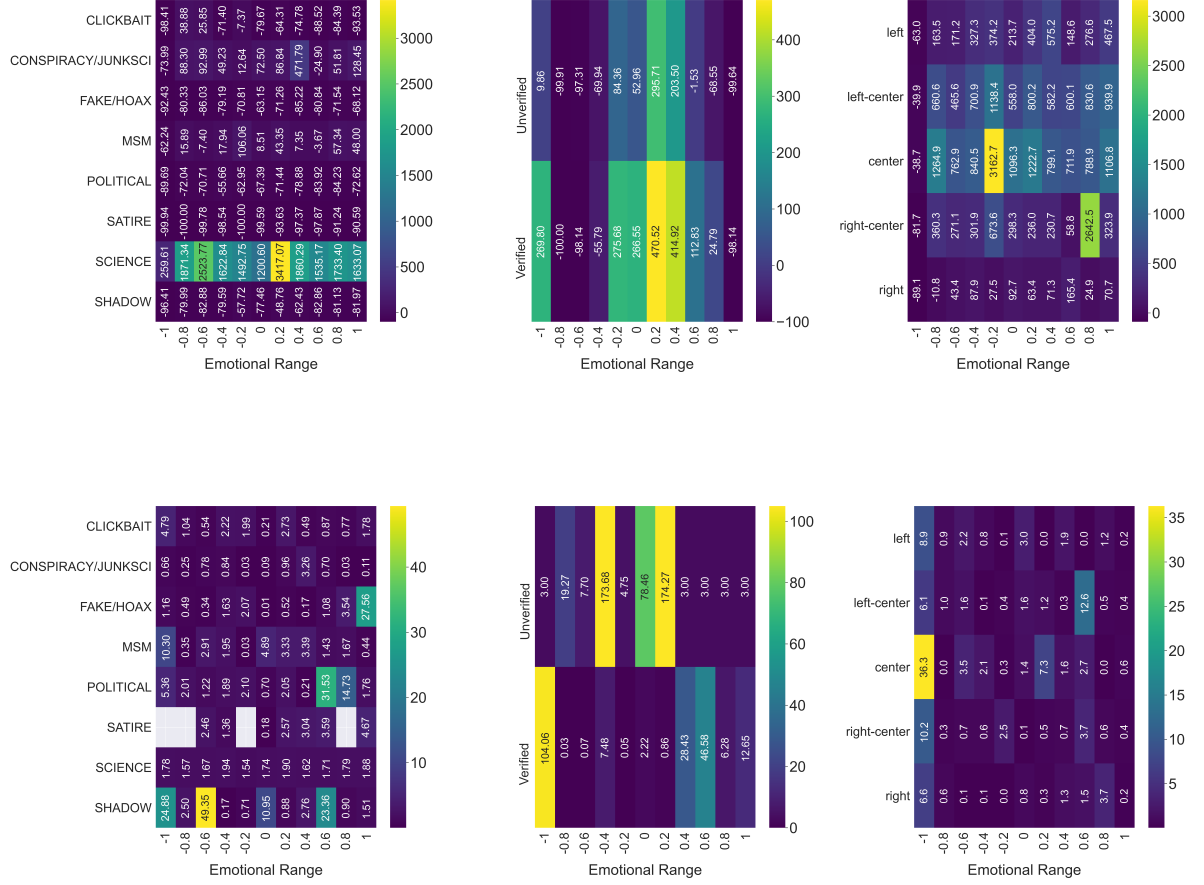


Figure 4.4: **Emotional range for messages posted on Twitter by verified-unverified accounts and with URLs linking to five different political leaning orientations and eight news media categories.** The top panel shows the percentage difference between the baseline and Long COVID data for all the categories considered: news media, verified-unverified accounts, and political leaning. The bottom panel shows the p-value results (in logarithmic scale) from performing the T-Test between the two distributions for each cell.

vation period. Each URL linked to the messages has been labeled to the political leaning of the media source (left, left-center, center, right-center and right) and the news media type of the content (science, mainstream media, satire, political, fake/hoax, clickbait, conspiracy/junksci). For each group considered, we quantified the distribution of media sources by political leaning and type, showing the results in the Appendix B. Most of the communities are characterized by different typologies of news sources and political orientations, and in all the communities one or more forms

of political leaning prevail in the total composition with respect to the least biased political orientation.

Following the analysis in [2], unreliable news may be further differentiated according to their relative levels of lack of accountability, in ascending order: Satire, Clickbait, Political, Fake or Hoax, Conspiracy/ Junksci. Less reliable sources, in principle, pose greater informational threats, and especially so in view of the fact that mention of threat-related informational content, however unreliable, could be taken as a signal of competence by some users [169]. In particular, communities 111 and 142 present a higher prevalence of conspiracy and junk science news compared to the other communities. In addition, 25 % of news circulating in the community 142 are right-center leaning. Another interesting community to mention is surely the one which, to the contrary, features 75 % of news coming from scientific sources (see Appendix B). It is also interesting to remark how, despite that most reliable news are typically classified as mainstream media sources, in all communities we can find some share of scientific news. As shown [114], online conversations are often driven by a small number of active users who make a big impact in terms of group polarization processes and of prevailing informational biases. Social activities and information shared are highly centralized in online social systems. We quantified how many users are responsible for how many news media shared, social activities, and politicized content during the unfolding of the Long COVID conversation. The results of our analysis highlight an evident centralization in all the spheres considered: social activities, news but also political centralization, quantified by means of an inequality index (the Gini Coefficient) consistently with the results in [114], as shown in the Appendix B.

Informational Cascades

Besides communities, we want to investigate how communication evolves by taking into account information cascades [170] as a process that originates at some point, to subsequently unfold and eventually end. More specifically, our goal is to identify those actors who originate information cascades at different levels. In this context, we quantified the degree (K) distribution of users starting a conversation on Twitter, corresponding to the user's number of followers. The degree of each node is a fundamental attribute since user exposure can be also evaluated by the number of followers. Figure 4.5 depicts the degree (k) of the first node considering 3 possible scenarios: 1) all the cascades; cascades that 2) involve at least 10 or more users, or 3) 50 or more users. As the number of the users involved in the cascades increase, the degree of the first node be-

comes higher, looking at the median values of the first node degree in Long COVID discussions. From plot (A), the peak of the distribution of the degree is at about 1000 followers. Filtering the data further, we find that the distribution of the degree shows meaningful differences between our case study and the random sample. In particular, plot (C) shows how the distribution of the first node's degree presents higher values compared to the baseline sample. The bottom panel shows the distribution of the degree without considering the first node, so that we can analyze also the average degree of users involved in this type of conversation. In particular, the differences between the two samples are more marked as the cascade size increases.

In addition, another important attribute to consider is the time necessary to reach collective attention in online social media, in view of the relationship between the dynamics of collective attention and the perceived uncertainty about a topic on social media [171]. The speed in reaching collective attention gives an idea of how long the information cascade takes to develop and involve as many users as possible, that is, its maximum breadth. One may think that there are differences in this respect between the Long COVID and baseline samples, provided that Long COVID discussions might be led by experts or influencers who can command attention from other users faster than in discussions about non-contentious topics, at least in the early phase of the debate [172]. Figure 4.5 (panel G) shows the distribution of time needed to reach the maximum numbers of users involved in the cascade. More precisely, we quantify the moment when the highest number of people were involved in retweeting content from the cascade, and then we compute how long (in minutes) it takes to reach collective attention. The Long COVID discussion reaches collective attention more quickly, taking only 10 minutes (median value) to get maximum attention.

Digital Health

A very important point to consider is how online discussions about Long COVID are related with its actual prevalence in the population. To this purpose, we used a dataset on the prevalence of Long COVID in 2020 and 2021, and we compared the prevalence of the disease with the number of related scientific sources shared on Twitter in Long COVID conversations. In Figure 4.6, we quantified the logarithmic amount of news classified as Science in each country against the prevalence of Long COVID in 2020, whereas the right panel reports data for 2021. We observe a positive relation for each of the years considered, reaching a Spearman correlation of 0.33 in the first year, that decreases to 0.24 in the second year. The two regressions are both statistically significant

4.3. Results

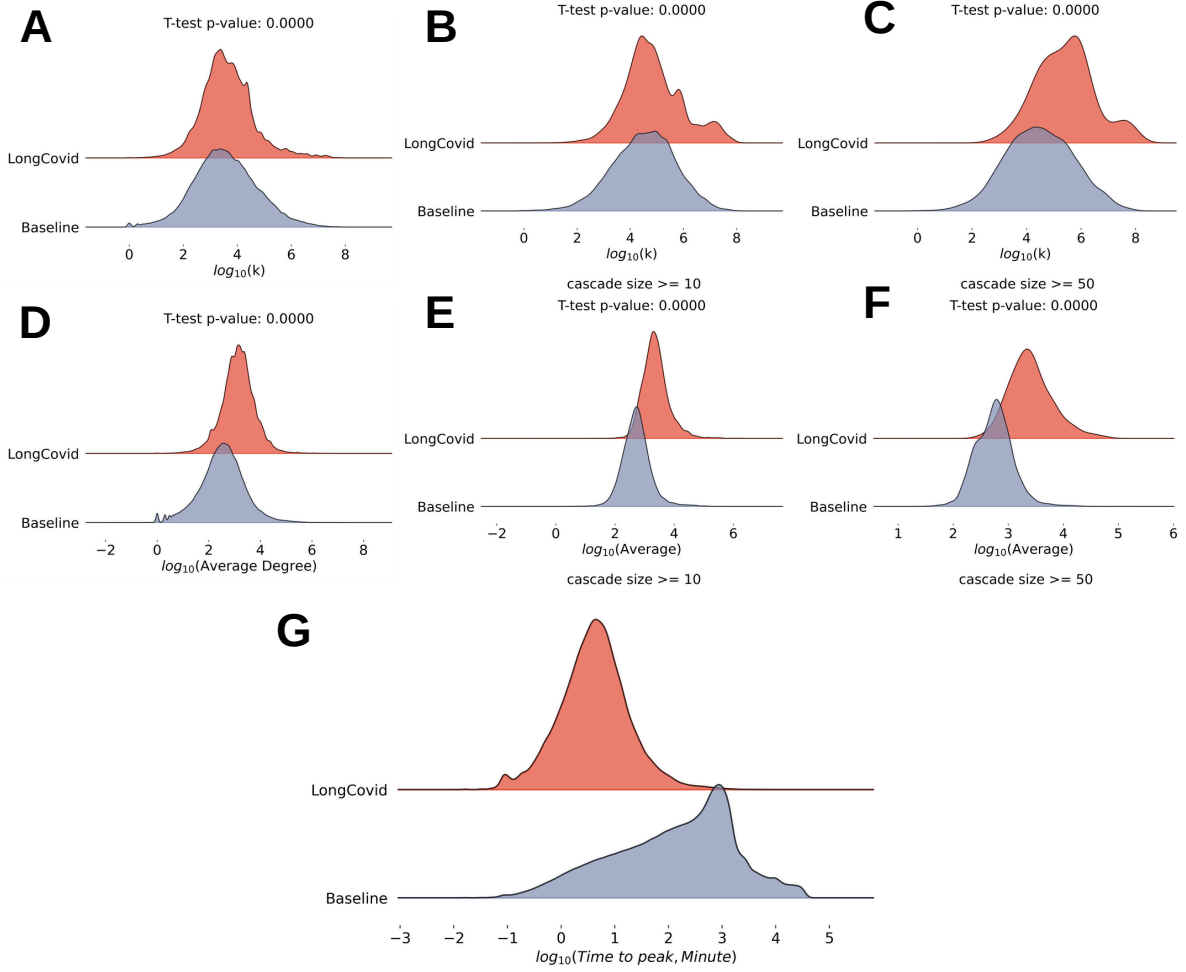


Figure 4.5: Degree Distribution and Collective Attention in Information Cascades. For each panel (A-C), we quantified the average degree of the first node for Long COVID and baseline samples, considering respectively 1) all the cascades, cascades that 2) involve 10 or more users, or 3) 50 or more users. For each panel (D-F), we quantified the average degree of the nodes involved in the cascades (but the first node) for the Long COVID and baseline discussions in all three scenarios, as described above. Panel G shows the time needed to reach collective attention for Long COVID and baseline samples.

with a p-value lower than 0.005. While this is only a statistical association, it shows that there could be mechanisms at play to relate the epidemiology of the disease to its awareness across the population [173], a topic that has been investigated from a more theoretical perspective so far [174, 175, 176, 177, 178].

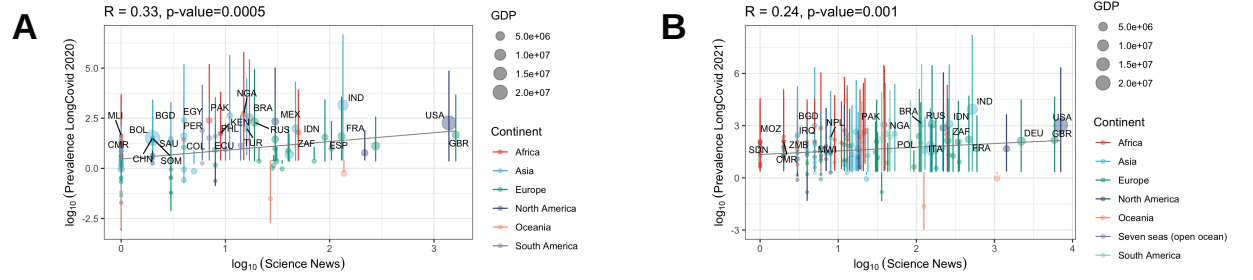


Figure 4.6: **Long COVID Prevalence vs Science News Regression.** Testing the hypothesis that in countries where Long COVID is more prevalent the volume of science-based news is higher. For each panel, we quantified the Spearman Correlation and the linear regression between Prevalence of Long COVID and news classified as Science, for 2020 and 2021 respectively.

4.4 Discussion

In this study, we have analyzed online Twitter conversations on Long COVID, contrasting them against discussions on non-contentious topics, to assess whether the former present significant structural differences from general online conversations. A key finding is that collective attention towards Long COVID has been increasingly high, and characterized by distinct information cascades that involve a variety of actors. Most interestingly, despite the noisy online environment where misinformation and conspiracy theories also circulate, as it typically happens for all topics related to the global pandemic [179, 180], mainstream media and especially scientific coverage predominate in the shared content. We have also employed enriched data, which include measures of content reliability and natural language analysis, revealing meaningful behavioral patterns in how the online community engages with Long COVID information. Verified users, in particular, play a crucial role in disseminating scientifically sound information. Moreover, the communication network analysis highlights significant deviations in the nature of media sources linked in Long COVID discussions compared to a baseline of less contentious discussions. This difference is further evident in the distribution of degrees of nodes in the communication network, with more significant degrees observed in Long COVID discussions, implying more substantial engagement and possibly influence in these networks. Another element of interest, again of particular significance when compared to the wider literature on COVID-related online behavior, is the relatively weak influence of political leaning in driving online conversations. One could therefore advance the hypothesis that, in the case of a poorly understood condition with potentially serious consequences such as Long COVID, the general public is relatively

more inclined to seek scientifically reliable sources to obtain sound information. This is consistent with evidence that increases in perceived health threats lead people to look for more reliable and corroborated information, even if its content is unwelcome or threatening [181, 182]. In view of the relatively high prevalence of Long COVID among the population and of the distinctiveness and persistence of related symptoms [183], this may have sparked a more focused public concern with the specific health threat than the more general topic of COVID-related health threats, reflecting a more general shift across the course of the pandemic from politically charged conversations to a growing interest toward health outcomes and concrete lifestyle changes brought about by the pandemic [184]. However, as believing and spreading misinformation is typically linked to an emotional state of anxiety [185], we would expect that the prevalence of scientifically reliable information in the Long COVID conversations would be associated with a different emotional tone. And in fact, we find that Long COVID online conversations are emotionally charged, but generally positively valenced. There seems to be a mutually reinforcing link, therefore, between the perceived availability of actual coping and mitigation strategies provided by exposure to sound scientific information about Long COVID, and the emergence of positive anticipatory emotional states that improve the daily management of anxiety and other negative feelings in stressful situations [186], motivating further engaged search for reliable information as knowledge gaps are gradually filled [187]. This result provides new hints on the value of social media for the discovery of emergent patterns that could prove pivotal in identifying major shifts in public sentiment and emerging behavioral trends on issues of major public health concern [188].

Our results have several implications of interest for shaping health policies in response to online discussions about Long COVID. First, they suggest that online social networks can be leveraged as a global laboratory for identifying and monitoring emergent diseases and their epidemic patterns in real-time, providing valuable insights for public health officials [189, 190]. This could empower more timely and data-driven decision-making processes in the public health domain where timely response is vital [191]. Integrating such tools in public health surveillance systems might prove of special importance in this respect [192]. Moreover, as we find a prevalence of mainstream media and scientific coverage in the shared content, this suggests the possibility for health policy makers to use such platforms more effectively to disseminate scientifically sound information about Long COVID, counteracting misinformation from conspiracy theories and junk science – a goal of primary importance in the light of the current, concerning trends in social media health information [193, 194].

There are also important behavioral insights implications: understanding the behavioral responses and attention patterns of online users towards Long COVID can help in planning targeted health campaigns and interventions. Moreover, prevailing emotional trends also provide very useful signals on how to optimally design and deliver public health communication campaigns. Tailoring communication strategies that resonate with the online community's concerns, mood, and information consumption habits could enhance the effectiveness of public health messages. Also of great importance is the understanding of how information centralizes around experts, influencers or key community members during health crises, with major implications for policy makers' capacity to understand vital features of the social dissemination of reliable versus unreliable information. Consequently, there is a need for further research into online social network activities as a means to track and understand the progression of diseases like Long COVID. Policies could support the permanent monitoring of these platforms to swiftly detect changes in disease prevalence or public sentiment. Finally, insights from online discussions can inform the allocation of resources. For instance, heightened online attention on Long COVID may indicate an emerging social demand for more clinical research, better healthcare services, and support for those affected by the condition.

The main limitations of our study can be summarized as follows. The study's dataset is derived exclusively from Twitter, which might not represent the entire spectrum of global online conversations about Long COVID. This focus on a single platform may lead to biases or an incomplete picture of the broader online discourse. In particular, Twitter demographics mostly consists of highly educated users, so that the prevalence of scientifically reliable sources found here might not carry over to other platforms with different educational demographics. In addition, the database was built using messages that contained specific relevant keywords. This method might miss out on discussions that do not explicitly use those keywords, thereby potentially overlooking significant portions of the conversation.

Also, while the study makes use of state-of-the-art measures for the reliability of shared content, distinguishing between scientifically sound information and conspiracy/junk science can be at times complex and subjective, and this might affect to some extent the overall assessment and interpretation of the data. Another delicate issue is that, although the study suggests a promising future for online social networks as a tool for monitoring diseases, the disparity in access to and usage of these networks across different regions and demographics might limit ecological and external validity, and such disparity might influence the generalizability of the findings. It is also worth noticing

that our analysis relied upon a simplified classification of complex emotional responses into a few basic categories, which are the same for the whole database independently of geographical location and cultural background. More granular, and possibly culturally sensitive emotional classification models might help us further improve our understanding of the emotional response patterns for key contentious public health issues such as Long COVID.

This research also suggests some promising avenues for future research. By analyzing the behavioral responses and digital fingerprints of online communities, researchers can gain deeper insights into how information about diseases like Long COVID spreads and how different types of content (mainstream media, scientific coverage, conspiracy theories, junk science) influence public discourse. Additionally, the study indicates that verified users on platforms like Twitter can serve as crucial sources of scientifically sound information about poorly understood conditions, suggesting that future research could focus on how to leverage such authoritative voices more systematically and effectively to improve public understanding and management of such diseases.

As a prospective direction for future research, it would be also interesting to compare the Long COVID discussions with those of other diseases. Such a comparative analysis could contribute to the formulation of a concept of “media diseases”. Investigating whether similar patterns of attention, emotional engagement, and network dynamics emerge across multiple health topics could shed light on the mechanisms that drive public discourse, collective perception, and the amplification of health-related information in digital environments.

Chapter 5

Decoding the News Media Diet of Disinformation Spreaders

This section is based on a journal article. Anna Bertani, Valeria Mazzeo, Riccardo Gallotti, Decoding the News Media Diet of Disinformation Spreaders. *Entropy* 26.3 (2024): 270. Moving from the emotional tone of online conversations, this chapter offers a new methodological framework to quantify the variety and regularity of news sources based on entropic measurements. The main contributions of this chapter are the following:

- we introduce a methodological framework using entropy-based measures to quantify the variety and regularity of users' news media consumption
- we analyze a large multilingual and multicultural Twitter dataset to characterize users' news media diets during the first year of the COVID-19 pandemic
- we reveal that high-risk users sharing fake or conspiracy content display a more diverse yet highly regular pattern of news consumption compared to reliable or low-risk users.
- we uncover structural patterns in the news media environment, showing that unreliable content is often concentrated in alternating sequences of fake and conspiracy news, highlighting mechanisms of misinformation propagation.

5.1 Introduction

As introduced in Chapter 2, with the advent of online social media [195], the access to information has drastically increased, allowing ideas and news to circulate rapidly and

to reach a global audience. This shift has prompted individuals to actively seek and consume information directly through online social media platforms. In contrast to the pre-digital era, where people relied on traditional news media outlets and printed journals for information, the digital age has multiplied and simplified access to these sources. News media journals, once accessible primarily through physical copies, are now readily available with just a few clicks, marking a fundamental change in the way individuals obtain and engage with information [196]. In this context, online social media serves as a digital space where people can look for online news articles as well as come across articles posted by other users [197]. New technologies have increased the possibilities for how people receive and send information [198], The increasing number of accessible online news media articles has surely contributed to the problem of information overload [199, 200, 201, 202], which is also caused by the absence of a proper content regulation on the Internet [203] and might affect the different spheres of democracy, such as the freedoms of thought, belief, and expression [204]. In the current scenario, the spread of misinformative and disinformative content exerts its influence across various social domains, including, but not limited to, the political sphere, international relations, and the formulation of public policies related to climate change [205]. Moreover, it has significant ramifications for numerous health-related issues, as evidenced by the dissemination of misleading narratives by the anti-vaccination movement [206, 207] and, notably, the onset of the COVID-19 pandemic in early 2020 [208]. In this context, the information overload might be exacerbated during highly debated and contentious topics, such as political elections [32] or the COVID-19 pandemic [209], which have been characterized by a notable increase in the dissemination of unreliable content [180]. Based on these premises, the aim of this work is to investigate the intricate news media digital landscape by analyzing the type of news media sources that dominated the online conversation during the first year of the COVID-19 pandemic. In addition, our interest focuses on the characterization of the behavioral patterns that each user displays in the selection of the news media types and web-domains chosen to read and share online. While extensive research has been conducted on news media consumption [210], particularly concerning the transformative impact of the Internet on information acquisition through websites and online articles, a notable gap exists in quantifying the diversity of users' news media diets within the digital and social landscape. Consequently, our interest goes beyond a mere descriptive analysis of user behavior in accessing and staying informed through online social media. We aim to introduce two measures capable of capturing the *variety* of news media consumption and the *regularity* of accessing this information, drawing inspiration from two distinct entropic measures. A deeper insight

into users' news media habits might shed light on the extent to which there is a variety in choosing different news media journal or if they are more likely to stick to a few of them. Decoding which sources online users rely on and share might raise our knowledge about how misinformative and disinformative content spread online and how users are more likely to be attracted by certain web-domains rather than others over time.

5.2 Materials and Methods

5.2.1 Overview of the Dataset

We analyzed online social media data gathered by the COVID-19 Infodemic Platform [211], a comprehensive database containing messages posted on Twitter throughout the COVID-19 pandemic. Specifically, we considered about 9,157,655 messages containing at least one URL (of which we were able to classify 2,549,226) posted in 2020.

We only consider the original messages, defined as tweets, thus not including in our analysis replies, retweets, or quotes. To ensure a robust sample, we considered in our analysis only 25 countries with an average of at least 500 tweets per day collected on our platform. The countries included in our analysis are those associated with the following iso3 codes: COL, IRL, ITA, VEN, TUR, SWE, ESP, BRA, NLD, DEU, JPN, NGU, POL, CHE, AUS, AUT, CAN, CUB, ECU, GRC, MEX, PRT, ROU, SLV, ZAF. In total, we gathered information about messages posted by 211,493 users from the beginning of February until the end of 2020.

5.2.2 News Reliability

In our analysis, we evaluate the reliability of news domains by matching the URLs included in the textual content of messages against information manually verified from various publicly accessible databases that cover scientific and journalistic sources, utilized data from MediaBiasFactCheck [74], as explained in Section 2.3.5. Their methodology involves a comprehensive evaluation of the ideological leanings and factual accuracy of media and information outlets, employing both quantitative metrics and qualitative assessments, <https://mediabiasfactcheck.com/methodology/> (accessed on 9 March 2024). Their evaluation is based on the web-domain itself, without considering the actual news. To assess the reliability score of each source, they review several headlines and news stories. The classification proposed by MediaBiasFactCheck has been further

integrated with a number of other publicly available sources (see [2] for more details), identifying a total of 4417 domains. In the process of developing this domain classification, we also assessed the language representation of the classified web-domains to take into consideration the intrinsic multilingual and multicultural aspects of our analysis. As demonstrated by Gallotti et al. [2], comparing web traffic statistics for the different countries and focusing on the top 50 most visited websites, the domains we classified match the top-tier websites across several countries in different native languages, suggesting the reliability of the results for a comprehensive multilingual and multicultural analysis.

The URLs within the tweets were automatically scrutinized using this domain list, and each URL was categorized based on its source type, such as political, satire, mainstream media, science, conspiracy/junk science, clickbait, fake/hoax. Naturally, a fraction of domains were not present in our list: the most frequent cases being ((i) shortened URLs for which the original link was not possible to reconstruct) and (ii) all the web-domains that have not been classified by external experts.

Building on the approach outlined in [2], we classified news sources as reliable (belonging to the science or mainstream media categories), low-risk (satire, clickbait, political, other), or unreliable (fake or hoax, conspiracy or junk science), as shown in Figure 2.3. The distinction between web-domains, which represent the actual URLs pointing to specific news media sources in the messages, and the categorized types within each domain are pivotal in our analysis, which wants to investigate both the macro and microscopic perspectives of the phenomenon of news “consumption” considering both the aggregation in the news categories and the individual domains.

5.2.3 Building the Information Networks

Each user is characterized by a sequence X of web-domains and news media categories shared during the time considered (see Figure 5.1). Based on these sequences, we reconstructed the undirected weighted networks where the nodes are web-domains and news media types, and the weight assigned to the edge is the frequency of co-occurrence, aggregated by the total number of users. An undirected network is a type of network where connections between nodes do not have a specific direction, meaning interactions are reciprocal between nodes. In our context, it implies that the network is built based on the subsequent posting of web-domains and news types without specifically tracking which one was mentioned first and which one second. Moreover, we quantified the number of times users iteratively shared the same web-domain and news media type,

inspecting the role of *self-loops* in the network, i.e., connections where a node is linked to itself, in depth to gain a better understanding of their appeal towards certain web-domain or news media categories.

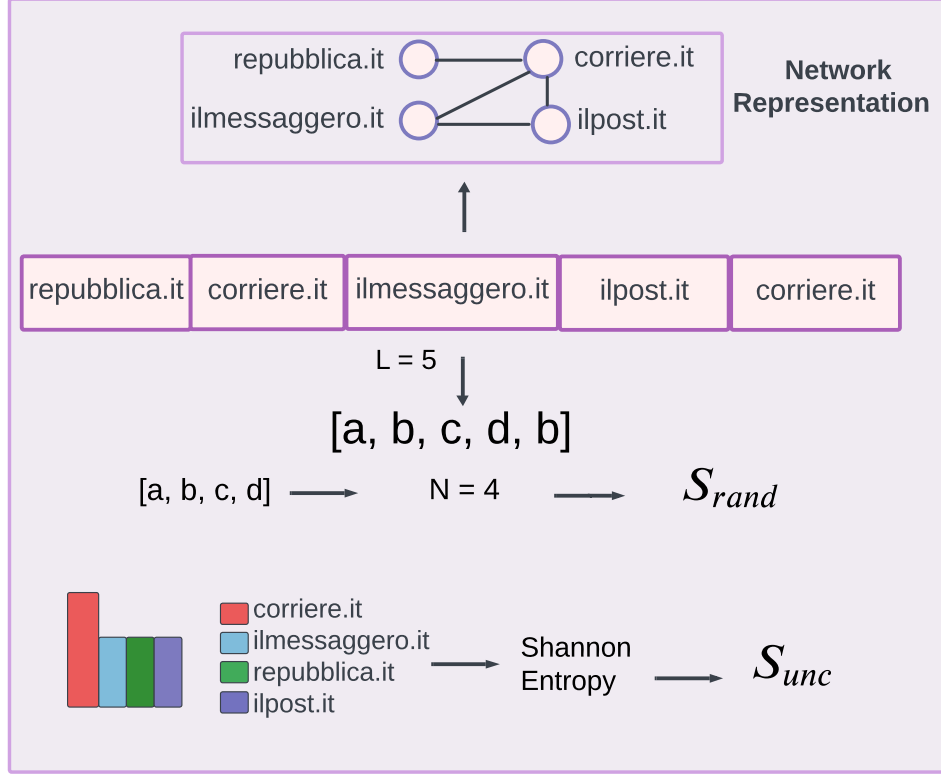


Figure 5.1: **The News Media Diet.** Network and list of the sequence of web-domains posted by a random user on Twitter. We defined the length of each sequence with the term L . Each sequence is decoded in a list of identified letters. On these lists, we calculate the entropy value to obtain information about the variety of each sequence. In particular, we performed both random entropy S_{rand} and Shannon entropy S_{unc} calculations for each sequence.

5.2.4 User Categories

Given a sequence of different web-domains and news media categories (see again Figure 5.1), we studied the news media diet of each user based on the variety of web-domains and news categories shared. We are interested in exploring whether there are any differences in the selection of specific domains and news media categories based on the type of users considered. For this reason, we distinguish users into 5 categories: reliable low-risk users (not sharing any news coming from fake or conspiracy news sources) and high-risk users (sharing at least one fake or conspiracy news piece), classified further

based on their increasing engagement with unreliable sources (more than 1, 5, 10, or 100 tweets in total). In the Appendix C, the same analysis is re-proposed considering more levels of engagement with high-risk sources.

5.2.5 Entropy

Entropy is a key measure to quantify the uncertainty of sources of information and it has played a central role in the field of communication theory [212]. For each user u , we denote by:

$$D_u = (d_1, d_2, \dots, d_L)$$

the ordered sequence of web-domains or news media categories shared by that user, where L is the total number of shared items and each d_ℓ represents a domain or media category.

We employed three different measures of entropy to explore to what extent there is a diversity in the choice of web-domains and news media categories.

First is the random entropy $S_{rand} = \log_2 N_i$, where N_i is the number of different characters in the sequence considered, in this case representing distinctive web-domains or news media types shared by user i . This measure describes the size of the ensemble of different news shared by a user and is useful to illustrate the predictability of the user's news media diet if each web-domain or news media category is shared with equal probability. Second, the temporal-uncorrelated entropy, computed as the Shannon entropy

$$S_{unc} = - \sum_{j=1}^{N_i} p(x_j) \cdot \log_2(p(x_j))$$

where $p(x_i)$ is the historical probability that a web-domain x_i was shared by user i . This second measure thus allows us to characterize the heterogeneity of the news media diet. Lastly, actual entropy, S , depends not only on the frequency of web-domains chosen but also on the order in which each web-domain is shared, capturing the temporal order present in a user's news media diet [213]. By definition [214], these three measures should respect the following inequality

$$S \leq S_{unc} \leq S_{rand}$$

In our data, this inequality did not hold in some cases for the actual entropy S estimated

using the Lempel–Ziv algorithm, which is an accurate estimator for the real entropy of strings sufficiently long [215, 216].

For each user u , the actual entropy is:

$$S(u) = - \sum_{T'_u \subseteq D_u} P(T'_u) \log_2 P(T'_u),$$

where:

- T'_u denotes a time-ordered subsequence of D_u consisting of one or more consecutive elements (e.g., (d_3, d_4) or (d_5, d_6, d_7)),
- $P(T'_u)$ is the empirical probability of observing that subsequence within the user’s history.

However, this is not always the case in our data. For this reason, we decided to take into account only users who had at least three different web-domains and news media categories in order to exclude flawed entropy estimates from our analysis.

To complete our analysis, we also calculated the number of repetitions of web-domains and fact types over the total number of potential couples over the length of each sequence in order to explore the weight of these repeated sequences of web-domains and news media categories shared by each user.

More specifically, given a sequence X of length L , where $X = [x_1, x_2, \dots, x_n]$, we counted the number of repeated pairs $x_j = x_{j+1}$ over the total number of potential pairs $L - 1$ to gain insights into the behavioral patterns in the sharing of information, considering different type of users.

5.3 Results

We consider the sequence of messages containing a URL and the corresponding news media category for more than 211,493 user accounts which have posted at least two different original tweets during 2020 (Figure 5.1). Each sequence has length L and has a specific number N of unique domains posted. In this work, we decided to remove all users having posted only a unique web-domain during the period considered. These user accounts represent 24% of our sample, leaving 160,228 unique users. In a second moment, as discussed above, we also decided to remove all users who shared only three different types of web-domains or news media categories in order to preserve the inequality of the three entropy measures [214], leaving almost 18,000 unique users.

We first characterize the sequences. Figure 5.2 shows the probability distribution function of the sequence length L of the (panel A) web-domains, considering the entire dataset (Total), the dataset containing only those users who have two or more unique domains shared ($N > 1$), and the different types of users we introduced in the Methods Section. Panel B shows the distribution of the number of tweets associated with each news media category in our dataset. The results confirm that the mainstream media and “other” categories are the main news media categories shared. Panel C reveals the distribution of the number of tweets associated with the different news types shared by one particular user who had shown a great activity during the time considered.

This user was classified as *reliable/low risk* since he/she did not post any URL associated with conspiracy and fake news sources.

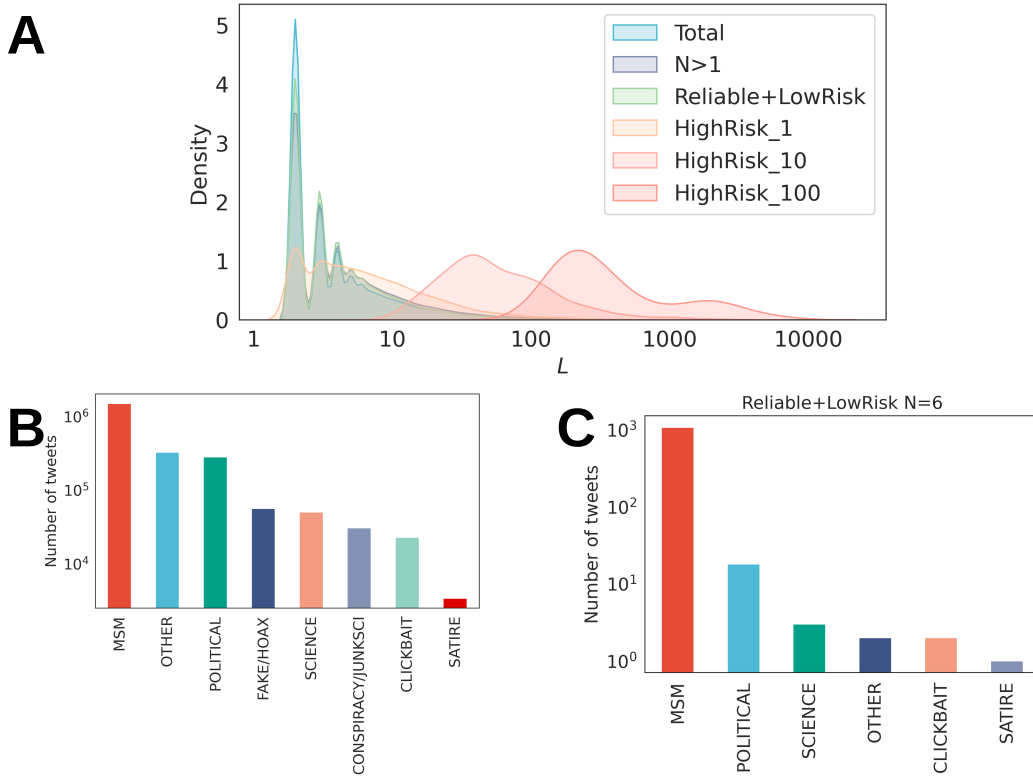


Figure 5.2: **Overview of the dataset.** (A) Distribution of the length sequence of web-domains, respectively, for the entire dataset, for users having posted more than 2 distinctive web-domains ($N > 1$) and users defined according the categories of reliability. (B) Distribution of the number of messages classified in one of the eight categories of news media types. (C) Distribution of the news media categories used by a particularly active user who has posted reliable and low-risk content, with 6 distinctive categories of news.

To give an overview of the news media environment as a whole, we reconstructed the

undirected network of interactions among the web-domains shared by each user, where the nodes are represented by the web-domains and the weight is given by the number of times those couples of web-domains have appeared in the sequence of posting. The network has been built at a coarser granularity having media categories as nodes in order to describe the inter-relations of web-domains of the different categories posted. In Figure 5.3, panel A shows that it is more likely to have posted a message containing a mainstream media source and news with a strong political bias, while panel B displays a heatmap quantifying the mean value of the relative value of the weight among the different news media categories.

It is expected that links connecting the larger categories (MSM, political, and other) would have larger weights and would appear dominantly in the network we built. To compensate the effect size of the web-domains distributed across the different news categories, we compare our findings with a null model that takes into account the relations normalized by the proportions of news that belong to each category, assuming a random sequence of news shared. Differently from the previous analysis, we found that, after compensating for the category dimensions, a strong tendency to alternate between fake and conspiracy news emerges among users, as shown by panel C and by the corresponding values presented in panel D. Differently, the relationship between the category science and the other categories is more equally distributed, as shown in panel D.

In light of this strong relation between fake and conspiracy news, we refer to the users who have posted fake or conspiracy news as *high risk*, as already introduced in the Methods Section.

Moving from this first result, we decided to gain better insight into the intra-relations among domains and news media classes, also called *self-loops*. The analysis displayed in Figure 5.4) reveals that the distribution of mainstream media is much broader with respect to the other classes when considering the web-domains chosen. Panel A highlights that users sharing MSM articles are more likely to be further sharing other articles from the same news category than users sharing other categories of media. Differently, the category “other”, encompassing general content that is not easily classified, such as videos on Youtube or posts on Instagram, shows that the percentage of domains chosen is limited to a few within this category. Conversely, panel B shows that users sharing MSM have a lower tendency to repeatedly share the same particular domain, while repetitions of the same domain are more frequent for sequences of conspiracy theory or clickbait content.

Beyond this analysis about the general patterns of news sharing in online media,

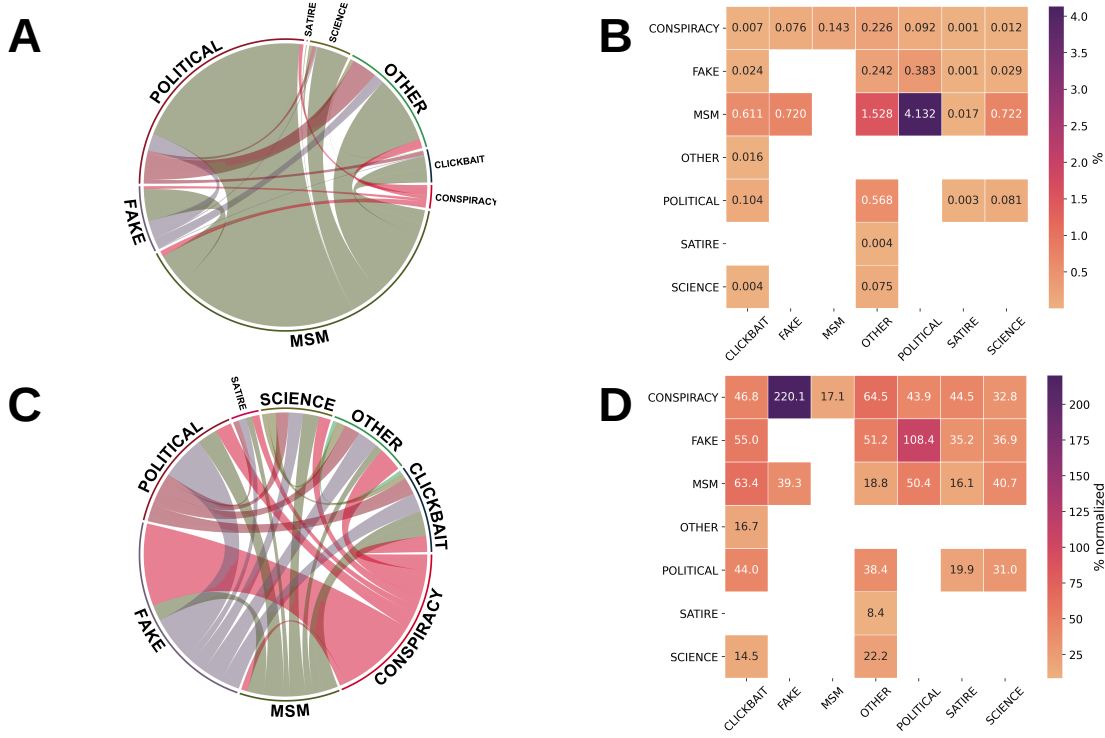


Figure 5.3: **The news media environment.** (A) Weighted networks of the interactions among different news media types (also known as intra-relations) for 25 countries in 2020. (B) Heatmap showing the corresponding value of the inter-relations among the eight categories. (C) Weighted networks normalized and compared with a null model accounting for the proportion of the number of news pieces belonging to each category. The chances of sharing conspiracy and fake news by the same users is much higher than the strong relation observed between mainstream media and political news (A). (D) Heatmap showing the corresponding normalized value of the inter-relations among the eight categories.

our focus extends to a more detailed exploration of users' news consumption habits. In this regard, we classified each user into distinct groups: those not sharing fake and/or conspiracy news and those showing different levels of unreliability (*high risk*, 1, 5, 10, and 100) based on their increasing engagement with high-risk news (see Methods). In this context, we assigned to each user the three measures of entropy to better understand the news media diet and to gain insights into the potential differences between different kinds of users.

Figure 5.5 shows the distribution of the random entropy S_{rand} (panel A), the compression rate between the uncorrelated entropy S_{unc} and the random one S_{rand} (panel B), and the compression rate between S and S_{rand} for different types of users (panel C). To validate our results, we calculated the proportion of web-domains and news media

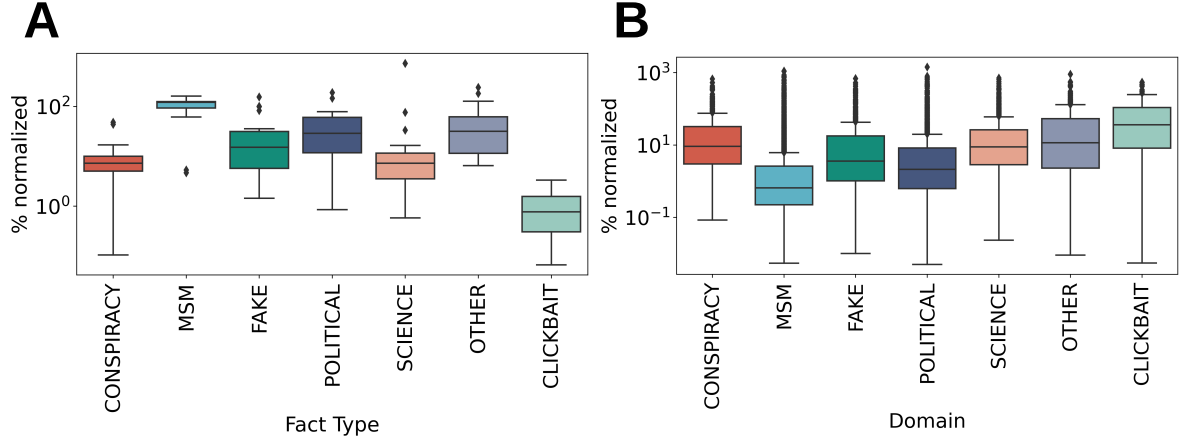


Figure 5.4: **Self-loops of web-domains and types of news shared.** (A) The percentage distribution of the number of self-loops, pointing to the same web-domains for all the categories considered. (B) The percentage distribution of the number of self-loops of users pointing to the same category of news, regardless of the different web-domains shared.

types classified as high risk relative to the total within their respective classes. This allowed us to extract a control sample of random web-domains with a similar size to that of high risk domains. We also tested that our results are not influenced by any potential selection biases associated with the definition of our user categories by randomly reshuffling the news categories in our tweet ensemble and re-assigning users in these shuffled categories. This process resulted in a constant value of entropy across all categories and confirms absence of bias. These mean entropy values obtained from reshuffled samples (indicated by the dashed lines) are consistently lower than those of the four high-risk groups, especially for the S_{rand} , which we interpret as a measure of variety in the news media diet. Users sharing reliable or low-risk content seems to have a more restricted news media diet not only when considering the news media categories chosen but also when looking at the number of domains shared. Differently, users classified as high risk show higher values of random entropy as their engagement with these types of content increases, indicating that the consumption and sharing of high-risk content can be associated with wider exploration of a larger number of different news categories and domains shared. To compare the different entropy measures and quantify the degree of structure in a user’s news-sharing sequence, we compute the *compression*

rate. In this study, we focus on two specific ratios:

$$C_{\text{unc,rand}}(u) = \frac{S_{\text{unc}}(u)}{S_{\text{rand}}(u)}, \quad C_{S,\text{rand}}(u) = \frac{S(u)}{S_{\text{rand}}(u)}.$$

The first ratio, $C_{\text{unc,rand}}$, measures how heterogeneous a user’s empirical distribution of web-domains is compared to the maximally uniform case. The second ratio, $C_{S,\text{rand}}$, captures the extent to which the temporal order of shared domains introduces additional regularity in the sequence.

The compression rate between S_{unc} and S_{rand} appears relatively flat for all the types of users considered, but here the pattern for high-risk users is more compressible than the random expectations. This tendency is even stronger in the third panel (C), which displays the compression rate between S and S_{rand} as a measure of regularity of the news media diet. In this case, the compression rate for high-risk users is much lower than the second panel (B). This indicates that the same heavy consumers and spreaders of high-risk content that we see have a tendency to explore a larger number of news categories and domains, displaying at the same time a stronger regularity in their choice of web-domains and news media categories. To validate our results, we also performed the same analyses by differentiating our users into verified and unverified users as well as taking into consideration a wider classification of high-risk users, demonstrating that the same pattern is recognizable in such scenarios, as shown in the Appendix C

In order to gain deeper insights into this counter-intuitive finding, we study the relations between S_{rand} and the percentage of news media types and web-domains repeated, shedding more light on users’ tendencies to either explore new web-domains and categories or stick to a few of them. Figure 5.6 illustrates the values of N of different news media category (or web-domains) shared against the percentage of news media repeated for reliable/low risk and high-risk users, as defined in the methods section . Panel A demonstrates that as the number of news media types increases, the percentage of fact types repeated grows. The last distribution characterizing the users with the highest value of S_{rand} entropy are the ones with a higher number of mainstream media and political journals repeated. The expectation of repeating the same fact-type assuming a homogeneous distribution is instead expected to be decreasing for growing N , as it is trivially given by the ratio $1/N$, highlighted by the dashed line in all panels. The distribution of repeated news media categories for users posting six unique news media categories therefore displays a different tendency and is significantly higher than the random expectation for larger N . However, the same behavior is not shown when looking at users posting high-risk content, as shown in panel B. Indeed, the distribution

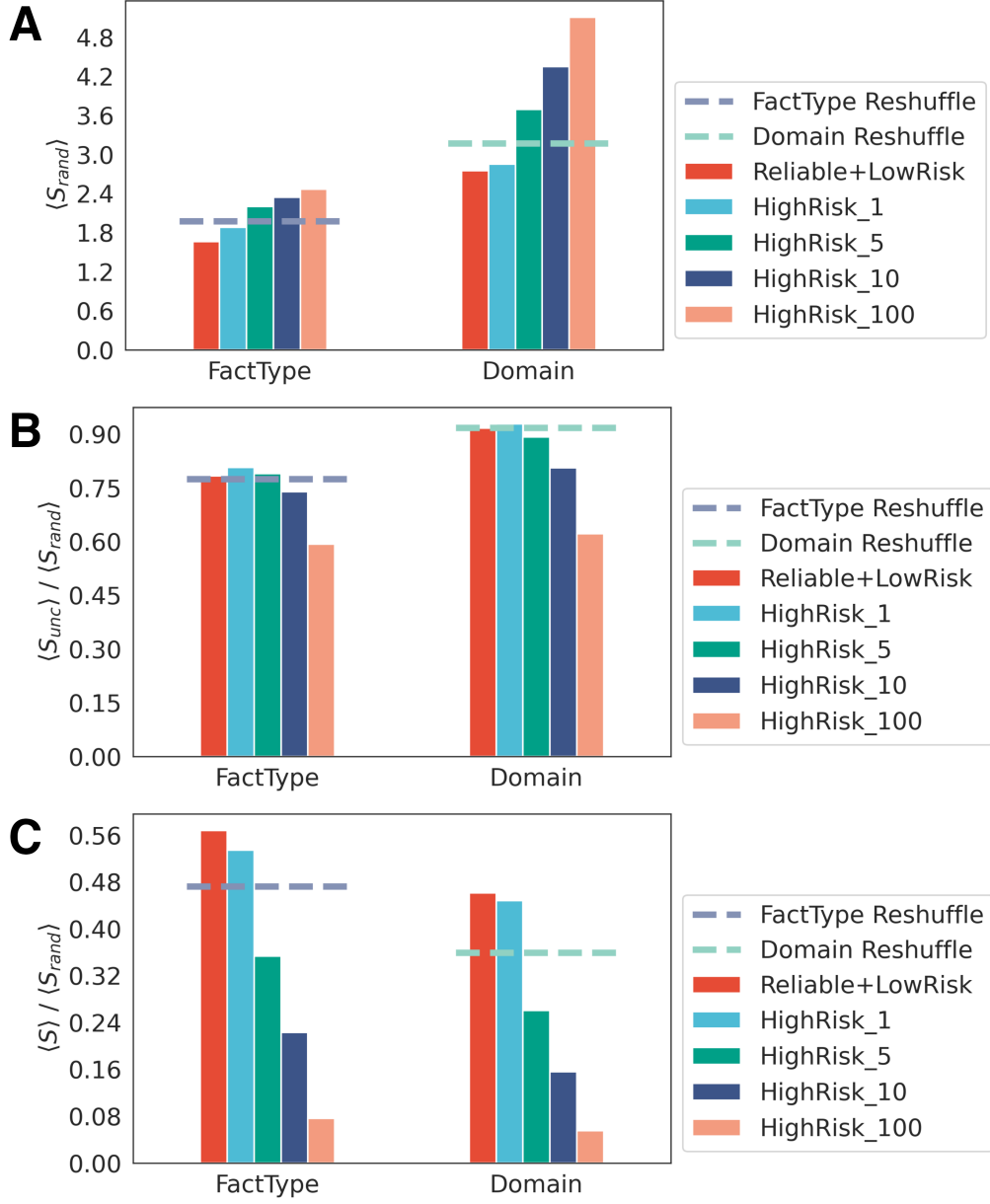


Figure 5.5: **News media diet for different types of users.** (A) Random entropy S_{rand} calculated for different types of accounts: users posting reliable or low-risk content and users posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon entropy S_{unc} calculated for the domain and the type of news shared by different types of users accounts: users posting reliable or low-risk content and users posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated for the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content.

5.3. Results

of the percentage of news media types shared is slightly higher for users with a lower S_{rand} . The last distribution is represented by people having posted all the possible news categories ($N = 8$). Besides the news media classes, we analyzed the percentage of web-domains repeated with respect to the number of unique domains shared for reliable or low-risk users (panel C) and high-risk users (panel D). In this case, the distribution of domains repeated shows to be fixed at 25 %, independent from the size of $\langle N \rangle$, both for reliable and low-risk and high-risk users. Interestingly, this result indicates that the percentage of repeated domains is the same regardless of the type of user and the number of domains or news media categories shared.

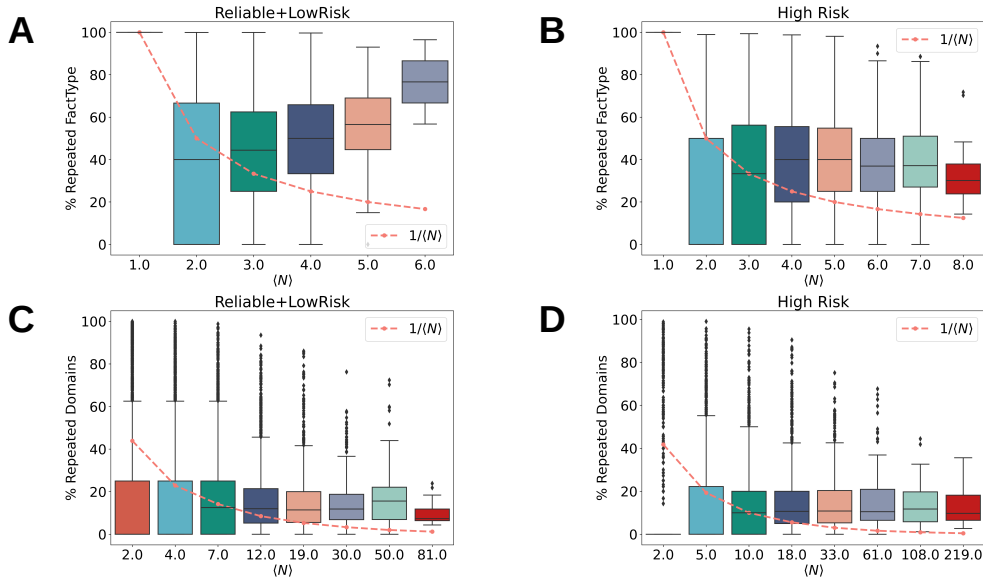


Figure 5.6: Fraction of repeated news media categories and web-domains for users with different N . In these boxplots, we represent the percentage of news media categories (A,B) or web-domains (C,D) repeated in subsequent tweets with respect to the total length of the sequence associated with an individual user. Each boxplot is aggregated over users with the same N . The dashed line shows the random expectation of repeated news media categories with respect to the increasing number of N unique news media categories. (A) Patterns of news media categories, aggregated for users classified as having posted only reliable and/or low-risk content on Twitter. There are 8 box-plots here since we have 6 distinctive categories and 2 are filtered out. (B) Patterns of news media categories, aggregated for users having also posted high-risk content. There are 8 box-plots here since we have all 8 distinctive categories of news media types. (C) Patterns of web-domains, aggregated for users classified as having posted reliable and/or low-risk content. (D) Patterns of web-domains, aggregated for users classified as having also posted high-risk content in their sequence of messages posted.

5.4 Discussion

Our research reveals that the news media digital landscape is a complex environment where users tend to be attracted to and thus repeat specific web-domains or news media categories. In particular, when we inspected the news media environment, we found that there is a strong relation between the probability of sharing mainstream media journals and political media. This relation is just apparent because if we compare our results with a null model that takes into account the effect size of each category, our analysis highlights how users sharing fake or hoax news are more likely to alternate with conspiracy and junk science types of news, illustrating the tendency for unfortunate misinformation hot streaks. This result also suggests how easy it is to fall into this vicious circle where alternating fake and conspiracy news becomes the norm. Differently, the news categories representing the most reliable ones (mainstream media and science) do not show this strong relationship, leaving the most unreliable categories at the center of the news media digital landscape. The habit of alternating between fake news and conspiracy theories suggests a strong inclination towards reinforcing one's own opinions, often by seeking out narratives that align with pre-existing beliefs in the digital environment. This phenomenon reflects a broader trend of confirmation bias [217], wherein individuals gravitate towards information that confirms their worldview while disregarding contradictory evidence. In the digital scenario, where an abundance of information is readily accessible and algorithms personalize content recommendations based on user preferences, individuals may find themselves immersed in echo chambers where their perspectives are constantly reinforced [218]. This echo chamber effect can amplify the consumption of unreliable and misleading narratives, perpetuating the misinformation hot streaks. Furthermore, the sensational and provocative content can make fake news and conspiracy theories particularly appealing to certain audiences, evoking strong emotional responses and exploiting cognitive biases [46], such as the availability heuristic or the illusion of explanatory depth, which can lead individuals to accept them uncritically. The tendency of alternating fake and conspiracy news has raised the need for better understanding which are the characteristic features of each user, specifically when we consider different types of users. In this scenario, one might suggest that different types of users, based on their interactions with different type of content, have completely different news media habits. This is what our research shows.

In this context, we introduced two different measures to characterize and quantify the users' behavior in the way they access and retain information online, shaping their news media diet. The *variety* in the so-called news media diet is given by S_{rand} , while

the *regularity* is calculated by the compression rate between S and S_{rand} . By employing these measures of entropy, we found that reliable or low-risk users have more restricted news media habits when searching for news media categories and web-domains. On the contrary, high-risk users at different levels reveal having a more varied news media diet, especially for users sharing more than 100 high-risk web-domains. Interestingly, the ratio between S_{unc} and S_{rand} demonstrates that reliable or low-risk users have less compressible sequences if compared with high-risk users. This is even more explicit when we observe the compression rate between the real entropy S and the random entropy S_{rand} , which highlights how high-risk users display a remarkable regularity in their news consumption, even if they are characterized by having a more diverse diet if we just consider the random entropy, which takes into account only the number of unique web-domains and news media categories. Getting deeper into this study, we found that the repetitions of news media categories is in general higher than expected if we look at the users characterized by a huge variety of news media categories selected. Interestingly, this effect is stronger for reliable and low-risk users. In that context, we observed that those users are usually also repeating more often political news media categories and represent particularly active users. On the contrary, the repetitions of web-domains seems independent to the variety of the news diet, whether the user is a reliable or low-risk or high-risk one, showing that there is the same probability of sharing web-domains, independently of how much the user has wider news media habits.

This work has shed light on news media behavior and the patterns of the users during highly contentious events, such as the COVID-19 pandemic, revealing how much certain types of users have more varied but at the same time regular news diets as they are more likely to share and to repeat the same news media sources. In particular, we found that this happens for users heavily sharing high-risk disinformation sources. At the same time, we also observed from an aggregated perspective how the news media digital environment is shaped according the strong relationship of fake and conspiracy news: highly debated and unprecedented topics can generate debates driven by fake and conspiracy news, which can monopolize the attention of particular users. These users with a greater tendency to engage with this high-risk content are also those more likely to have a wider content base but are also the ones more likely to express a strong regularity in the way they share news media content online over time.

Our findings are based on an aggregated sample of countries having at least 500 tweets per day, showing an aggregated and global tendency. For the purpose of comparison, we decided to pick some countries with lower and higher values of press freedom in order to observe if certain trends change by considering only specific countries. As

shown by comparing S5 and S6 and S7 and S8 (see Appendix C for more details), this is not the case; we observe a consistent pattern across countries, indicating a significant presence of misinformation and conspiracy narratives within their respective media landscapes. The process of collecting data and integrating different sources of user data provides us with the opportunity of analyzing complex human behavior phenomena. However, it is important to acknowledge that our study does not claim to be complete. Indeed, we decided to analyze only Twitter data, which might contain population biases. Twitter is indeed an online social network mainly used by well-educated males (63% of Twitter users [219]) between the ages of 18 and 34 (56% of Twitter users, according to Statista [220]). Further research might explore the news media environment and diet of users in other online contexts, highlighting differences and similarities among several social media platforms. These findings underscore the imperative for a more nuanced comprehension and proactive analysis of the ramifications of online social systems, alongside a discerning evaluation of the offline risks engendered by the dissemination of streaks of misinformative content within the digital sphere. This is certainly a topic that merits additional analyses and holds significant potential for enhancing our comprehension of the news media habits of users during contentious debates.

Chapter 6

The interplay between socio-economic indicators and infodemic risk indices

While Chapter 5 examined information diversity at the individual level, the following work scales up the perspective by exploring the dynamics of several countries in infodemic patterns. In particular, this work proposes a validation of previous infodemic risk indexes using socio-economic indicators and the framework of news diversity. The main contributions of this work are the following:

- we identify multidimensional socioeconomic factors that shape cross-national vulnerability to COVID-19 infodemics using dimensionality reduction and clustering techniques.
- we reveal distinct temporal patterns of infodemic risk and show that news media diet diversity mediates the relationship between socioeconomic development and misinformation exposure.
- we demonstrate that institutional quality and democratic stability reduce infodemic volatility, stabilizing information environments during crises.

6.1 Introduction

In this chapter, our focus is still on the COVID-19 Pandemic, which has been a major phenomenon followed by an increase amount of misinformation. However, the concept of information disorder during public health emergencies is not entirely new; historical instances such as the HIV/AIDS epidemic [221], H1N1 outbreak [222], and Ebola crisis [223] have all been accompanied by waves of misinformation. However, the scale,

speed, and impact of the COVID-19 infodemic has been dramatically amplified by the digital transformation of our information ecosystem, particularly the rise of social media platforms as primary news sources for significant portions of the global population [224].

The consequences of the infodemic extend far beyond mere confusion. Studies have linked exposure to COVID-19 misinformation with reduced compliance with public health measures [225], increased vaccine hesitancy [226], exacerbated social tensions [227], and even physical harm from dangerous remedies promoted through false claims [228]. As [2] argue, infodemics can substantially increase epidemic spread by undermining confidence in scientific evidence and public health authorities, thereby hampering coordinated social responses needed for effective pandemic control.

While early research on misinformation often emphasized technological factors such as algorithmic amplification [229] [218] or the absence of traditional gatekeepers in digital media [230], a growing body of literature recognizes infodemics as complex socio-technical phenomena embedded within broader social, political, and economic structures [231] [232]. This shift in perspective is crucial, as it suggests that societal vulnerability to information disorders cannot be addressed through technological solutions alone, but requires an understanding of the underlying socioeconomic conditions that create fertile ground for misinformation to flourish.

Recent studies have begun to explore how various social factors might influence misinformation dynamics. [233] found that countries with strong public service media, higher levels of trust in news, and lower levels of political polarization showed greater resilience against online misinformation. [234] highlighted how existing social divisions and identities shape individuals' receptivity to misinformation. [39] demonstrated that false news travels faster and reaches more people than accurate information, particularly when it evokes emotional responses of fear, disgust, or surprise. These studies collectively suggest that social, cultural, and economic factors play critical roles in determining how information and misinformation spread through populations.

However, empirical research systematically examining the relationship between comprehensive sets of socioeconomic indicators and infodemic vulnerability across multiple countries remains scarce. Much of the existing literature focuses on single countries, such as Brazil [235], or on individual-level factors rather than structural conditions [236] [237]. While these studies provide valuable insights, they offer only a partial understanding of how broader societal characteristics might influence a nation's overall information environment during a global crisis.

Social media platforms have become central to crisis communication during the

COVID-19 pandemic [238] [239], serving as critical channels for rapid information dissemination from health authorities [240], spaces for public discourse and sense-making [241], and unfortunately, vectors for the spreading of misinformation [208]. Twitter, in particular, has emerged as a significant platform for COVID-19 communication due to its real-time nature, public accessibility, and adoption by key stakeholders including government officials, health organizations, journalists, and public health experts [242] [114].

Analyses of Twitter data have yielded important insights into information flow patterns during the pandemic. Various studies have examined the prevalence of misinformation in COVID-19 related tweets [243] [244], the role of bots in spreading false claims [245] [246], emotional responses to pandemic developments [247] [248], and the evolution of topics in public discourse [249] [250]. [2] developed a quantitative measure of “infodemic risk” based on the reliability of news sources shared on Twitter, creating a framework for cross-national comparisons of information quality during the pandemic.

While these studies have advanced our understanding of digital communication dynamics during the COVID-19 pandemic, they have generally not systematically explored how these patterns correlate with the underlying socioeconomic characteristics of different countries. This represents a significant gap in our knowledge, as understanding these relationships could provide crucial insights for developing more effective strategies to combat infodemics.

The broader literature on media systems and information environments suggests numerous ways in which socioeconomic factors might influence a society’s vulnerability to infodemics. Educational attainment, particularly levels of scientific and digital literacy, may affect a population’s abilities to critically evaluate health information [251] [252]. Political polarization can lead to motivated reasoning and selective exposure to ideologically congruent information, regardless of its accuracy [253] [254]. Economic inequality may exacerbate information disparities and create segmented information environments [255] [256]. Media freedom and pluralism shape the diversity and quality of available information sources [233].

Countries differ substantially across these dimensions, suggesting potential variation in their vulnerability to infodemics. However, these factors likely operate in complex, interrelated ways rather than in isolation. For instance, Hallin and Mancini’s [257] influential work on comparative media systems demonstrated how media institutions are deeply embedded in broader political and social structures, creating distinctive patterns across different types of democracies. Similarly, information ecosystems during crises likely reflect complex interactions between multiple socioeconomic dimensions rather

than simple correlations with individual factors.

Despite the growing recognition of infodemics as a socially embedded phenomenon, several important research gaps remain. First, few studies have systematically examined how comprehensive sets of socioeconomic indicators relate to infodemic patterns across multiple countries. Second, existing research has largely overlooked the temporal dynamics of infodemics, i.e., how they evolve over time and whether their relationship with socioeconomic factors changes during different phases of a crisis. Third, the role of media diet diversity as a potential mediating factor between socioeconomic conditions and infodemic vulnerability remains relatively underexplored.

Our study addresses these gaps by investigating the relationship between 20 socioeconomic indicators and patterns of infodemic risk across 37 OECD countries during the COVID-19 pandemic. Using dimensionality reduction techniques, we identify underlying structures in our socioeconomic data and examine how these dimensions correlate with various measures of infodemic risk at different time points. Additionally, we explore how the approach proposed in 5, related to the diversity of news media diets relates to both socioeconomic factors and infodemic vulnerability. Specifically, our research aims to answer the following questions. First, what underlying dimensions can be identified in the socioeconomic indicators of OECD countries, and how do countries cluster along these dimensions? Moreover, do countries exhibit distinct clusters based on their temporal patterns of infodemic risk during the COVID-19 pandemic? In addition, how do socioeconomic factors correlate with different measures of infodemic risk (overall risk, dynamic risk, and unreliability), and do these correlations change over time? Furthermore, what role does news media diet diversity play in mediating the relationship between socioeconomic factors and infodemic vulnerability? And finally, how does infodemic volatility relate to socioeconomic factors, and does this relationship evolve over the course of the pandemic? By addressing these questions, our study contributes to both theoretical understanding and practical approaches to the management of infodemics. Theoretically, we advance a more nuanced framework for understanding infodemics as phenomena embedded within complex socioeconomic contexts, moving beyond purely technological or individual-level explanations. Practically, the identification of socioeconomic correlates of infodemic vulnerability can inform more targeted and effective interventions to build societal resilience against misinformation during future public health emergencies. In the following sections, we describe our methodology for collecting and analyzing Twitter data to measure infodemic risk across countries, as well as our approach to gathering and processing socioeconomic indicators. Subsequently, we present results from our dimensional analysis of socioeconomic factors, our identifi-

cation of infodemic risk clusters, our examination of correlations between socioeconomic dimensions and infodemic measures, and our exploration of the role of news media diet diversity. Finally, we discuss the implications of our findings for understanding and addressing infodemics in the context of global public health crises.

6.2 Methods

6.2.1 Twitter dataset

As explained in 2.3.5, also the analysis of this chapter relies on data collected through the COVID-19 Infodemic Platform [211], using the reliability assessment provided by MediaBiasFact [74]. The URLs in our dataset were classified into seven distinct categories: reliable sources (Mainstream Media, Scientific Journals), unreliable sources (Satire, Clickbait, Political, Fake/Hoax, Conspiracy/Junk Science), and unclassified sources. Building on this classification system, we employed three metrics developed by [2] to quantify different dimensions of infodemic risk (see 2.3.5 for more information about Infodemic Risk Index (IRI), Dynamic Infodemic Risk Index (DynIRI), and Unreliability Index).

6.2.2 Socioeconomic indicators dataset

To examine the relationship between infodemic risk and socioeconomic factors, we compiled an extensive dataset of socioeconomic indicators for 37 OECD countries. This selection ensured data consistency and reliability across all indicators. Our initial compilation encompassed 14 socioeconomic indicators with complete data for all countries in our sample. We subsequently enriched this dataset with six additional indicators from the OECD’s comprehensive database, resulting in a final set of 20 indicators spanning multiple dimensions of social, political, and economic development.

The selected indicators capture diverse aspects of national development: democratic quality (EIU Democracy Index, EIU Political Participation); civil liberties (Freedom House Freedom Index); societal stability (Global Peace Index, Global Terrorism Index); economic prosperity (GDP per capita); educational attainment (percentage of adults with tertiary education, percentage of foreign students, PISA mathematics and science scores); social welfare (social security expenditure, social spending); institutional trust (trust in government); media freedom (Press Freedom Index); digital participation (E-Government Participation Index); demographic structure (median age); social

cohesion (political polarization, social polarization); digital development (internet use percentage); and economic inequality (Gini coefficient). A detailed description of each indicator, including its source and measurement methodology, is provided in D.

6.2.3 Dimensionality Reduction Techniques

The *curse of dimensionality* arises in high-dimensional datasets where the number of features greatly exceeds the observations, causing issues such as sparsity, computational inefficiency, overfitting, and challenges in visualization. Dimensionality reduction techniques mitigate these problems by mapping data to a lower-dimensional space while preserving essential information. This enables an easier visualization of patterns and clusters in two or three dimensions and reduces noise and redundancy in features. By simplifying the dataset, these methods enhance interpretability and improve the performance of analytical models on complex data.

UMAP

We employed Uniform Manifold Approximation and Projection (UMAP) as our primary dimensionality reduction technique. UMAP represents a state-of-the-art algorithm that effectively visualizes high-dimensional data while maintaining an appropriate balance between local and global structure preservation [258]. UMAP is moreover computationally fast and scales well with regard to both dataset size and dimensionality. It works by approximating the underlying manifold of the data and projecting it into a low-dimensional space, typically two or three dimensions, through two key phases: graph construction and graph projection. In the first phase, a high-dimensional graph is created using a fuzzy simplicial complex, where the edge weights indicate the connection strength between the data points. The second phase optimizes this graph to a low-dimensional Euclidean space using cross-entropy. UMAP's tunable hyperparameters, such as `n_neighbors`, `min_dist`, and the distance metric, enable users to control the embedding's granularity and structure, though careful tuning is required to avoid distortions. Compared to other techniques, UMAP excels in capturing non-linear patterns and is faster than t-SNE, while also supporting a variety of distance metrics [259].

6.2.4 k-means

K-means clustering is a popular and straightforward method for partitioning data into k distinct and non-overlapping clusters. The algorithm works by minimizing the total

within-cluster variation, a measure of similarity of the data points within each cluster, often calculated using the squared Euclidean distance. The process begins with an initial random assignment of cluster labels to the data points. Then, it iteratively refines the clusters by computing the centroids of each cluster and reassigning each data point to the cluster with the closest centroid. This optimization process continues until convergence, ensuring that the objective function (minimizing within-cluster variance) decreases at each step.

While k-means is computationally efficient, intuitive, and widely applicable, it has significant limitations. One major drawback is its sensitivity to the initialization of cluster centroids; different initializations can lead to varying outcomes, as the algorithm is only guaranteed to find a local optimum. Another challenge is the need to specify the number of clusters (k) in advance, which is often not straightforward. Choosing an incorrect value for k can lead to poor results, such as clustering noise or failing to capture meaningful groupings in the data. Several heuristic methods, like the elbow method and silhouette score, have been developed to guide the selection of k . The elbow method involves plotting the within-cluster sum of squares against different values of k and identifying the point where the decrease in variance begins to level off, forming an “elbow”. The silhouette score evaluates how well each data point fits within its assigned cluster compared to other clusters. However, these methods are not always consistent or reliable, and there is no definitive consensus on the best approach to determine k . Despite these challenges, k-means remains a robust and widely used clustering technique, especially for large datasets where simplicity and speed are critical.

6.2.5 News Media Diet

To better understand the relationship between the Infodemic Risk Index and socioeconomic indicators, we used uncorrelated entropy as a measure to capture the variety of news outlets in each country. We recall here the concept of news media diet, as discussed in [260]. With the term *news media diet*, we refer to the collection of web domains or URLs shared in each country. In particular, we employed the temporal-uncorrelated entropy, computed as the Shannon entropy

$$S_{unc} = - \sum_{j=1}^{N_i} p(x_j) \cdot \log_2(p(x_j))$$

where $p(x_i)$ is the historical probability that a web-domain x_i was shared by user i . This measure enables us to characterize the heterogeneity of the news media diet.

6.2.6 Volatility

To assess the stability of information environments across countries, we calculated the volatility of the Infodemic Risk Index, defined as the standard deviation of the IRI over time for each country. Initial analysis revealed a strong negative correlation between the number of tweets posted daily and the volatility of the index, indicating that countries with higher tweet volumes exhibited more stable IRI values over time. This correlation follows an expected statistical relationship, where the standard deviation scales with the square root of the sample size $\sigma \approx (N^{0.5})$. To account for this effect and obtain a measure of volatility independent of sample size, we adjusted our volatility metric by multiplying the standard deviation by the square root of the average daily tweet count: $\sigma^* = \sigma * \langle N \rangle^{-0.5}$. This adjustment enabled more meaningful cross-national comparisons of infodemic volatility by controlling for differences in tweet volume across countries.

6.3 Results

This study explores the relationship between infodemic risk indices and 20 socioeconomic indicators across 37 OECD countries, employing dimensionality reduction techniques to identify underlying patterns and correlations. Our analyses reveal complex relationships between the socioeconomic fabric of nations and their vulnerability to infodemics during the COVID-19 pandemic.

Firstly, we used UMAP, a dimensionality reduction method, to project data from socioeconomic indices and countries in a two-dimensional space, allowing us to discern the underlying structure of the data more effectively.

The application of UMAP dimensionality reduction to socioeconomic data, presented in Figure 6.1, reveals distinct structural patterns that transcend simple geographic or economic categorizations. The country projection (Figure 6.1, left panel) demonstrates clear geopolitical and cultural clustering: Scandinavian and Central European nations congregate in the upper right quadrant, Anglosphere countries position to the right, while South American nations cluster in the bottom left. This visualization captures deeper institutional and cultural similarities that unite countries beyond their geographic proximity, reflecting historical patterns of institutional development and socioeconomic evolution.

The arrangement of socioeconomic indicators (Figure 6.1, right panel) along a curvilinear distribution provides equally meaningful insights. Welfare and demographic in-

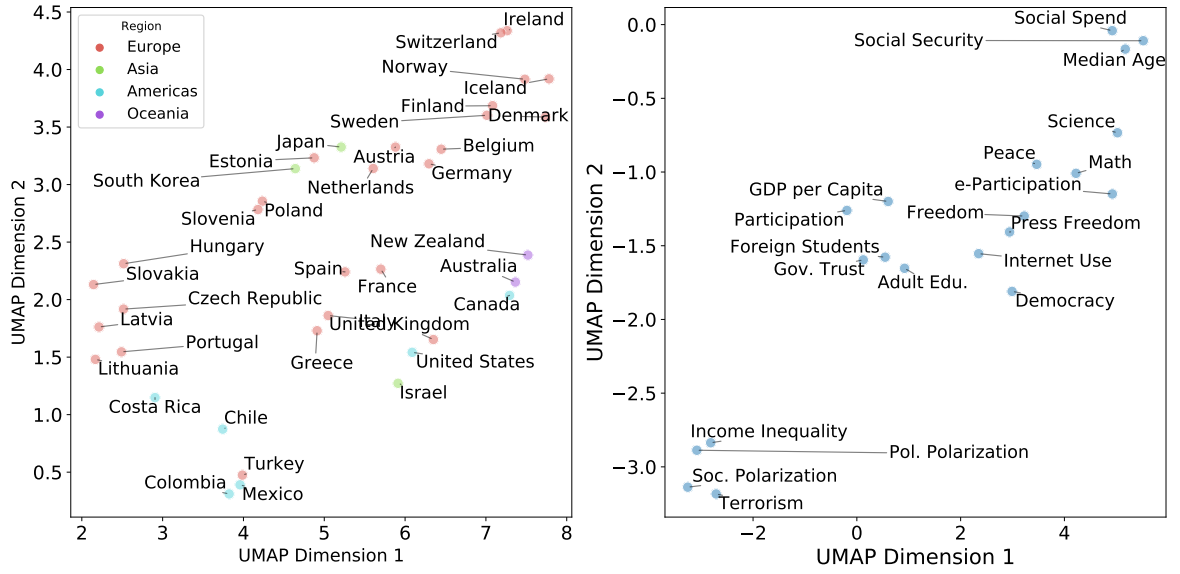


Figure 6.1: **UMAP visualization of OECD countries and the socioeconomic indices.** UMAP visualization of the OECD* countries based on the selected indices (left) and of the selected socioeconomic indices (right). The employed hyperparameters are $n \text{ neighbors} = 5$ and $\text{min dist} = 0.01$.

dicators (median age, social security, social spending) position at the upper region, followed by democracy, education, and economic development indicators in the central area. Notably, indicators of social and political polarization appear distinctly separated in the bottom left, alongside measures of income inequality and terrorism.

This separation suggests that polarization constitutes a distinct dimension of social development that can vary independently of traditional development metrics. The effectiveness of UMAP in discerning these multidimensional relationships shows how certain societal characteristics cluster together, forming coherent socioeconomic profiles that may influence information environments during crises.

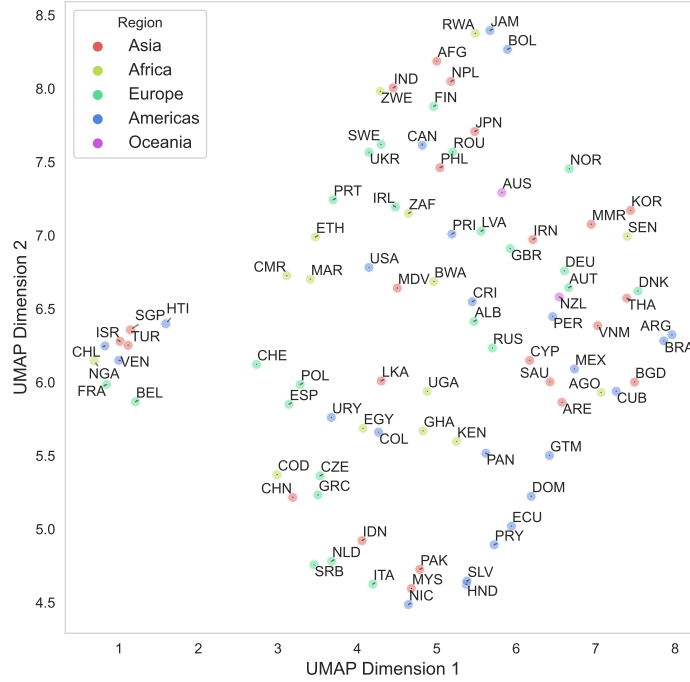


Figure 6.2: **UMAP visualization of IRI timeseries.** UMAP plot of the IRI time series of the countries featuring at least 500'000 tweets in the relevant time frame. The selected parameters are $n\text{ neighbors} = 5$ and $\text{min dist} = 0.1$. The 9 countries belonging to Cluster 2 are grouped on the left, whereas the remaining ones belong to Cluster 1.

Our temporal analysis of infodemic risk patterns across countries yielded significant findings regarding how information environments evolved during the pandemic. Figure 6.2 illustrates the UMAP projection of countries based on their Infodemic Risk Index (IRI) time series, revealing two distinct clusters that transcend conventional socio-economic categorizations. The larger cluster encompasses several countries on the right, while the second cluster, comprising nine countries (Belgium, Chile, France, Haiti, Israel, Nigeria, Singapore, Turkey, and Venezuela), appears isolated on the left side of the projection.

To better understand these distinctive clustering patterns, we applied a *k-means* clustering to the temporal evolution of the Infodemic Risk Index (Figure 6.3). The analysis confirms two clearly differentiated infodemic risk profiles: Cluster 1 maintains a relatively stable risk profile throughout the observed period, whereas Cluster 2 exhibits a pronounced spike in risk during early 2021, with z-scores fluctuating between -0.5 and 0.5. This temporal divergence suggests that certain combinations of socioeconomic factors may create vulnerability to specific types of infodemic events, potentially related to vaccine rollout controversies that emerged during this period. The heterogeneity of

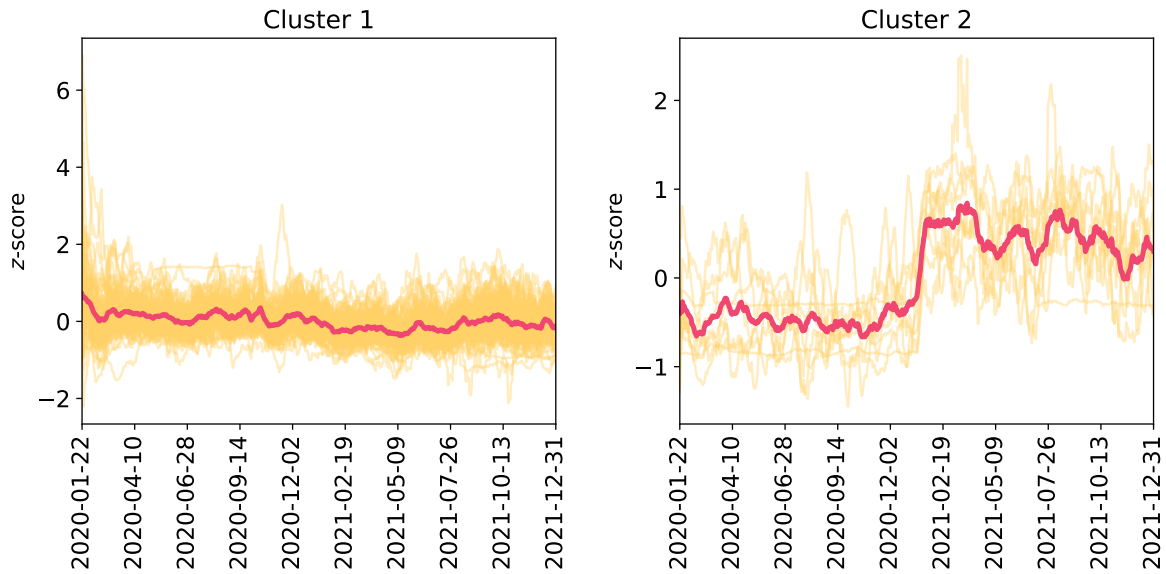


Figure 6.3: **k-means clustering with Euclidean distance and UMAP preprocessing on the IRI time series for the 88 selected countries.** The UMAP parameters employed were n neighbors = 5, min dist = 0.1 and n components = 102. The cluster average is plotted in red. All the time series are smoothed using a 30-day rolling average.

countries within Cluster 1, spanning diverse regions, development levels, and political systems, indicates that infodemic vulnerability follows complex patterns that are not reducible to single socioeconomic dimensions.

To investigate how socioeconomic factors relate to infodemic dynamics over time, we extracted the first principal component of our socioeconomic indicators and analyzed its correlation with three infodemic metrics (IRI, Dynamic Risk Index and Unreliability Index) over the course of the pandemic (Figure 6.4). This principal component represents the linear combination of socioeconomic indicators that captures the maximum variance in our dataset, serving as a composite proxy for socioeconomic development.

The first panel of Figure 6.4 reveals distinct temporal patterns in these correlations. During the initial phase of the pandemic (2020), the Dynamic Risk Index and the Unreliability Index demonstrate stronger correlations with socioeconomic factors, with coefficients reaching approximately 0.7. However, as the pandemic progresses into 2021, a marked shift occurs: these correlations diminish while the correlation between socioeconomic factors and the overall Infodemic Risk Index increases substantially, stabilizing around 0.5 through May 2022.

The correlation of Dynamic Infodemic Risk and Unreliability Index follows the same pattern and decreases with the beginning of 2021. The second panel, depicting p-values

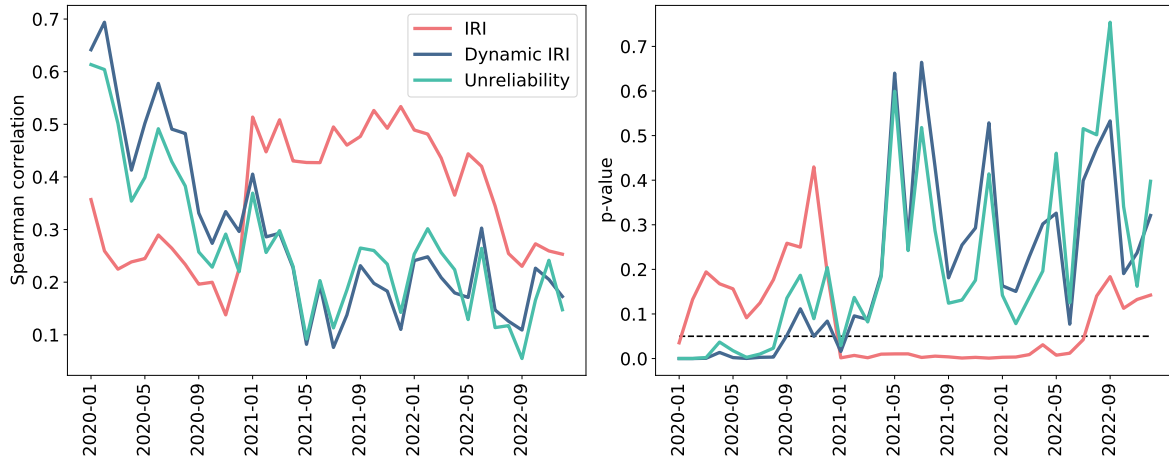


Figure 6.4: **Evolution of correlation between First Principal Component and Infodemic Indexes** The first panel shows the Spearman correlation of the first Principal Component and the monthly average of the infodemic metrics over time for OECD* countries. The second panel displays the p-value of the corresponding correlation with a black dashed line representing the significance level of 5%.

for these correlations, confirms the statistical significance of these temporal patterns. Two distinct phases emerge: an early phase characterized by significant correlations between socioeconomic factors and measures of information dynamics (Dynamic Risk) and content reliability (Unreliability Index), followed by a later phase where overall infodemic risk becomes more statistically significant and more strongly associated with socioeconomic conditions. This evolution suggests that different aspects of information environments become salient at different stages of a crisis, with initial responses dominated by dynamic information behaviors that later crystallize into more stable patterns of overall risk.

Interestingly, our analysis shows that the public conversation on COVID-19 evolved distinctly over the three-year pandemic period. As illustrated in Figure 6.2, countries clustered into two clearly differentiated groups based on their infodemic risk profiles. These clusters reflect substantive differences in how information environments responded to the pandemic across nations. Figure 6.3 further demonstrates that these differences manifested in distinct temporal patterns, with one cluster exhibiting marked fluctuations in Infodemic Risk Index values while the other maintained relatively stable information environments throughout the crisis period.

Based on these results, we decided to add another variable in order to explain the variability of the Infodemic Risk Index with respect to the socioeconomic indicators. Figure 6.5 shows the correlation between the two-dimensional space of the socioeco-

nomic indicators and the news media diet of each country analyzed.

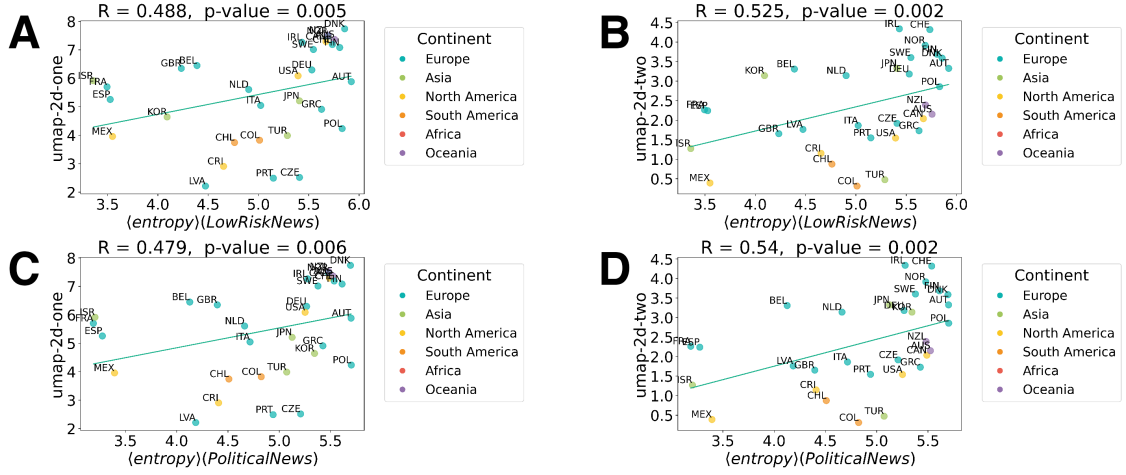


Figure 6.5: **Correlation between the news media diet and the UMAP of the socioeconomic.** Panels A and B show the correlation between the socioeconomic indicators with the entropy of all the domains classified as low-risk (satire, clickbait and political), while Panels C and D represent the correlation between the socioeconomic indices and the news classified as political.

A critical dimension of our analysis concerns the relationship between news media diet diversity and socioeconomic factors. Figure 6.5 illustrates the correlation between Shannon entropy of shared web domains (a measure of news source diversity) and the two-dimensional UMAP projection of socioeconomic indicators. Panels A and B report a robust positive correlation between the diversity of low-risk news consumption (encompassing satire, clickbait, and political content) and both dimensions of the socioeconomic space, with Spearman coefficients of 0.48 for the first dimension and 0.52 for the second, both statistically significant.

This relationship strengthens further when focusing specifically on political news diversity (Panels C and D), particularly with respect to the second dimension. These findings suggest that countries positioned in socioeconomic contexts characterized by stronger democratic institutions, higher education levels, and greater economic development tend to exhibit more diverse news consumption patterns. The pronounced correlation with the second dimension indicates that institutional quality and governance factors may be particularly influential in fostering media pluralism. The stronger correlation observed for political news diversity suggests that the pattern of engagement with political information may be especially sensitive to underlying socioeconomic conditions. This reinforces the idea that open, pluralistic societies with well-functioning democratic structures encourage a broader and more varied media landscape, whereas

6.3. Results

lower entropy may be indicative of more concentrated or polarized information environments.

Our final analysis examined the relationship between infodemic volatility, measured as the standard deviation of the Infodemic Risk Index, and socioeconomic factors, namely the two-dimensional socio-economic space derived from UMAP (Figure 6.6). This approach allowed us to investigate whether countries with different socioeconomic profiles exhibited distinct patterns of stability or fluctuation in their exposure to misinformation over the course of the pandemic.

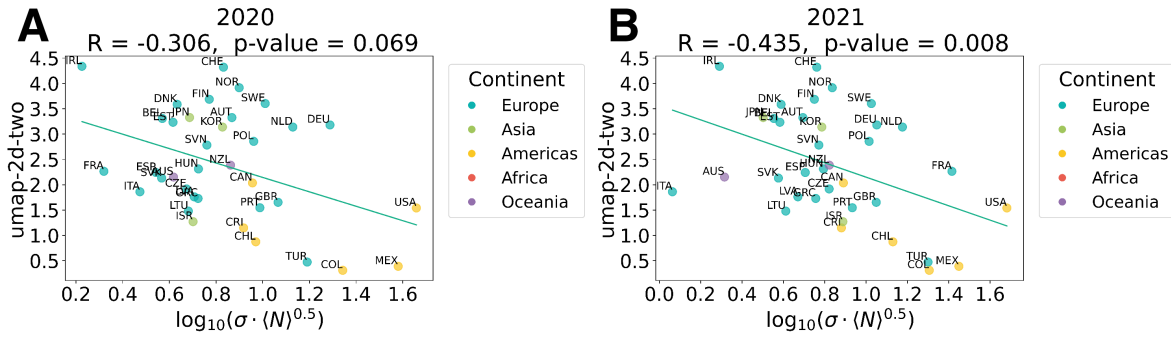


Figure 6.6: **Correlation between the second dimension of UMAP and the volatility of Infodemic Risk Index.** Panel A shows the correlation between the volatility of the Infodemic Risk Index in 2020, while Panel B shows the same relations in 2021.

We observed a pronounced negative correlation between volatility and the second dimension of the UMAP socioeconomic projection, which was consistent across both 2020 and 2021. In 2020, this correlation amounted to -0.304 with a marginally significant p-value of 0.069, while in 2021 it strengthened to -0.435 with a highly significant p-value of 0.008. This intensification of the relationship over time suggests that as the pandemic progressed, the stabilizing effect of robust institutional frameworks on information environments became more pronounced.

These results indicate that countries positioned higher along the second socioeconomic dimension, which captures aspects of institutional quality, democratic function, and media environment, exhibited significantly more stable information landscapes during the crisis. Conversely, countries lower on this dimension experienced greater fluctuations in their infodemic risk, suggesting that institutional strength may serve as a buffer against volatile shifts in information quality during prolonged crises. This finding highlights the importance of institutional frameworks not only in reducing overall infodemic risk but also in maintaining consistent, reliable information environments that

can withstand external pressures and evolving crisis conditions.

6.4 Discussion

Our analysis reveals complex, multidimensional relationships between socioeconomic factors and infodemic dynamics during the COVID-19 pandemic. By examining these relationships through several complementary analytical approaches, we have uncovered patterns that provide significant insights into how societal characteristics shape information environments during global crises. In this section, we interpret our findings in detail, addressing each of our five research questions while connecting our results to existing literature and drawing out theoretical implications and practical applications for policy interventions.

6.4.1 Socioeconomic dimensions and country clustering

Addressing our first research question about the underlying dimensions in socioeconomic indicators and country clustering patterns, the UMAP visualization presented in Figure 1 reveals a striking dimensional structure in our socioeconomic data. The arrangement of countries (left panel) shows clear geographic and cultural clustering patterns, with Scandinavian and Central European nations positioned in the upper right, Anglosphere countries to the right, and several South American countries in the bottom left. This clustering reflects deeper similarities in social, political, and economic structures that transcend simple geographic proximity, supporting theories of path-dependent institutional development [261] [262] and varieties of capitalism [263].

More revealing is the arrangement of socioeconomic indicators (right panel) along a curved distribution. The positioning of welfare indicators and median age at the top, followed by democracy, education, and economic development indicators, with social and political polarization measures distinctly separated in the bottom left, suggests a multidimensional structure of social development. This structure aligns with Welzel and Inglehart’s [264] theory of value change, which posits that socioeconomic development follows patterns that move from survival values to self-expression values, with corresponding shifts in institutional arrangements.

The separation of polarization indicators from other socioeconomic measures is particularly noteworthy. It suggests that polarization represents a distinct dimension of social organization that can vary independently of traditional development indicators, a finding that resonates with recent research showing how even advanced democra-

cies can experience significant polarization under certain conditions [265] [266]. This dimensional independence helps explain why infodemic risk does not correlate straightforwardly with conventional development metrics, as nations with similar levels of economic development or democratic institutions can nevertheless experience very different levels of social and political division.

The effectiveness of UMAP in capturing these complex relationships demonstrates the value of dimensionality reduction techniques for understanding multifaceted socioeconomic data. Rather than analyzing individual indicators in isolation, this approach reveals underlying structures that better reflect the interconnected nature of social, political, and economic development.

6.4.2 Temporal infodemic patterns and country clusters

In response to our second research question regarding distinct country clusters based on temporal patterns of infodemic risk, Figure 2 presents perhaps our most intriguing finding: the emergence of two distinct clusters of countries based on their infodemic temporal patterns. The smaller cluster of nine countries (Belgium, Chile, France, Haiti, Israel, Nigeria, Singapore, Turkey, and Venezuela) displays markedly different infodemic dynamics compared to the larger group. This clustering cuts across traditional categories of economic development, geographic region, and political systems, suggesting that infodemic vulnerability follows patterns not easily reducible to conventional development metrics.

The temporal analysis in Figure 6.3 further illuminates these differences, showing that while Cluster 1 countries maintained relatively stable infodemic risk levels, Cluster 2 countries experienced a dramatic spike in early 2021. This timing coincides with the global rollout of COVID-19 vaccines, suggesting that vaccine-related controversies may have triggered distinct infodemic patterns in these countries. In particular, vaccine misinformation can significantly reduce vaccination intentions, particularly in certain national contexts [267] [268].

The heterogeneity of countries within Cluster 1 is remarkable, including nations from different continents, political systems, and development levels. What might unite these diverse countries in their infodemic vulnerability? One possibility is that they share certain institutional features or social dynamics not captured by individual socioeconomic indicators but revealed through their information behavior patterns. For instance, they might share particular configurations of media systems, political discourse, or trust relationships that create similar vulnerabilities to infodemic spikes despite their differences

in other aspects.

This finding challenges simplistic narratives about information vulnerability being primarily a function of economic development or democratic institutions. Instead, it suggests that infodemic resilience emerges from complex interactions between multiple societal factors, creating distinctive national patterns that transcend conventional categorizations. Significantly, there is a growing recognition in the literature that information disorders are shaped by context-specific institutional arrangements and historical trajectories [269].

6.4.3 Evolving correlations between socioeconomic factors and infodemic measures

Examining our third research question about how socioeconomic factors correlate with different measures of infodemic risk over time, Figure 6.4 reveals a fascinating temporal evolution in these relationships. The first panel shows that different infodemic measures correlate with socioeconomic indicators in distinct patterns that change over time. During the pandemic's first phase (2020), the Dynamic Risk Index and Unreliability Index show stronger correlations with the first principal component of socioeconomic indicators, while in the second phase (2021 onwards), the correlation shifts to the Infodemic Risk Index.

This temporal shift suggests that different aspects of information environments become salient at different stages of a crisis. In the early, acute phase of the pandemic, socioeconomic factors appear more strongly related to dynamic measures of information engagement and sharing, perhaps reflecting how societies initially respond to novel threats with varying levels of information seeking and sharing behaviors. As the pandemic became a chronic condition, socioeconomic factors increasingly correlated with overall infodemic risk, suggesting that deeper structural characteristics of societies became more determinative of information quality as the crisis progressed.

The second panel of Figure 6.4, showing the statistical significance of these correlations, reinforces this interpretation. The p-values for Dynamic Risk Index and Unreliability Index are consistently significant in the pandemic's first year, while the Infodemic Risk Index becomes significant in the second year. This provides some statistical support for the changing relationship between socioeconomic factors and different aspects of information environments over time.

This temporal dimension has been overlooked in previous research on misinformation dynamics, which has tended to analyze phenomena at single time points rather than

as evolving processes. Our findings suggest that understanding infodemics requires an attention to these temporal patterns, in particular about how information environments evolve through different phases of a crisis, and how their relationship with underlying societal factors changes over time.

6.4.4 The protective role of media diet diversity

Turning to our fourth research question regarding the role of news media diet diversity in mediating the relationship between socioeconomic factors and infodemic vulnerability, Figure 6.5 presents evidence for the protective effect of information pluralism. The positive correlations between the Shannon entropy of shared web domains and both dimensions of the UMAP socioeconomic space (panels A and B) indicate that countries with more diverse information ecosystems tend to have stronger democratic institutions, higher education levels, and greater economic development. This relationship is even stronger when focusing specifically on political news (panels C and D), suggesting that exposure to diverse political viewpoints may be particularly important for information resilience.

These findings support and extend previous work suggesting that media diversity contributes to democratic health and information quality. At the individual level, higher news literacy tends to be associated to richer media diets [270] and improves the identification of fake news [271]. At the societal level, media legitimacy and trust emerge as key factors in national resilience against misinformation [272].

Our results therefore provide some support to the “information pluralism” hypothesis, namely the idea that more diverse information ecosystems create conditions where misinformation is less likely to dominate the public discourse. When citizens are exposed to a variety of sources and perspectives, they may develop greater critical evaluation skills and become less susceptible to misleading narratives from any single source. This challenges “echo chamber” theories that suggest that greater information choice necessarily leads to selective exposure and polarization [273]. Instead, our findings are more consistent with research showing that most citizens’ media diets are actually more diverse than commonly assumed [274], and that this diversity can serve protective functions and favors moderation.

The stronger correlation observed for political news diversity is particularly interesting. It suggests that exposure to varied political viewpoints, rather than a narrow ideological information diet, may be especially important for building societal resilience to misinformation. This emphasizes the value of cross-cutting exposure in democratic

discourse [275] and contradicts concerns that political diversity necessarily exacerbates polarization. Rather, it is polarization that poses a threat to democracy by reducing interest and information diversity [276].

6.4.5 Infodemic volatility and institutional stability

Addressing our fifth research question on how infodemic volatility relates to socioeconomic factors and whether this relationship evolves over time, Figure 6.6 reveals an important negative correlation between infodemic volatility (standard deviation of the Infodemic Risk Index) and the second dimension of our UMAP socioeconomic space. Countries positioned higher on this dimension, which captures aspects of institutional strength and media environment quality, demonstrate more stable information landscapes even amid the uncertainty of a global pandemic. This stability became more pronounced in 2021 (panel B, $r = -0.435$) compared to 2020 (panel A, $r = -0.304$), suggesting that institutional frameworks tend to become increasingly important for maintaining information integrity as crises evolve.

Institutional theories of resilience that emphasize how formal structures provide stability during disruptions [277]. In the context of information disorders, robust institutions appear to help societies maintain more consistent information environments, reducing the dramatic fluctuations in reliability that characterized high-risk countries. The strengthening of this relationship over time further suggests that institutional factors become more determinative of information quality as crises progress from acute emergency to chronic challenge.

This finding also contributes to debates about what constitutes information system resilience during crises. Rather than merely looking at average levels of misinformation, our analysis suggests that volatility, i.e. the tendency of information environments to fluctuate dramatically, represents another important dimension of infodemic vulnerability. Societies with greater institutional stability appear better able to maintain consistent information environments even amid uncertainty, avoiding the dramatic swings between reliable and unreliable information that can disorient public understanding and undermine trust.

6.4.6 Theoretical and policy implications

Our findings contribute to an emerging theoretical framework that views infodemics as a complex socio-technical phenomenon, deeply embedded in broader social structures. Rather than treating misinformation as primarily a problem of individual psychology or

platform design, this perspective emphasizes how historical institutional developments, social trust, educational systems, and political cultures create conditions that either foster or inhibit information disorders during crises.

The multidimensional nature of socioeconomic factors revealed in our UMAP analysis suggests that simple models of information vulnerability based on single indicators (such as education levels or economic development) are insufficient. Instead, countries occupy positions in a complex socioeconomic space with multiple dimensions that interact to create distinctive patterns of vulnerability or resilience. The emergence of distinct clusters based on temporal infodemic patterns, cutting across traditional development categories, suggests that information vulnerability follows its own logic, not entirely independent from other socioeconomic factors, but not reducible to them either. This challenges both technological determinism (the idea that digital media platforms inherently produce certain information outcomes regardless of context) and simple socioeconomic determinism (the notion that information quality mechanically reflects development levels).

The changing correlations between socioeconomic factors and different infodemic measures over time point toward a dynamic model of information disorders, where relationships between structural conditions and information environments evolve through different phases of a crisis. This temporal dimension has been undertheorized in most existing frameworks, which tend to present static models of misinformation vulnerability.

Finally, the strong relationship between news media diet diversity and positions in socioeconomic space suggests a possible causal pathway, that needs to be further investigated in future research, through which broader societal characteristics influence information environments. Societies with certain institutional arrangements, educational systems, and political cultures may foster more diverse information ecosystems, which in turn provide greater resilience against infodemics during crises.

Our findings have significant implications for policy approaches to managing infodemics and building societal resilience against information disorders. We can identify several key areas where policy interventions might be most effective, based on our empirical results.

As to media diversity and pluralism, the strong correlation between news media diet diversity and reduced infodemic risk suggests that policies promoting media pluralism could enhance societal resilience against misinformation. Policy approaches should support diverse media ecosystems through funding for public service media, subsidies for local news, and anti-monopoly regulations in media markets. Additionally, developing

media literacy programs that encourage citizens to consume information from multiple sources would strengthen individual information evaluation skills. Creating incentives for cross-cutting exposure to diverse political viewpoints and implementing platform design features that promote exposure to diverse content could further enhance information ecosystem health. The particularly strong relationship between political news diversity and socioeconomic indicators associated with lower infodemic risk suggests that ensuring pluralism in political information sources may be especially important. This challenges approaches focused primarily on removing “harmful” content, suggesting instead that ensuring a healthy diversity of perspectives may be more effective for long-term resilience.

Concerning the strengthening of the institutional framework, the negative correlation between infodemic volatility and the institutional dimension of our socioeconomic space indicates that stronger institutions help stabilize information environments during crises. This suggests policies aimed at building trust in key epistemic institutions such as science, journalism, and public health authorities. Supporting independent fact-checking organizations and quality journalism provides additional safeguards against the spreading of misinformation. Designing institutional arrangements that aim at reducing political polarization can mitigate one of the key drivers of information disorder. Furthermore, creating rapid response systems that can provide authoritative information during emerging crises enables a more timely correction of misinformation before it gains traction. The strengthening of this relationship over time suggests that institutional factors become increasingly important as crises progress from acute emergency to chronic challenge. Institutional development inevitably pairs with a long-term investment in information resilience rather than with the search for quick fixes to immediate information disorders. The distinct separation of polarization indicators from other socioeconomic measures in our UMAP analysis suggests that this dimension requires specific policy attention. Countries may have well-developed economies, strong democratic institutions, and high education levels, yet still experience high levels of polarization that create vulnerability to infodemics. This makes a case for policies focused on reducing affective polarization through interventions that humanize political outgroups. Additionally, developing political institutions that incentivize cooperation rather than zero-sum competition could mitigate some of the structural drivers of polarization. Creating shared information spaces that transcend partisan divides would provide common ground for diverse groups to engage with the same factual basis. Supporting trust-building initiatives across political and social boundaries further addresses the social-psychological dimensions of polarization. The distinct clustering

of polarization indicators separate from traditional development metrics suggests that these interventions may be necessary even in highly developed societies with otherwise strong institutions. The changing correlations between socioeconomic factors and different infodemic measures over time suggest that effective interventions may need to adapt to different phases of a crisis. During early, acute phases, when dynamic information sharing behaviors are more strongly correlated with socioeconomic factors, interventions should focus on the rapid dissemination of accurate information through trusted channels. Supporting accurate information sharing behaviors among influential nodes in social networks becomes crucial in these initial stages. Providing clear guidance to reduce uncertainty can prevent information vacuums that misinformation often fills. As crises become chronic, and overall infodemic risk becomes more strongly correlated with socioeconomic factors, interventions might shift toward addressing deeper structural factors that shape information environments. Building long-term resilience through educational initiatives enables preparation for future information challenges. This temporal dimension suggests that infodemic management requires adaptive response strategies rather than one-size-fits-all approaches.

6.4.7 Limitations and future directions

Our results are subject to several limitations. First of all, our reliance on Twitter data implies that our analysis captures only a portion of the total information ecosystem, with known demographic biases toward younger, more educated, and more politically engaged populations [278]. Considering more data sources, including other social media platforms, traditional media, and offline communication channels, is essential to provide a more comprehensive view of information flows across different demographic segments.

Moreover, the cross-sectional nature of our socioeconomic indicators limits causal inferences about their relationship with infodemic patterns. While we observe correlations between socioeconomic factors and infodemic measures, we cannot establish causal pathways, and our preliminary results should be seen mainly as a useful source of testable predictions for future longitudinal studies tracking changes in both socioeconomic factors and information environment characteristics over time.

Additionally, our focus on country-level analysis necessarily obscures within-country variations that may be substantial, particularly in larger or more diverse nations. Regional disparities in education, economic development, political attitudes, and media access likely create uneven information landscapes within countries. A closer focus on how regional differences in socioeconomic factors within countries relate to more

localized patterns of infodemic spread is therefore needed.

Our analysis includes a comprehensive but not exhaustive set of socioeconomic indicators. Some potentially relevant factors, such as historical experiences with epidemics, cultural attitudes toward authority, or specific features of healthcare systems, were not included in our dataset. Enriching the panel of socioeconomic indicators with additional relevant dimensions that might influence infodemic vulnerability would be another major step forward.

Finally, while our study identifies correlations between socioeconomic factors and infodemic patterns, an understanding of the precise mechanisms through which these factors influence information environments is still largely uncharted territory. Mixed-methods approaches combining quantitative analyses with qualitative case studies could be useful for hypothesis development in this respect. Comparative case studies of countries with similar socioeconomic profiles but different infodemic outcomes would also be particularly valuable for identifying specific institutional arrangements or cultural factors that enhance infodemic resilience.

6.5 Conclusion

Our study demonstrates that the vulnerability of societies to infodemics during the COVID-19 pandemic was significantly shaped by underlying socioeconomic factors, with patterns that evolved over time as the crisis progressed. The arrangement of socioeconomic indicators in a multidimensional space, the clustering of countries based on temporal infodemic patterns, the changing correlations between socioeconomic factors and infodemic measures over time, and the relationship between news media diet diversity and infodemic risk all point toward a complex, context-dependent model of information vulnerability during crises.

These findings challenge simplistic narratives about information disorders seen as primarily technological problems related to digital platform design and functioning, or straightforward reflections of the levels of economic development. Instead, they suggest that infodemics emerge from complex interactions between technological systems, social structures, institutional arrangements, and cultural factors, creating distinctive patterns of vulnerability and resilience across different societies.

The diversity of media diets, particularly regarding political information, emerged as an especially important protective factor against information disorder, highlighting the value of pluralistic information ecosystems for societal resilience. The negative correlation between institutional stability and infodemic volatility further emphasizes

the importance of robust institutions for maintaining consistent, reliable information environments during crises.

These insights underscore the need for more comprehensive approaches to infodemic management that address not only immediate manifestations of misinformation but also the deeper structural factors that create favorable conditions for information disorders. Building societal resilience against future infodemics will require coordinated efforts across educational, institutional, and media domains, guided by a nuanced understanding of how socioeconomic factors shape information environments during crises.

Chapter 7

A Stochastic Model for Understanding Volatility in Infodemic Risk Index

7.1 Introduction

While the previous chapter 6 focused on the validation of infodemic risk indexes using socio-economic indicators, the work proposed here aims at understanding the instability of Infodemic Risk Index (IRI) and contributes to the following points:

- we introduce volatility as a complementary dimension to the Infodemic Risk Index, moving beyond average exposure to capture fluctuations in misinformation risk across countries and over time
- we empirically show that IRI volatility cannot be explained solely by random sampling, revealing the role of structural and behavioral factors, including super-spreaders and engagement dynamics.
- we propose a stochastic Monte Carlo model that integrates follower distribution, disinformation ratio, engagement scaling, and superspreader behavior to reproduce empirical patterns of IRI instability.
- we provide insights into the socio-technical mechanisms driving infodemic dynamics, highlighting the disproportionate influence of highly connected users and potential avenues for targeted interventions to stabilize information environments.

This index provides a measure of a country's exposure to misinformation by aggregating the number of followers reached through users who share unreliable content

online. In this context, the IRI represents the average daily level of risk that a given country faces in terms of circulating false or misleading information. Importantly, the index incorporates not only the frequency of misinformation but also its potential reach, as a proxy of the audience size (i.e., follower counts). While the IRI offers a useful first step toward quantifying infodemic risks across countries, it also presents only one insight about the infodemic risk. By focusing primarily on average values, it does not consider the variability and volatility of exposure. Beyond the average value, we can also take into account the higher-order moments of the distribution, such as the standard deviation, which provide insight into how strongly the risk fluctuates across time and across countries. In this context, our analysis aims to investigate the temporal evolution of the IRI, assessing both its trends and its stability over different periods. Indeed, some countries such as Venezuela exhibit a remarkable volatility of the Infodemic Risk Index, while other countries such as Italy show a stable evolution of such index. In this context, we propose a stochastic model to understand the key features that produce these different outcomes in different countries. This serves not only as a way to reproduce the dynamical aspects of the infodemic risks but also to provide insights about key aspects of the online communication networks that might impact the stability of the index itself.

7.2 Methods

7.2.1 Overview of the data

Our analysis relies on data collected through the COVID-19 Infodemic Platform [211], as described in the chapters 2 and 6. This dataset encompasses messages containing URLs, each evaluated for reliability through comparison with various publicly accessible databases covering scientific and journalistic sources. We specifically employed reliability classifications from MediaBiasFactCheck [74], an organization that maintains an extensive and regularly updated database evaluating the ideological leanings and factual accuracy of media outlets through both quantitative metrics and qualitative assessments. The URLs in our dataset were classified into seven distinct categories: reliable sources (Mainstream Media, Scientific Journals), unreliable sources (Satire, Clickbait, Political, Fake/Hoax, Conspiracy/Junk Science), and unclassified sources. In this specific study, we analyzed only 51 countries different in terms of average number of daily tweets. We specifically selected those countries because our aim is to be able to compare countries with different size in terms of volume.

7.2.2 Infodemic Risk Index

As recalled in the chapter 2, the Infodemic Risk Index is a measure to quantify the exposure to unreliable contents, weighted by the number of followers that have been reached by these news. Conversely, the exposure to reliable news (E_r) is given by aggregating the number of followers potentially reached by tweets containing reliable sources. The IRI is computed as

$$IRI = \frac{E_U}{E_U + E_R}$$

7.2.3 Volatility

The concept of volatility has been defined mainly in economics and finance as the degree of variation of a trading price series over time, usually measured by the standard deviation of logarithmic returns. In this context, we employ the standard deviation of IRI as a measure of the instability of the risks over time.

7.2.4 Monte Carlo Simulation

To better understand the dynamics underlying the Infodemic Risk Index (IRI) and the role of different parameters in shaping its volatility, we implement a Monte Carlo simulation framework. This stochastic approach has been widely adopted in the study of complex social and economic systems to capture uncertainty and variability in diffusion processes [165].

Our simulation is designed with two primary goals. First, we want to reproduce the scaling value to test whether the observed empirical anti-correlation between tweet volume and volatility of the IRI can be reproduced by stochastic sampling effects. In theory, volatility should scale as $\sigma \approx N^{0.5}$ with the number of daily tweets, but our empirical estimates indicate a weaker dependence $\sigma \approx N^{0.24}$, indicating an unexpected dependence of volatility on tweet volume. Second, our goal is the characterization of average and variability to reproduce the empirical distributions of the IRI in different countries and assess how structural features of online communication networks impact both the average risk and its temporal fluctuations.

We simulate populations of N users ($N = 20, 40, 80, 160, 320, 640, 1280, 2560$) and run 1000 simulations for each setting. Each user can share either reliable or unreliable content, with a probability determined by a set of structural parameters:

- **Follower distribution (α):** The number of followers per user follows a power-law

distribution, consistent with empirical observations of social media platforms. The exponent α controls the degree of heterogeneity, with smaller values producing heavier tails and more dominant hubs.

- **Disinformation ratio:** A proxy for the media diet of each country, defined as the ratio of unique unreliable sources over the total number of unique sources. This parameter captures cross-country differences in exposure to unreliable information.
- **Engagement scaling:** Based on the empirical correlation between user reach and engagement, we assume that the probability of amplification (retweets, replies) scales with the number of followers.
- **Superspreader activity:** To account for the disproportionate role of highly connected accounts, we estimate the probability of unreliable sharing among the top 5% of users (by follower count). This reflects the empirical observation that a small elite of accounts drives the majority of misinformation diffusion.

7.3 Results

The average value of Infodemic Risk Index can vary from country to country. In the Figure 7.1, four countries are analyzed with respect to the evolution of the Infodemic Risk Index, showing completely different patterns of such index. While Peru seems to be pretty stable with an average higher value of IRI, Venezuela is a case example in which both the average and the variability of the index are pretty high. Different is the case of Italy that has a stable and yet lower value of IRI, France seems to have two different phases: one characterized by lower and stable value, and another one subsequent to the vaccination against Covid-19 characterized by having higher and more unstable values.

As introduced above, we suspect a relationship between standard deviation of IRI and the average value of daily tweets. Countries with a higher tendency of posting tweets on a daily basis are also countries recognized to be more stable due to the effect size of our data. This is clearly shown by Figure 7.2 when considering all the countries as well as when considering only the subsample of 51 countries we decided to focus on Figure 7.3.

The standard deviation of the Infodemic Risk Index (IRI) is expected to scale with the number of daily tweets N as

$$\sigma \sim N^\beta,$$

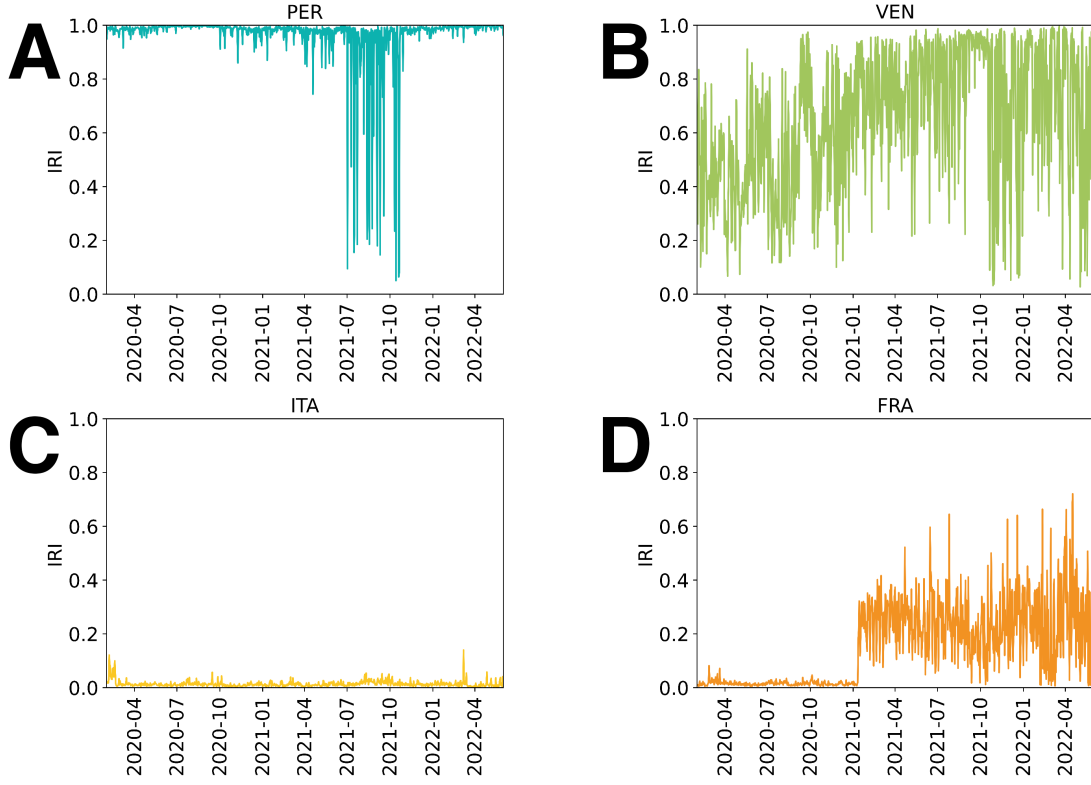


Figure 7.1: **Time Series of Infodemic Risk Index.** Evolution of the Infodemic Risk Index over time for 4 different countries: Perù, Venezuela, Italy and France.

where σ is the standard deviation of the IRI and N is the number of daily tweets. If fluctuations were driven solely by random sampling, variance would scale linearly with N , giving

$$\sigma \propto \sqrt{N} \quad \Rightarrow \quad \beta = 0.5,$$

as expected from the central limit theorem. However, our empirical analysis shows a weaker dependence:

$$\sigma \sim N^{0.24},$$

indicating that structural and behavioral factors, not just statistical variability, influence the IRI volatility.

This weaker dependence suggests that fluctuations in the IRI cannot be explained solely by random variability, but are instead influenced by additional structural and behavioral factors within each country's online communication ecosystem. In this context, our aim is to investigate the key mechanisms that lead to such different value of instability for several countries considered.

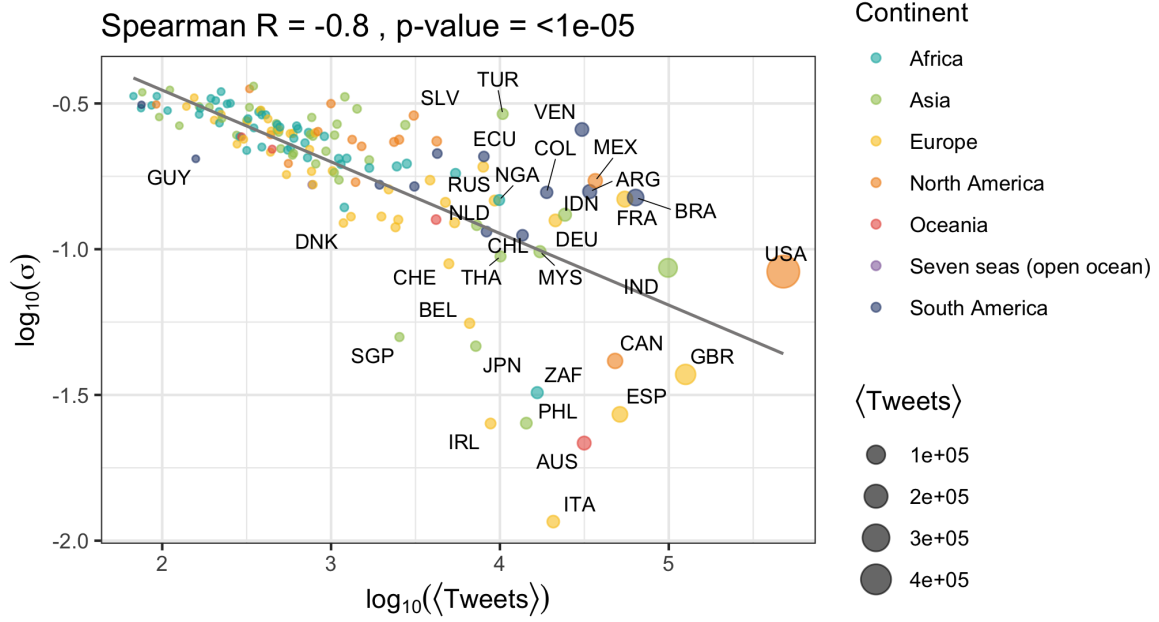


Figure 7.2: **Correlation between average number of tweets and the Volatility of Infodemic Risk Index.** We calculated the spearman correlation between the volatility, defined as standard deviation of IRI, and the average number of tweets, showing that countries with more tweets tend to be more stable in terms of Infodemic Risk Index, due to the effect size of the messages' volume.

7.3.1 Followers Distribution

One key element is the role of *superspreaders*. With this term, we refer to accounts with disproportionately large follower bases that are able to amplify unreliable content far beyond the average user. Because of their high reach, superspreaders can introduce large fluctuations in exposure and destabilize the otherwise expected scaling. We rank the distribution of followers for each country and we fit a power-law distribution with *powerlaw* package [279], extracting the scaling exponent α . Each user's number of followers is assumed to follow a power-law distribution:

$$P(k) \sim k^{-\alpha},$$

where α controls the heterogeneity of the network: smaller α values produce heavier tails, meaning a few users hold a disproportionately large number of followers. This analysis has been done for all the countries of interest, differentiating by users sharing reliable and unreliable content as last tweet. Figure 7.4 shows an average value of 2.1 for users sharing reliable content and 2.07 for users sharing unreliable content. However,

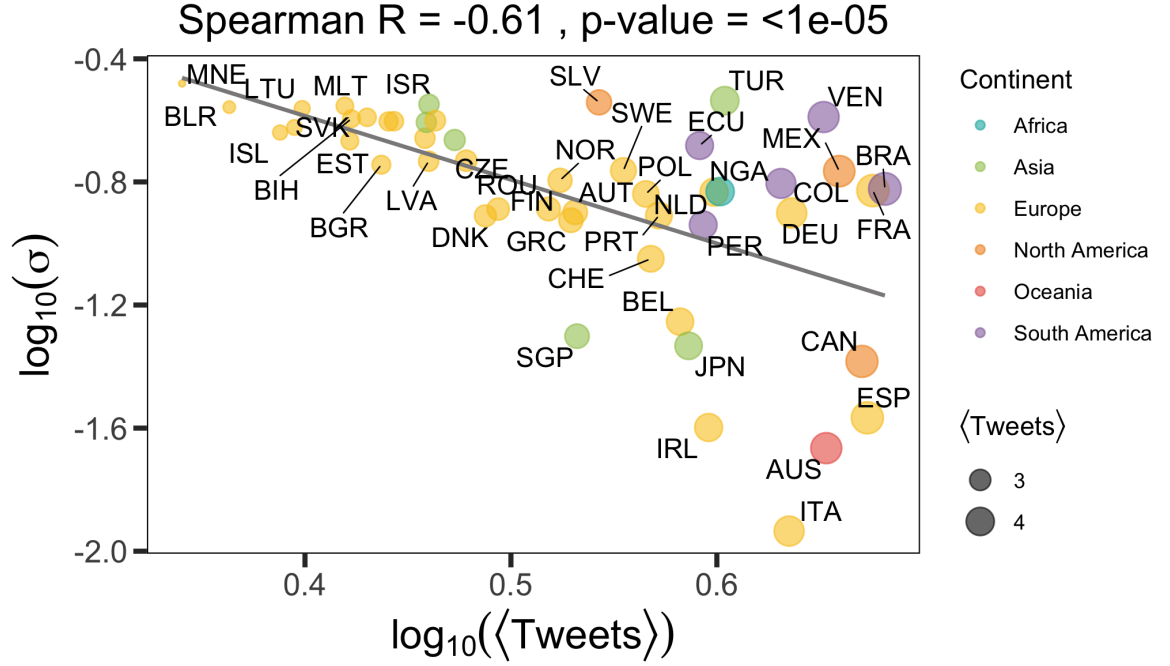


Figure 7.3: **Correlation between average number of tweets and the Volatility of Infodemic Risk Index only on our subset of data.** We calculated the spearman correlation between the volatility, defined as standard deviation of IRI, and the average number of tweets, showing that countries with more tweets tend to be more stable in terms of Infodemic Risk Index, due to the effect size of the messages' volume.

this element was not sufficient to explain the scaling value found, when incorporated in the model simulation, the follower distribution leads to a scaling exponent of 0.11, which diverges significantly from the empirical estimate of 0.24. This result suggests that additional mechanisms must be considered to account for the observed scaling behavior.

7.3.2 Disinformation Ratio

Another element that we took into consideration as a parameter of our model is the *disinformation ratio* as the ratio between the unique unreliable sources and the total unique reliable and unreliable news media outlets consumed in each country. This serves as a generalization of the concept of *news media diet*, discussed in the chapter 5. The average value of disinformation ratio is 0.4, so there is the 40 % of probability for a user to fall into an unreliable sources. Beside the type of news shared and the distribution of followers, we looked at the *engagement* of users based on their number of followers.

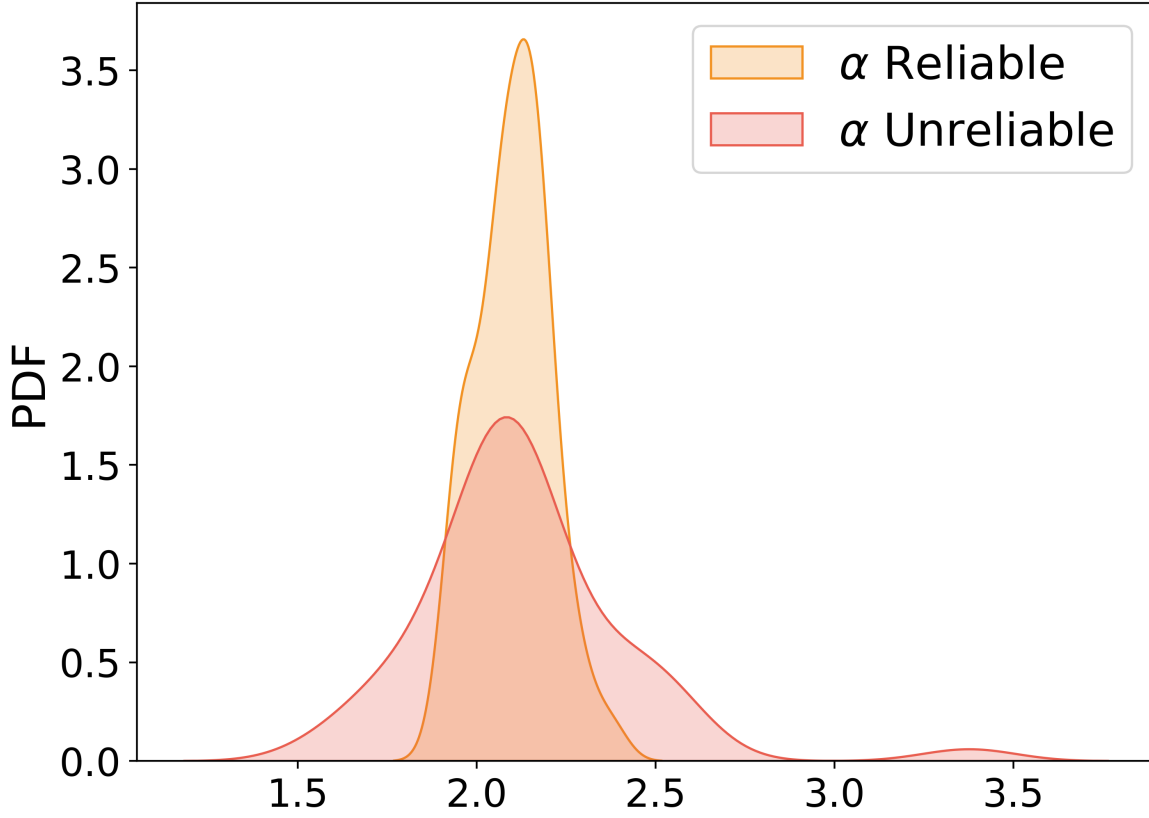


Figure 7.4: **Distribution of power-law alpha.** Distribution of the exponent value of powerlaw distribution for users sharing reliable content (2.11) and those sharing unreliable content.

7.3.3 Engagement

Figure 7.5 shows the relationship between number of followers and engagement received (calculated by summing the retweets and replies received by users). Interestingly, across all four countries examined (Peru, Venezuela, France, and Italy), we observe a consistent nonlinear pattern. For users with fewer than approximately 10^3 followers, engagement remains nearly constant and relatively low, suggesting that content produced by small or moderately connected users generates limited amplification regardless of their follower count.

7.3.4 Superspreaders posting behavior

However, beyond the threshold of roughly 10^3 followers (indicated by the dashed vertical line), engagement begins to scale super-linearly with follower count. This regime

shift indicates the emergence of a class of users whose messages propagate much more efficiently through the network. The fitted regression line beyond the threshold shows a steep positive slope, confirming that once users surpass a certain reach, their probability of generating significant amplification grows disproportionately. From a modeling perspective, this finding supports the inclusion of an “engagement scaling” parameter that depends on follower count, as introduced in our Monte Carlo simulations, modeling the engagement as a function of number of followers.

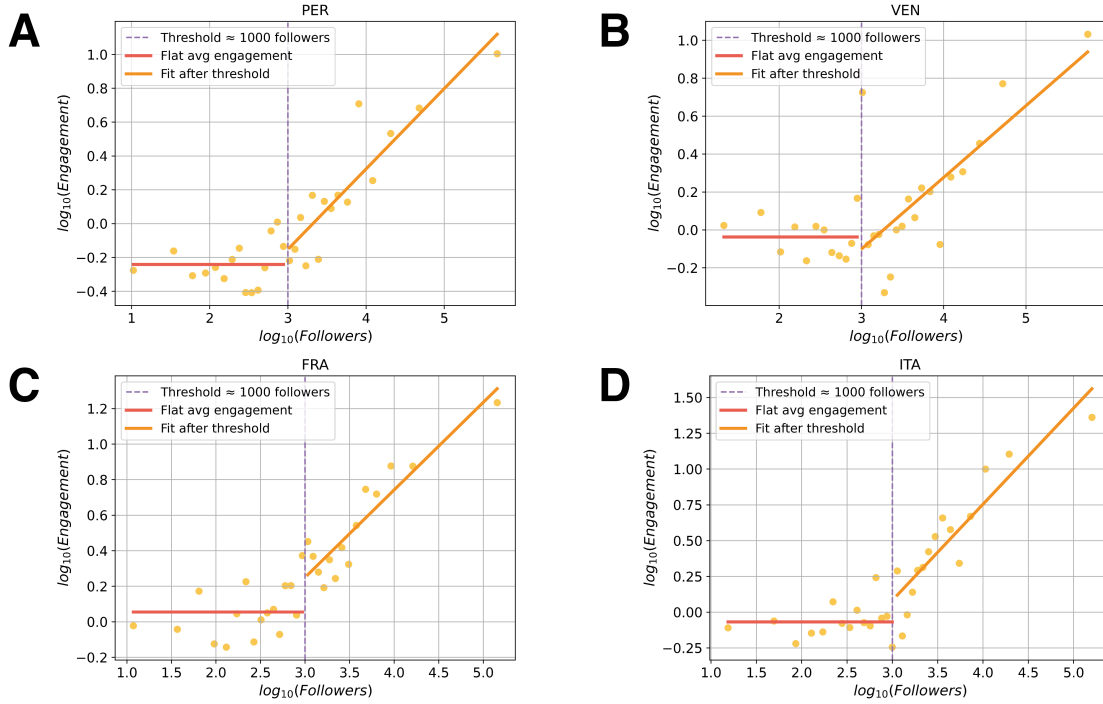


Figure 7.5: **Relationship between number of followers and engagement received.** We found for all the 4 countries (PER, VEN, ITA, FRA) a strong correlation between number of followers and engagement received (especially with users with more than 10^3 followers).

Beside the distribution of followers and the engagement that superspreaders might reach, it is interesting to observe the type of information that they share online. On average, we assume the so-called disinformation ratio. But how does it change with respect to the superspreaders? Indeed, countries might differ in how the most influential users share news online. To address this question, we extracted the top 5 % of users in terms of followers in order to quantify the percentage of unreliable sources shared by users with a high number of followers. Figure 7.6 shows the relationship between volatility and unreliable sources shared by superspreaders with a result of 0.48 spearman

correlation.

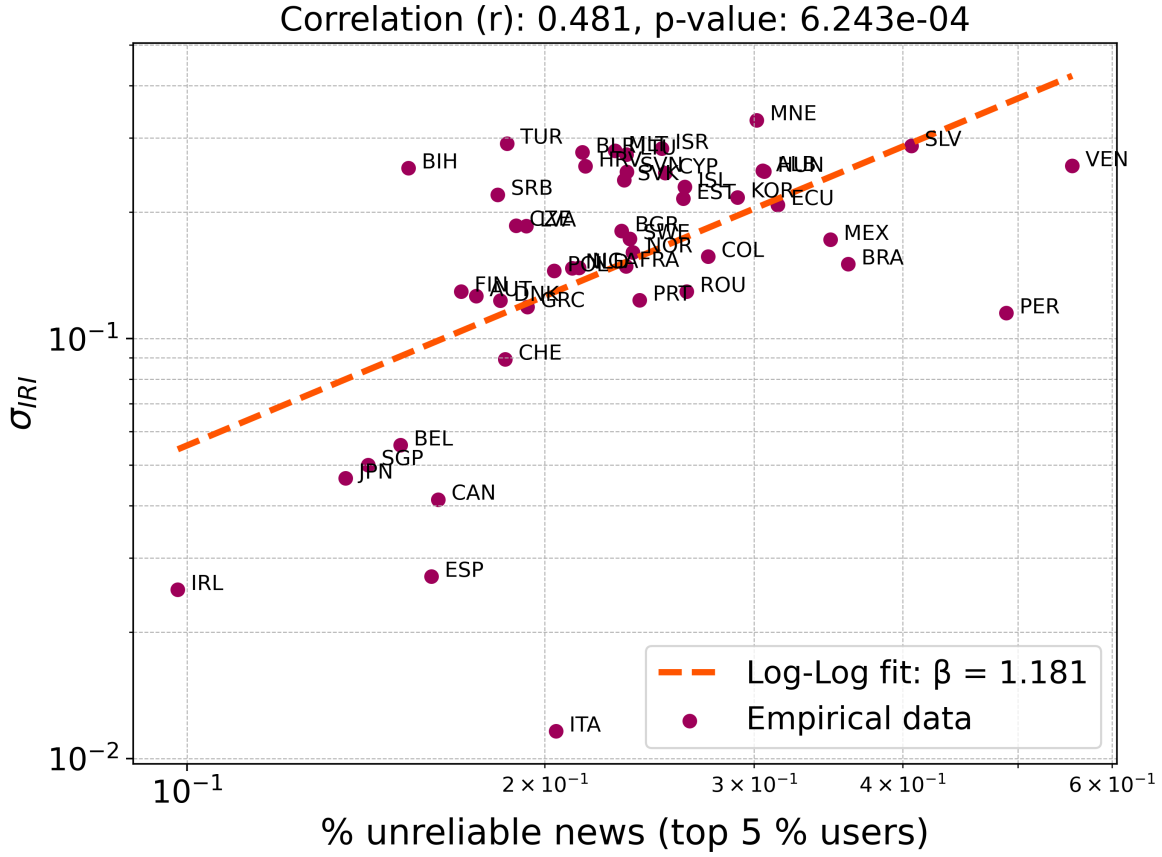


Figure 7.6: **Unreliability vs Top 5 % of superspreaders** Correlation between superspreaders users (defined as the top 5 % of users in terms of followers distribution) and the volatility of the countries. Countries with higher value of superspreaders sharing unreliable content are also the ones with higher instability with respect to the Infodemic Risk Index.

To map better the relationship between volatility and news consumption of superspreaders, Figure 7.7 highlights two different aspects of the mechanisms of the volatility: the Panel A shows the level of volatility in the countries analyzed, while the Panel B reveals the percentage of unreliable news shared by *superspreaders*. It is interesting to see how there is an overlap between countries with higher value of volatility as well as higher value of unreliable news sources shared by users with an influential role in the digital ecosystem.

Thus, the model simulation proposed has taken four parameters: the exponent of the followers distribution, the disinformation ratio, the engagement and the average ratio of unreliable content shared by most influential users. By controlling these parameters,

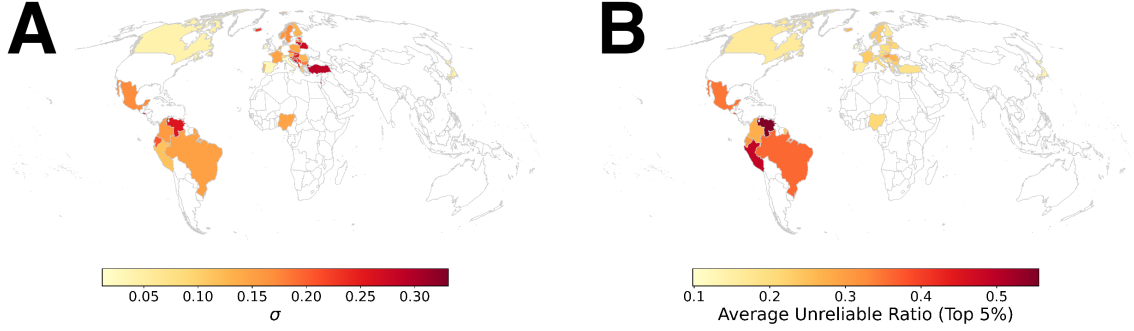


Figure 7.7: **Volatility and Superspreaders posting behavior.** Panel A maps the volatility of each country analyzed, while Panel B shows the average unreliable ratio of news posted by superspreaders, showing the relationship between instability of IRI and behavior of the most influential users in the digital ecosystem.

our model reproduces for a given set of N users a scaling exponent β of 0.16 with a range between 0.13 and 0.20, much closer to the one observed in our empirical data, as shown in Figure 7.8.

7.4 Discussion

The Infodemic Risk Index is a useful indicator that gives us an idea of the average value of being exposed to unreliable sources given the potential reach of users with a high value of followers. However, the variability of this indicator might be informative of another aspect of the same phenomena. How does this indicator change? How does it differ from country to country over time? We tried to answer to these questions in this chapter, trying to formalize key features of the socio-technical systems that allow us to reproduce the empirical scaling β exponent of 0.24. We found that superspreaders have an important role when it comes to have an impact on the stability of the index. First of all, they hold a huge number of followers that might be potentially reached by the information they share. The engagement that they produce is correlated with the number of followers they have, producing a bigger effect on the stability of the index itself. In addition, some countries seem to have superspreaders that share on average higher value of unreliable news, making those countries more likely to have higher variability in the Infodemic Risk Index. Our results demonstrate that the volatility of the IRI does not simply follow from random sampling effects. In principle, if fluctuations were driven solely by statistical variability, we would expect volatility to scale with

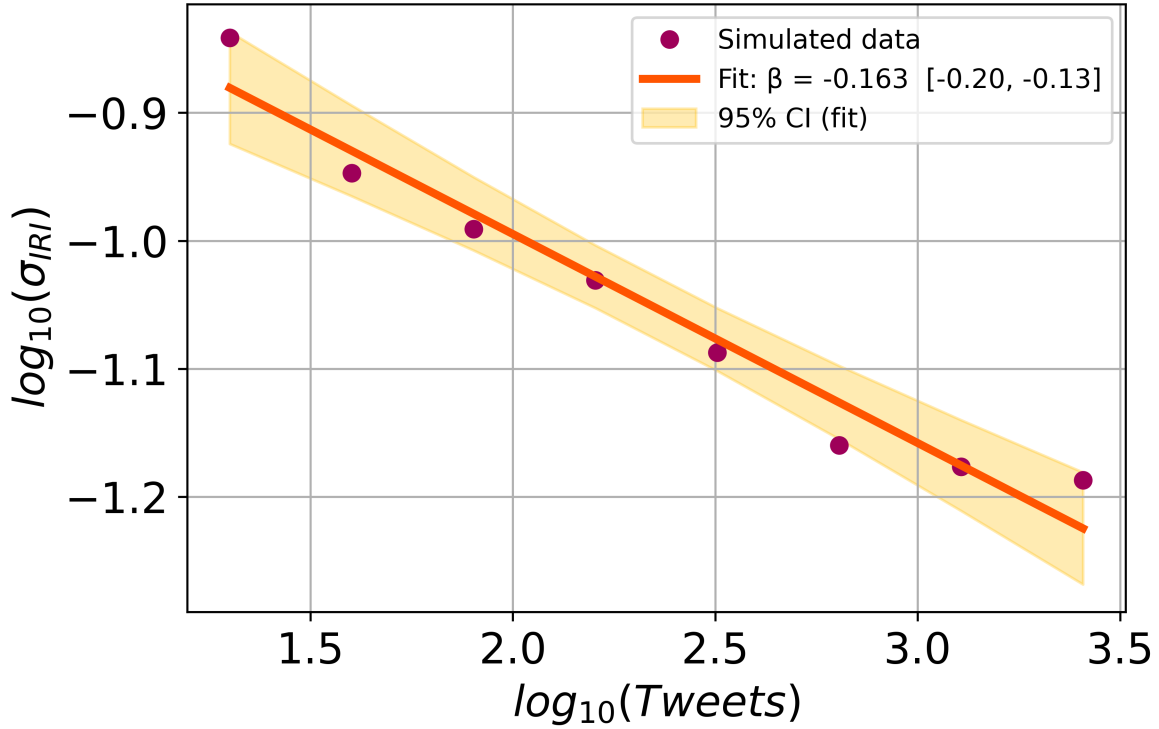


Figure 7.8: **Simulation of the empirical scaling β** We simulated a population of different users posting online, characterized by four parameters: disinformation ratio, exponent value of the followers distribution, engagement received and posting behaviour of superspreaders.

tweet volume according to 0.5, Instead, our empirical analysis reveals a much weaker dependence, with a scaling exponent of $\beta = 0.24$. This finding indicates that deeper structural and behavioral mechanisms are at play, shaping how misinformation spreads and stabilizes in national contexts.

This chapter contributes to the growing literature on infodemic dynamics in three key ways. First, it moves beyond average exposure metrics by highlighting volatility as a distinct and meaningful dimension of infodemic risk. Second, it shows that volatility cannot be explained by random fluctuations alone but must be understood in terms of socio-technical mechanisms embedded in digital platforms. Third, it proposes a stochastic modeling approach that captures how structural and behavioral factors interact to reproduce empirical patterns. From a theoretical standpoint, our findings underscore the importance of accounting for heterogeneity and behavioral asymmetries when modeling information diffusion. Simply scaling up random sampling processes misses the nonlinearities introduced by engagement thresholds and the disproportionate role of

superspreaders. From a practical standpoint, the results suggest that monitoring the posting behavior of highly connected accounts may be as important, if not more important, than monitoring aggregate volumes of misinformation. Policies or platform interventions that reduce the visibility of unreliable content shared by superspreaders could have outsized effects in stabilizing infodemic risks. Finally, our analysis opens several directions for future research. One is the incorporation of temporal shocks, such as elections or vaccination campaigns, which may alter both the disinformation ratio and the behavior of superspreaders in ways not fully captured by our model. Indeed, while we were able to reproduce the scaling behavior observed in our simulation, we are not fully capable of reproducing the average volatility for each country, demonstrating that we are still missing other parameters that might improve the results.

Chapter 8

The Radicalization Pipeline: Characterizing and Predicting online misogyny

This chapter opens a new line of research, where the focus is not on Twitter/ **X** but on Reddit, a social media or platform built on communities centered around specific topics. The main contributions of this chapter are the following:

- we trace the behavioral and linguistic trajectories of Reddit users before and after joining incel communities, revealing shifts toward higher toxicity, hate speech, and engagement with radicalized topics.
- we demonstrate that participation in certain subreddit categories, particularly “Social Issues”, increases sharply post-engagement, highlighting ideological drift and thematic reorientation as key mechanisms of online radicalization.
- we develop a machine learning framework that accurately predicts which users will join incel communities, showing that contextual interactions and incel-specific linguistic features are strong indicators of future radicalization.
- we provide evidence of social influence in radicalization, with future incel members exhibiting more overlapping participation in discussion threads, suggesting that networked exposure and reinforcement facilitate entry into extremist communities.

8.1 Introduction

As individuals navigate and participate in different subreddits, their interests and attitudes may evolve over time, potentially becoming more extreme. While many communities focus on relatively neutral or leisure-oriented topics such as video games, sports, or television series, others are centered around more contentious or politically charged topics, creating fertile ground for the reinforcement and escalation of radical opinions. One of the most debated subreddits is the *Incel* community (involuntary celibate). While the term “incel” originally emerged from a support forum created by a woman in 1997 [280], the corresponding subreddit on Reddit in August 2013. This group has its roots in the idea that looks have a fundamental role in finding a relationship and that men who do not conform to beauty standards are doomed to rejection and loneliness [281]. Young men perceive women as the privileged gender because they can be valued for their sexuality. On the other hand, they perceived themselves having no inherent value due to their perception of physical unattractiveness and lack of sexual success [282].

Scholars have described “being incel” in multiple dimensions: as a social identity (self-identifying as incel), as an ideology (adopting misogynistic beliefs consistent with incel views), and as group membership (belonging to incel communities) [283]. Over time, these online communities became increasingly radicalized, amplifying hateful discourse and normalizing violence against women [284]. In 2017, Reddit banned the largest incel subreddit, *r/incels*, citing repeated violations of its policies against harassment and hate speech. Yet, this intervention did not eliminate the movement: members migrated to other subreddits and to alternative platforms such as 4chan or dedicated incel forums, where moderation was weaker and extremist content could spread more freely [285]. The incel case highlights the role of online communities as incubators of radical ideologies and the challenges of platform moderation, which may inadvertently disperse users into less regulated spaces. Existing literature has primarily examined incel ideology, discourse, and representations of community members. However, an important gap remains: how do individuals become part of these communities? How does their online behavior change once they join? In this chapter, we address these questions by tracing the trajectories of users who joined the incel subreddit. We analyze their posting behavior and the linguistic features of their comments both before and after joining, and we investigate whether certain characteristics can predict a user’s transition into the incel community. To do so, we compare incel joiners with a sample of users they interacted with prior to posting or commenting in incel-related subreddits,

providing new insights into the mechanisms of radicalization and group entry online.

8.2 Methods

Overview of the Dataset

We collected Reddit data using the Pushshift API from the subreddit *Incels*. The original dataset includes 54,328 posts and 732,377 comments spanning the period from January 4, 2014, to September 30, 2017, period on which the community was banned. For our analysis, we focused on the most active users of the community. These are users who were active for more than four months, made at least one post, and contributed more than four comments. Using this subset, we traced their online behavior across other communities both before and after their first activity (post or comment) in the *Incels* subreddit. This resulted in a dataset comprising 27,200 posts and 519,001 comments from 195 unique users. To investigate differences between users who engaged in incel-related discussions and those who did not, we also extracted comments from conversations in which these prolific users were involved prior to joining the *Incels* subreddit. This other dataset consists of 661,852 comments across 2,135 threads, in different subreddits.

8.2.1 Classification Subreddits

To analyze the thematic structure of user participation across Reddit, we categorized each subreddit into broader topic domains. We employed a zero-shot classification approach using ChatGPT (GPT-3.5) to automate this task. In the initial phase, the model was prompted with only the name of the subreddit and asked to infer a suitable thematic label, without being given any predefined list of categories or labeled examples. After collecting these outputs, we identified the ten most frequently occurring or semantically consistent categories, which included categories such as "Entertainment", "Social Issues", "Gaming". In the second phase, we re-queried ChatGPT for each subreddit, this time instructing it to classify subreddits into one of these ten already defined categories.

8.2.2 Toxicity and Hate Speech

The toxicity score is from the Perspective API [286], a machine learning model that assigns a piece of text a score between 0 and 1, where 1 is considered most toxic. The

API defines “toxicity” as “a very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective”. Beyond the toxicity, Google Perspective API offers other level of toxicity, including severe toxicity, identity attack, insult, profanity, threat, sexually explicit. These scores were computed for each comment and post in our dataset and used as both individual and contextual features in our analysis. With individual features, we refer to the score of each author, while contextual features refers to the comments present in the same threads where each author commented.

To detect explicit hate speech, we used HateBERT, a language model based on BERT-base and further pre-trained on over one million posts from Reddit communities that were banned for violating platform-wide hate speech policies [287]. This model has been shown to outperform general-purpose language models on hate speech detection tasks by capturing domain-specific language patterns prevalent in toxic and hateful on-line spaces. Each comment in our dataset was passed through HateBERT to determine whether it contained hate speech, resulting in a binary label [0, 1].

8.2.3 Lexical Feature Analysis

Previous works have shown that Incel community has an extensive cryptolect, or set of ingroup words that community members can use to identify themselves and that reflect the group’s worldview [288]. Klein et al. developed a Lexicon that contains the words and building blocks of terminology common to incels, including terms such as **toilet** (derogatory term for woman) **rice** (prefix referencing East/Southeast Asians, e.g., “ricecel”) **manlet** (short man, used derogatorily) **foid** (female humanoid) The presence of these terms was tracked across posts and comments to quantify the adoption of incel-specific language.

8.2.4 Classifier Features

To investigate whether it is possible to distinguish users who eventually join incel communities from those who do not, we implemented a supervised machine learning framework. Our analysis is defined by user-level features extracted from posting behavior, interactions, and linguistic characteristics, with the binary target variable defined as incels, indicating whether a user joined the incel community (1) or not (0).

In particular, our features can be categorized in 5 different types: Toxicity and Sentiment at the individual level, Toxicity and Sentiment at the contextual level, so the average toxicity and sentiment before each target user commented on the threads,

Contextual user metrics, such as number of unique users that commented before the target user, User activity and engagement, such as likes that they received, number of comments, unique subreddits in which they commented, and linguistic features such as presence of lexicon used by incels. Given the class imbalance of our dataset (293098 users than never joined incels community and 195 who did), we employed a downsampling technique in order to balance the dataset, creating 10 different balanced datasets. Prior to training, features were standardized using a StandardScaler.

We compared different machine learning models, including Random Forest, Decision Tree, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (K-NN). We conduct 5-fold cross-validation repeated over 10 simulations to ensure robustness and stability of performance estimates and report on the mean precision, recall, and F1 values across folds along with AUC in Table 8.1.

8.3 Results

Characterization

To examine how users change before and after engaging with the Incel community, we analyzed their behavior in terms of (i) toxicity and hate speech, and (ii) posting patterns, considering both time and the distribution of posts across subreddits. We define pre-engagement as all comments and posts made by users before they became active in the Incel community, and post-engagement as those made only after their first comment or post within the Incel subreddit.

Figure 8.1 summarizes changes in toxicity and hate speech. Overall, users become more toxic after engaging with the Incel community. The average toxicity of their comments rises from 0.20 (pre-engagement) to 0.265 (post-engagement). A similar increase is observed in posts, which increase from 0.19 to 0.25. For comparison, the baseline group—users who interacted in the same threads as incels but never joined the community—shows consistently lower toxicity (0.18), even compared to Incel users in their pre-engagement phase. Panel B of Figure 8.1 shows the ratio of hate speech, calculated as the proportion of a user’s posts or comments labeled as hate speech in each period. Here, the trend differs from toxicity. For comments, users express slightly more hate speech before engagement (0.048) compared to after (0.034). However, for posts, users post more hate speech after engagement (0.044) than before (0.027). As with toxicity, the baseline remains much lower (0.018) than both pre and post-engagement levels.

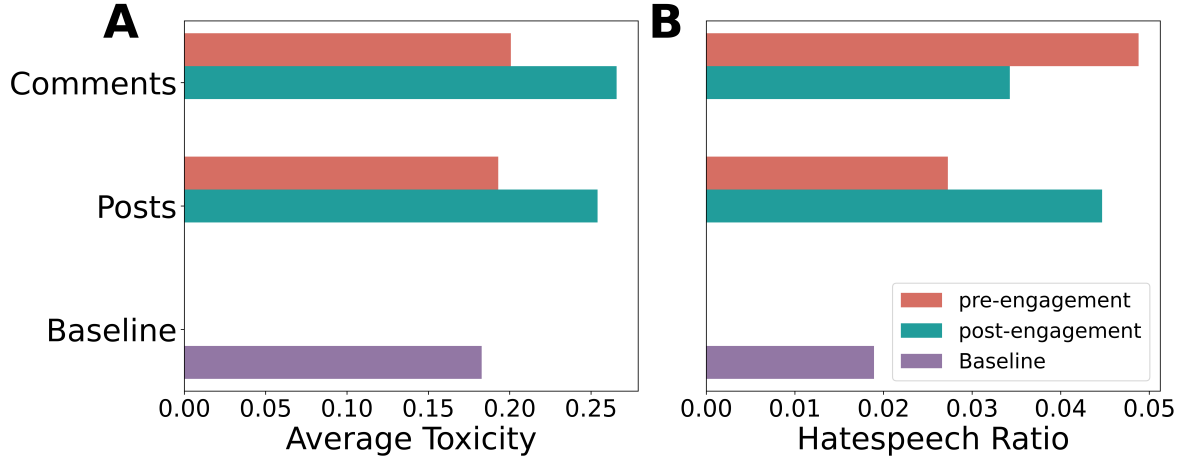


Figure 8.1: **Hate Speech and Toxicity** Distribution of Hate Speech and Toxicity for comments and posts before and after joining *incel* communities.

We then explored how users' posting patterns shift across different subreddit categories (Figure 8.2). For each user and time period (pre vs. post), we first counted their total comments and then computed the fraction falling into each content category. These fractions were then averaged across all users, yielding the mean distribution of activity across topics. Before joining the Incel community, users were primarily active in subreddits focused on relationships (0.45), entertainment (0.39), mental health and wellness (0.27), and gaming (0.24), lifestyle (0.19), showing interest in leisure activities and participation in advice and/or support-oriented groups. Interestingly, social issues, characterized by comprising the most radicalized communities, has a value of 0.19. On the contrary, users tend to concentrate more on relationship (0.35) and social issues (0.35), followed by entertainment (0.30) and gaming (0.19). It is interesting to notice that all the categories tend to decrease after the engagement, except for the Social Issues category, that increases by 85 % with respect to the pre-engagement moment, as shown in the C panel. With respect to the comments, relationship (0.41), entertainment (0.37), followed by social issues (0.21), gaming (0.16) and mental health and wellness (0.16) are the most frequent communities of users before the transition to the Incel subreddit. Similarly to the previous case, after the engagement, relationship (0.33) and social issues (0.33) are the most prominent channels in which users comment, while interest for entertainment (0.26) and gaming (0.12) decreases. The panel D shows the relative difference between the post and pre-engagement in terms of activity across various subreddits, highlighting how social issues themes are the only one increasing about 58 % with respect to the pre-engagement.

Beside insights about the posting behavior, it is interesting to analyze how on av-

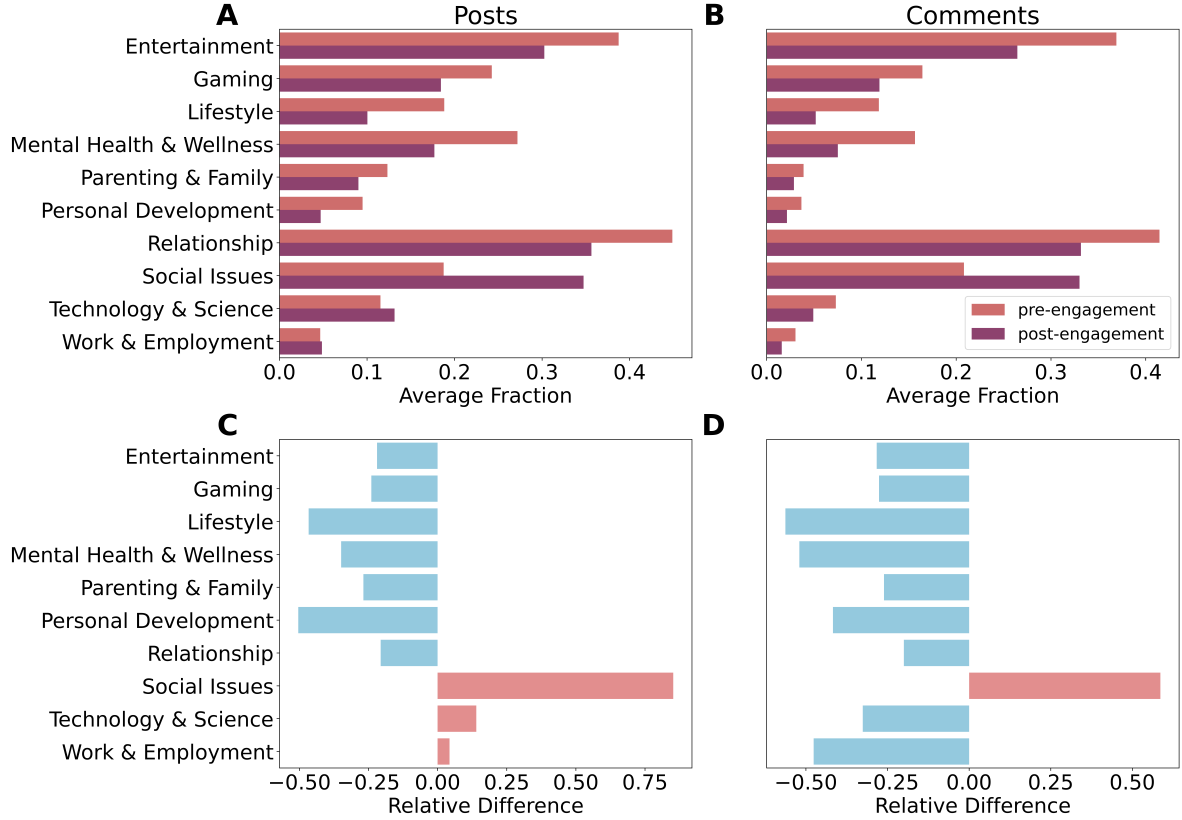


Figure 8.2: **Engagement in different communities.** Average fraction of posts and comments in different communities before and after joining *incel* community.

erage users post on those subreddits over time to observe a shift behavior in terms of temporal evolution. Figure 8.3 shows the absolute valeus in terms of messages (A) and comments posted per year, while providing the relative difference for posts and comments between the post and pre-engagement moment. We notice that users were more active in gaming(11.90) and entertainment (9.94 posts), followed by social issues (9.85) and relationship (8.93). After the engagement, the most activity category seems to be social issues with an average number of posts of 20.13, and a relative difference of 104 %. Another category that increases in a significant way is Personal Development going from an average of 1.8 to 5.17 posts per year with a relative increase of 187 %. With respect to the comments, the most common category Entertainment goes from an absolute value of 208.50 comments to 150.82, while the second one Relationship varies from 99.66 to 119.45. In terms of relative difference, social issues shows an increase of 201 % from pre and after engagement with the Incels community.

Beyond the activity pattern, our analysis highlights the difference in terms of hate speech and toxicity between different categories of subreddits. As before, given the

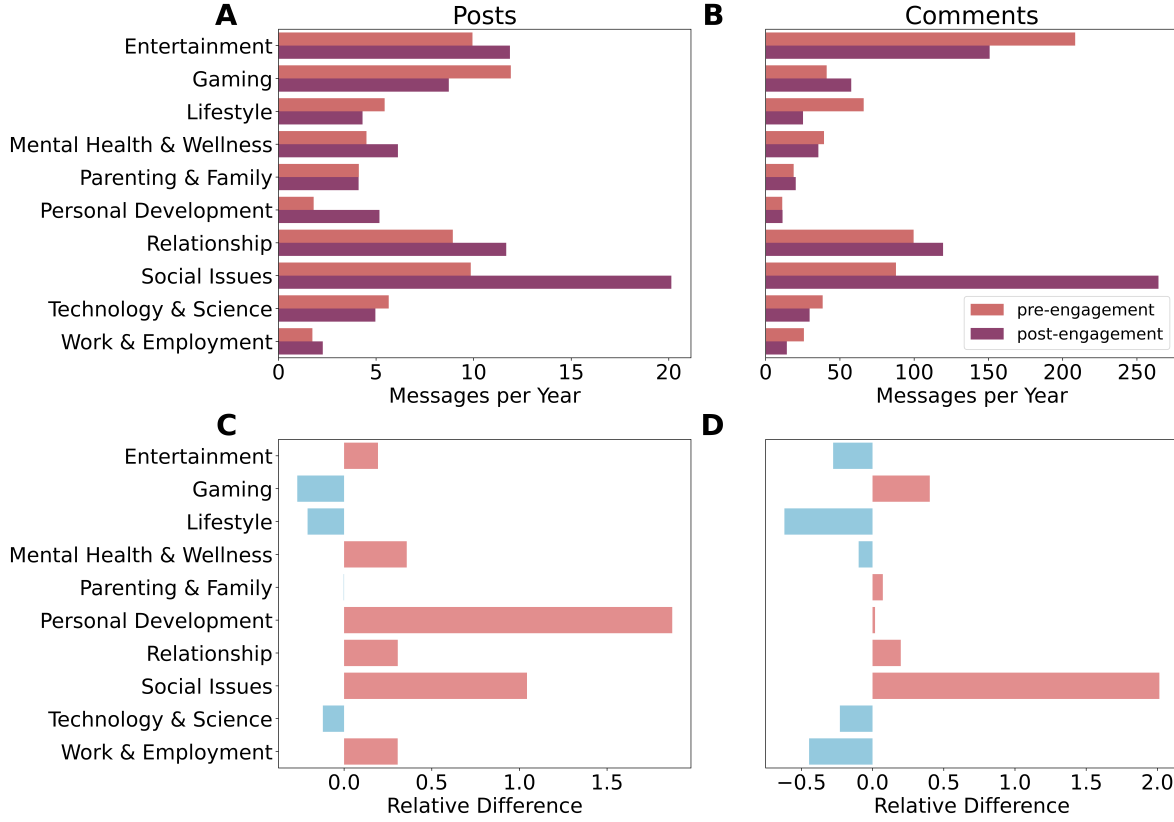


Figure 8.3: **Engagement in different communities normalize by time.** Average number of posts (A) and comments (B) normalized by time. Relative Difference between posts (C) and comments (D) over time.

10 categories of our interest, we calculated the average toxicity per user and for each category, resulting in the macro-average of toxicity both for pre and post engagement. Figure 8.4 presents the absolute value of the macro-average for both posts (A) and comments (B) and the corresponding relative difference between before and after engagement. With respect to the posts, the categories with the highest average value of toxicity are personal development (0.26), relationship (0.24), and social issues (0.22), when based on the pre-engagement. The shift between and post engagement is pronounced with respect to social issues (0.31), relationship (0.28) and mental health and wellness (0.22). Interestingly, the relative difference between pre and after engagement increases in all categories except personal development (-0.39) and lifestyle (-0.16). The highest increase is evident in categories like work and employment and technology and science, characterized both by having low values of toxicity in the pre engagement phase, respectively 0.06 and 0.07.

As before, the same analysis has been conducted with respect to the hate speech ra-

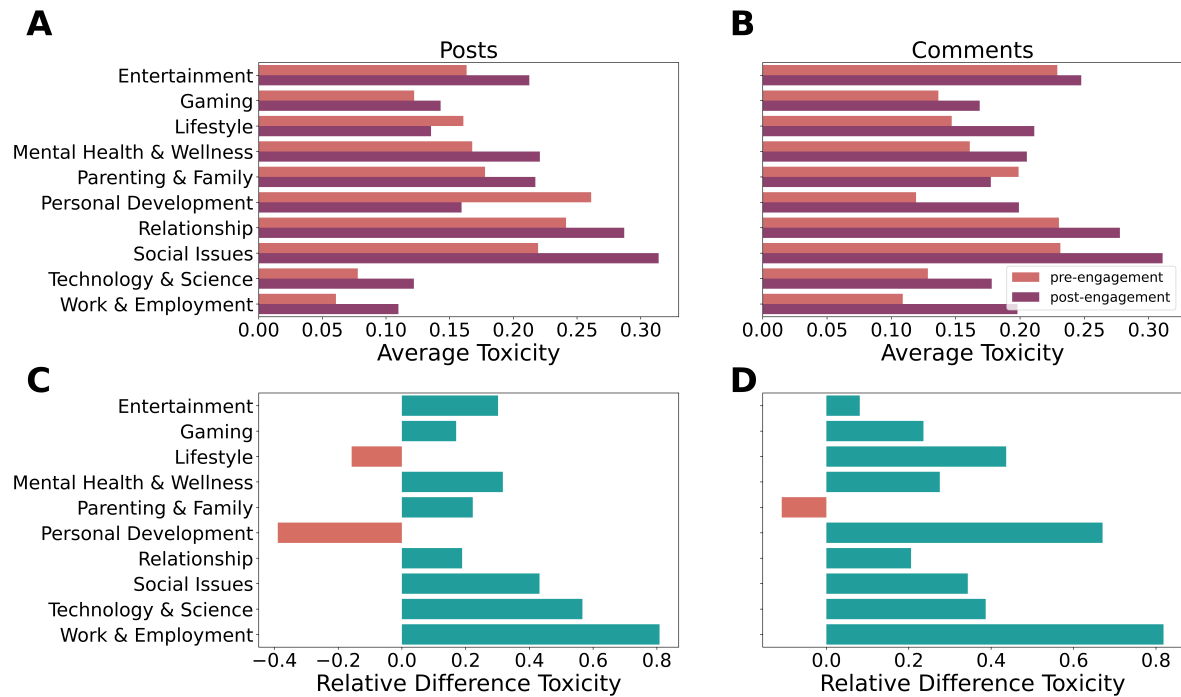


Figure 8.4: **Toxicity across subreddits.** Absolute values and Relative Difference of toxicity before and after the engagement with Incel community.

tio in several subreddits. Figure 8.5 shows the absolute values of hatespeech for both pre and after engagement and the relative differences between these two moments. With respect to the posts, social issues (0.08) and personal development (0.076), followed by lifestyle (0.04) and relationship (0.03), are the categories with more hate speech before the engagement. This trend is amplified after the engagement with the incels community, showing an increase of hate speech in many categories, especially personal development (0.09), social issues (0.08), lifestyle (0.05). In terms of relative differences, all categories have a substantial increase of hate speech except for the gaming subreddits. In particular, even categories related to leisure activities such as entertainment and technology and science show the highest increase in terms of hate speech with corresponding values of 80 % and 83 % between pre and after engagement. Similarly to this result, other categories, such as parenting & family and work & employment, have no presence of hate speech in the pre-engagement, while this is evident in the second phase.

Similar results have been found when it comes to comments, with social issues (0.09) and entertainment (0.06) as the primary channels where hate speech is spread. After the engagement, hate speech is still present mainly in social issues(0.085), relationship (0.05), personal development (0.04) and entertainment (0.04), highlighting how support

communities have a consistent pattern of hate speech in the comments section. Unlike posts, comments are increasingly becoming hatred with relative difference of 265 % for parenting & family and 194 % for technology & science.

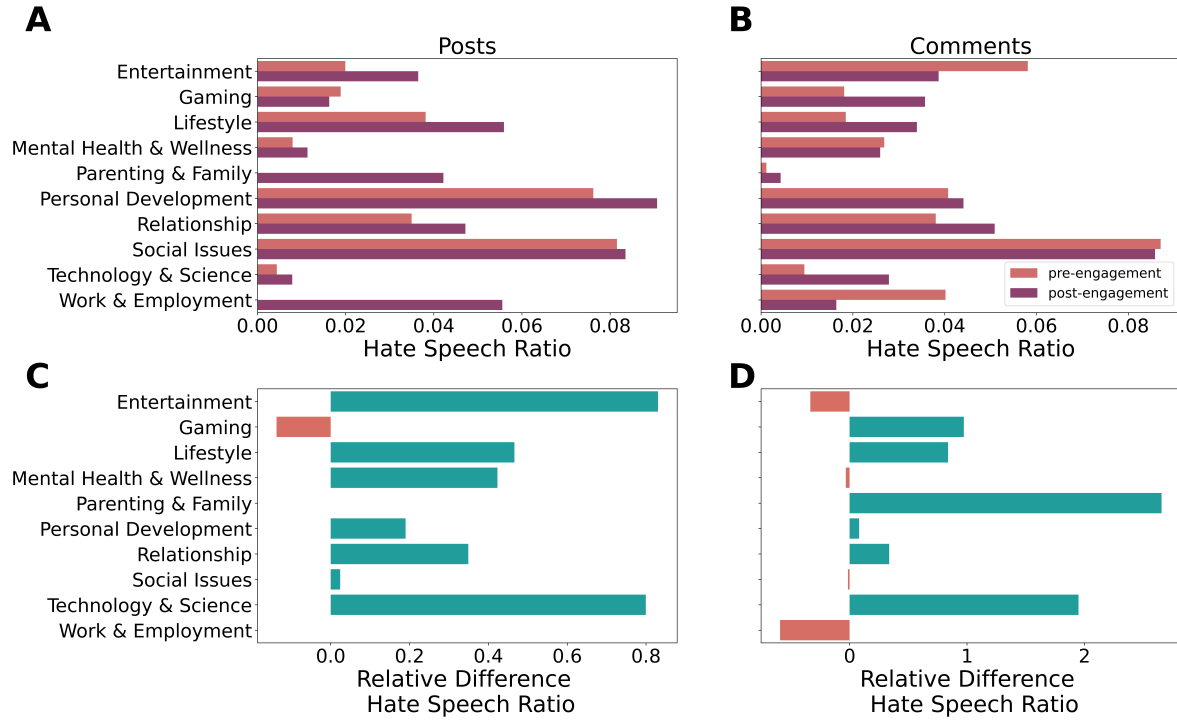


Figure 8.5: **Hate Speech across subreddits.** Absolute values and Relative Difference of hatespeech ratio before and after the engagement with Incel community.

Detection

We evaluated the performance of several machine learning models to predict whether a user would eventually join an incel community. Table 8.1 reports the average F1 score, precision, recall, accuracy, and AUC for each model across 5-fold cross-validation repeated over 10 simulations. Among all models, Random Forest achieved the highest predictive performance ($F1 = 0.895$, $AUC = 0.951$), followed by SVM ($F1 = 0.869$, $AUC = 0.926$) and K-NN ($F1 = 0.867$, $AUC = 0.912$).

To assess the contribution of different types of features to predictive performance, we trained the Random Forest model using only one feature set at a time. As shown in Table 8.2, contextual user metrics and linguistic features were among the most informative ($F1 = 0.84$), while toxicity and sentiment features, both individual and contextual, also contributed substantially ($F1 = 0.82$ – 0.83). In contrast, user activity and engagement alone were less predictive ($F1 = 0.746$), indicating that activity metrics by themselves

Model	F1 Score	Precision	Recall	Accuracy	AUC
Decision Tree	0.809	0.803	0.828	0.806	0.850
K-NN	0.867	0.914	0.831	0.874	0.912
Logistic Regression	0.843	0.875	0.827	0.847	0.924
Random Forest	0.895	0.913	0.887	0.899	0.951
SVM	0.869	0.892	0.856	0.873	0.926

Table 8.1: Comparison of model performance across evaluation metrics.

are not sufficient to identify potential incel users. Overall, these results suggest that a combination of contextual, linguistic, and toxicity-based features is most effective for predicting users who are likely to join incel communities, with ensemble models like Random Forest yielding the strongest overall performance.

Feature Set	Precision	Recall	F1 Score	AUC
Contextual User Metrics	0.8218	0.8742	0.8469	0.8797
Linguistic Features	0.8216	0.8483	0.8319	0.8749
Toxicity and Sentiment (Individual)	0.8653	0.7967	0.8271	0.9040
Toxicity and Sentiment (Context)	0.8805	0.7850	0.8230	0.9171
User Activity and Engagement	0.7415	0.7608	0.7458	0.8499

Table 8.2: Performance metrics for various feature sets.

Social Influence

To investigate whether prior interactions in online discussion threads might be predictive of a user’s transition to the incel community, we analyzed patterns of commenting behavior among users who joined the incel community and users who never joined that community in pre-engagement threads. We first constructed a user-thread matrix, where each row corresponds to a user and each column corresponds to a thread, with counts of comments as values. To normalize differences in activity across threads, we applied a TF-IDF transformation to the matrix [289], producing vector representations of user behavior. Pairwise cosine similarities were then computed between all incels community and, for computational feasibility, between non-incels users in randomly selected groups of equal size. These similarities quantify the degree of overlap in thread participation between users, providing a measure of behavioral similarity. Figure 8.6 shows the distribution of cosine similarity in Incels and Non Incels users. Despite the sparsity of user-thread interactions (most user pairs never commented in the same thread, yielding a median similarity of 0), the distributions differed significantly in their

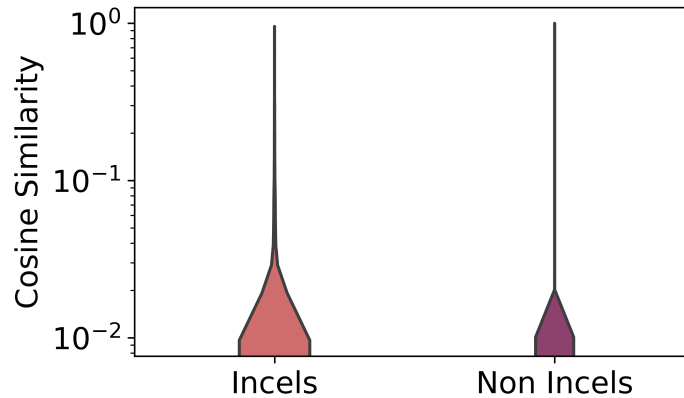


Figure 8.6: **Cosine similarity of participation of the same threads for Incels and No Incels users** We calculated the cosine similarity of threads commented on by Incels and Non-Incels users to understand whether users had engaged in prior discussions before joining the Incel subreddit.

upper tails. A Mann–Whitney U test confirmed this difference, with $U = 559,324,076$ and $p < 10^{-100}$, indicating that incel users are significantly more similar to each other in terms of their commenting patterns than non-prolific users. This analysis suggests that prior interactions and participation in overlapping discussion threads may provide a signal of social influence, highlighting that users who eventually transition to the incel community tend to engage in more overlapping or coordinated conversation patterns compared to non-prolific users.

8.4 Discussion

The findings of this study provide an understanding of how online behaviors and discursive patterns evolve during the radicalization process toward misogynistic online communities such as r/Incels. By analyzing users' activity before and after joining the Incel subreddit, this research demonstrates a clear behavioral and linguistic shift that reflects broader processes of ideological alignment and social reinforcement within extremist digital spaces. A key finding is the significant increase in toxicity following users' engagement with the Incel community. Both comments and posts displayed higher average toxicity scores after joining, indicating that exposure to radicalized discourses not only normalizes but amplifies hostile communication patterns. Interestingly, while hate speech within comments slightly declined post-engagement, hate speech in posts increases considerably. This difference may suggest a strategic shift in expression, users might adopt more hateful discourse when creating original content, while engaging in

subtler or coded forms of communication in discussion threads. The overall trend underscores how immersion in extremist communities reshapes communicative norms and lowers the threshold for aggressive or demeaning language use. The thematic analysis of subreddit participation further supports the notion of ideological drift. Prior to joining r/Incels, users were predominantly active in communities centered on entertainment, gaming, and relationships—spaces typically characterized by social exchange or leisure. After joining, however, there was a marked shift toward “Social Issues” subreddits, with an 85% increase in engagement relative to the pre-engagement phase. This finding highlights how users gradually reorient their attention from personal or recreational interests toward socio-political arenas, often those conducive to ideological polarization. Moreover, toxicity and hate speech were not uniformly distributed across categories. The post-engagement increase in toxicity within “Social Issues,” “Relationships,” and “Mental Health and Wellness” subreddits suggests that radicalized discourse permeates even those communities traditionally oriented toward support and discussion. The expansion of hate speech into otherwise non-political categories like “Entertainment” and “Technology & Science” further indicates the diffusion of extremist lexicons across the wider Reddit ecosystem. This diffusion effect represents a critical challenge for content moderation: radicalized narratives may propagate indirectly through mainstream spaces rather than remaining isolated within fringe communities. From a predictive standpoint, the machine learning results confirm that radicalization pathways are not random but can be anticipated from behavioral and linguistic signals. The Random Forest model achieved a F1 score of 0.89, outperforming other classifiers. Contextual user metrics and linguistic features, particularly those tied to Incel, specific lexicon—proved to be the strongest predictors of future membership. These findings imply that radicalization manifests not merely through toxic content but through structural engagement patterns and subtle shifts in vocabulary. Early detection frameworks leveraging such features could thus serve as valuable tools for identifying at-risk users before full immersion in extremist subcultures occurs. The social influence analysis reinforces this perspective. The significantly higher cosine similarity in commenting behavior among future incel members suggests that radicalization is not purely individual but socially mediated. Users who eventually joined the Incel subreddit were more likely to interact within overlapping discussion threads, forming a web of exposure and reinforcement that facilitated entry into extremist spaces. This aligns with theories of networked radicalization, where algorithmic curation and peer interaction strengthen ideological homogeneity.

This study underscores that online misogynistic radicalization is both detectable

and dynamic, emerging from a confluence of personal disaffection, social interaction, and exposure to normalized toxicity. Understanding this process not only highlight the mechanisms of digital extremism but also points toward data-driven interventions capable of mitigating its spread.

Chapter 9

Conclusion

The digital transformation of communication has changed the way in which information is produced, circulated, and consumed. The present work has examined this transformation through the lens of computational social science, focusing on the measurement and modeling of information disorders that emerge within large-scale digital ecosystems. The thesis aimed to develop a systematic understanding of how misinformation, disinformation, and related online harms propagate through social media, and how these processes interact with emotional, socio-economic, and structural dimensions of societies. By combining large-scale empirical analyses, methodological, and modeling approaches, the research has contributed to a growing interdisciplinary field that seeks to explain informational disorders not merely as isolated digital entities but as manifestations of complex sociotechnical systems. At its core, this research is motivated by the fact that the diffusion of false or misleading information represents one of the most pressing challenges of the contemporary information age given the easy spread of news by social media platforms. The acceleration and algorithmic mediation of communication have intensified the vulnerability of individuals and communities to manipulative content, especially in moments of social and political uncertainty such as the Covid-19 Pandemic. Building on this evidence, this thesis has attempted to measure these processes empirically, by developing tools and indicators capable of quantifying both the scale and volatility of informational risks across multiple temporal and geographical contexts. The analyses conducted in the first part of the thesis, focused on large-scale studies of the COVID-19 vaccination campaign and the emergence of Long COVID discourse, demonstrated that emotional responses constitute a fundamental aspect in the dissemination of information. In contrast to the assumption that misinformation spreads solely due to cognitive biases or political motivations,

the findings revealed that affective patterns, particularly those associated with anger, fear, and distrust, amplify the virality of low-quality or unreliable content. This is consistent with the theoretical frameworks of affective contagion [290] and emotional diffusion [291], which suggest that emotional intensity drives online engagement. The comparative analysis between verified and unverified users, as well as between different ideological orientations, further revealed how emotions not only circulate but polarize. Verified users, often representing institutional voices, exhibited more balanced or positively valenced expressions, whereas unverified and politically aligned users tended to express higher negativity, particularly when engaging with low-reliability news sources. The second chapter builds upon this work by exploring the conversations around Long Covid, demonstrating that attention towards this disease has been characterized by mainstream and scientific sources rather than unreliable type of news, highlighting also a weak influence of political leaning in driving such conversation. Extending beyond emotional and news consumption characterization, the thesis introduced a quantitative framework for analysing the diversity of information exposure through what has been defined as the concept of *news media diet*. Drawing on entropy-based measures, this approach provided an operational means of assessing how users engage with a variety or a subset of news sources. The findings indicated that users who systematically share unreliable content tend to operate within informational environments marked by low diversity and high repetition, reflecting a reduced informational variety. These results resonate with the literature on echo chambers and filter bubbles [273, 292], reinforcing the notion that the structure of one’s media consumption patterns can significantly constrain cognitive openness. The methodological contribution here lies in the ability to quantify such regularities across millions of digital traces, providing empirical support to theoretical claims that have often been difficult to verify at scale. In the fifth chapter, the focus expanded from the micro-level of user behaviour to the macro-level of systemic risk. The research examined the relationship between infodemic patterns, quantified through the Infodemic Risk Index (IRI), and the socio-economic characteristics of countries. The empirical results suggested that information disorder is not a random phenomenon but one embedded in broader societal structures that depend on a good amount of different variables. Countries with higher education levels, diversified media ecosystems, and stable institutions showed lower volatility in infodemic risk, while those with greater inequality and institutional fragility exhibited heightened exposure to unreliable information. This finding is consistent with earlier works linking misinformation susceptibility to socio-economic stress and governance quality [293]. After the macro relationship between countries and infodemic risk index, we further

explained the dynamic variability of the IRI, proposing a stochastic model to simulate the volatility of infodemic risk as a function of user engagement, follower distributions, and superspreaders behaviour. The model reproduced the empirical scaling properties observed in the data and highlighted how systemic instability emerges from a small number of highly connected users capable of amplifying unreliable content. The model clarified some of the mechanisms underlying observed volatility given the fact that it is able to reproduce the scaling factor, however, is still not fully capable of generating timeseries based on the parameters used. The final chapter of the thesis explored the relationship between information disorders and radicalization dynamics, focusing on online misogynistic communities as a case study. By tracking users' behavioural and linguistic evolution before and after their engagement with such communities, the research suggests that radicalization is a gradual process, marked by increasing toxicity, hate speech and social reinforcement. This work demonstrated the predictability of behavioural drift toward extremism, methodologically, it integrated linguistic, temporal, and network-based features into a unified analytical framework capable of identifying early warning signals of harmful online engagement. While these contributions advance our understanding of information disorders, the work also presents several limitations. The first limitation concerns our reliance on social media platforms, such as Twitter/X and Reddit. While providing rich datasets, they might not be fully representative of the entire digital discourse, given the fact that data have been collected with some specific keywords and on some specific subreddits. In addition, despite huge amount of data analyzed, many chapters of this thesis are based on the COVID-19 Infodemic Platform, restricting our case study to one specific event. Another limitation concerns the methods used for emotion detection in the early analyses, which relied on classical NLP techniques. While these approaches provided valuable insights, they are less capable of capturing nuanced emotional and contextual patterns compared to modern large language model (LLM)-based methods. Future work could use LLMs or hybrid approaches to improve the accuracy and depth of emotional analysis, potentially uncovering subtler affective dynamics in online information diffusion. Another aspect that has been at the center of the computational social science field is that changes in platform ownership, moderation policies, and data accessibility, such as those experienced by Twitter in 2023, introduce additional uncertainty into longitudinal analyses. Addressing these challenges will require a more integrated cross-platform incorporating decentralised that remain underexplored in current research. Future research should consider new platforms, such as Mastodon and Bluesky, to validate the approaches proposed as well as new emerging events. Despite these limitations, this research contributes to a broader

understanding of information disorders, both on the spread of misleading information and on harmful content spread in online social platforms. The findings suggest that the resilience of the digital ecosystems depends on the interplay between cognitive, technological and institutional factors. In this context, reducing the risks and the impact of misinformation requires not only algorithmic or regulatory interventions but also the development of socio-economic and cultural environments that promote trust, media diversity, and civic literacy. With respect to this regard, this thesis does not want to provide only new approaches in the quantification and modeling of information disorders but also be a starting point for interdisciplinary efforts in the mitigation of such phenomena, fostering collaboration across computational, social, and policy experts.

Bibliography

- [1] Claire Wardle et al. Information disorder: The essential glossary. *Harvard, MA: Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School*, 2018.
- [2] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. Assessing the risks of ‘infodemics’ in response to covid-19 epidemics. *Nature human behaviour*, 4(12):1285–1293, 2020.
- [3] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.
- [4] David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062, 2020.
- [5] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.
- [6] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.
- [7] Mason Walker and Katerina Eva Matsa. News consumption across social media in. *Pew Centre Research*, 2021.

- [8] Nathaniel Persily, Joshua A Tucker, and Joshua Aaron Tucker. Social media and democracy: The state of the field, prospects for reform. 2020.
- [9] Julie Posetti and Alice Matthews. A short guide to the history of ‘fake news’ and disinformation. *International Center for Journalists*, 7(2018):2018–07, 2018.
- [10] Global risks report 2025: Conflict, environment and disinformation top threats. <https://www.weforum.org/press/2025/01/global-risks-report-2025-conflict-environment-and-disinformation-top-threats/>. [Accessed 16-Sept-2025].
- [11] Nic Newman, A Ross Arguedas, Craig T Robertson, Rasmus Kleis Nielsen, and Richard Fletcher. *Digital news report 2025*. Reuters Institute for the study of Journalism, 2025.
- [12] Widespread global public concern about made-up news. <https://www.pewresearch.org/global/2025/04/24/widespread-global-public-concern-about-made-up-news/>. [Accessed 16-Sept-2025].
- [13] Kokil Jaidka, Tsuhan Chen, Simon Chesterman, Wynne Hsu, Min-Yen Kan, Mohan Kankanhalli, Mong Li Lee, Gyula Seres, Terence Sim, Araz Taeihagh, et al. Misinformation, disinformation, and generative ai: Implications for perception and policy. *Digital Government: Research and Practice*, 6(1):1–15, 2025.
- [14] Kalina Bontcheva, Symeon Papadopoulos, Filareti Tsalakanidou, Riccardo Gallotti, Lidia Dutkiewicz, Noémie Krack, Denis Teyssou, Francesco Severio Nucci, Jochen Spangenberg, Ivan Srba, et al. Generative ai and disinformation: recent advances, challenges, and opportunities. 2024.
- [15] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.
- [16] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
- [17] Anti-Defamation League. Online hate and harassment. the american experience 2024. *Center for Technology and Society*, 2024.

- [18] European Commission against Racism and Intolerance (ECRI). Hate speech and violence. <https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence>. [Accessed 15-Sept-2025].
- [19] United Nations. ‘hate speech is poison in the well of society,’ says guterres. <https://news.un.org/en/story/2025/06/1164501>. [Accessed 15-Sept-2025].
- [20] Christopher J Ferguson. Does the internet make the world worse? depression, aggression and polarization in the social media age. *Bulletin of Science, Technology & Society*, 41(4):116–135, 2021.
- [21] Sharon Meraz. The fight for ‘how to think’: Traditional media, social networks, and issue interpretation. *Journalism*, 12(1):107–127, 2011.
- [22] Number of internet and social media users worldwide as of february 2025. <https://www.statista.com/statistics/617136/digital-population-worldwide/>. [Accessed 25-Aug-2025].
- [23] Fabrizio Gilardi, Theresa Gessler, Maël Kubli, and Stefan Müller. Social media and political agenda setting. *Political communication*, 39(1):39–60, 2022.
- [24] Elena Pavan. Collective action and web 2.0. an exploratory network analysis of twitter use during campaigns. *Sociologica*, 7(3):0–0, 2013.
- [25] Danah Boyd. Social network sites as networked publics: Affordances, dynamics, and implications. In *A networked self*, pages 47–66. Routledge, 2010.
- [26] José Van Dijck and Thomas Poell. Understanding social media logic. *Media and communication*, 1(1):2–14, 2013.
- [27] Tarleton Gillespie. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, 167(2014):167, 2014.
- [28] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.
- [29] Caitlin O’kane. Twitter is officially ending its old verification process on april 1. to get a blue check mark, you’ll have to pay. <https://www.cbsnews.com/news/twitter-blue-check-verification-ending-new-subscription-april-1-elon-musk/>, CBS News, 3 2023. [Accessed 5-Oct-2025].

- [30] Shelley Boulianne, Mireille Lalancette, and David Ilkiw. “school strike 4 climate”: Social media and the international youth protest on climate change. *Media and Communication*, 8(2):208–218, 2020.
- [31] Alvin B Tillery Jr. What kind of movement is black lives matter? the view from twitter. *Journal of Race, Ethnicity, and Politics*, 4(2):297–323, 2019.
- [32] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.
- [33] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019.
- [34] Elihu Katz and David Foulkes. On the use of the mass media as “escape”: Clarification of a concept. *Public opinion quarterly*, 26(3):377–388, 1962.
- [35] Elihu Katz, Hadassah Haas, and Michael Gurevitch. On the use of the mass media for important things. *American sociological review*, pages 164–181, 1973.
- [36] Denis McQuail. *Mass communication theory: An introduction*. Sage Publications, Inc, 1987.
- [37] Anita Whiting and David Williams. Why people use social media: a uses and gratifications approach. *Qualitative market research: an international journal*, 16(4):362–369, 2013.
- [38] Carol F Scott, Laina Y Bay-Cheng, Mark A Prince, Thomas H Nochajski, and R Lorraine Collins. Time spent online: Latent profile analyses of emerging adults’ social media use. *Computers in Human Behavior*, 75:311–319, 2017.
- [39] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.
- [40] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone’s an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74, 2011.
- [41] Bence Bago, David G Rand, and Gordon Pennycook. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general*, 149(8):1608, 2020.

-
- [42] Eytan Adar and Lada A Adamic. Tracking information epidemics in blogspace. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 207–214. IEEE, 2005.
 - [43] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management science*, 62(1):180–196, 2016.
 - [44] Jonas L Juul and Johan Ugander. Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences*, 118(46):e2100786118, 2021.
 - [45] Paul Hendriks Vettehen, Koos Nuijten, and Allerd Peeters. Explaining effects of sensationalism on liking of television news stories: The role of emotional arousal. *Communication Research*, 35(3):319–338, 2008.
 - [46] Ming M Boyer. Aroused argumentation: How the news exacerbates motivated reasoning. *The International Journal of Press/Politics*, 28(1):92–115, 2023.
 - [47] Siyi Guo, Zihao He, Ashwin Rao, Eugene Jang, Yuanfeixue Nan, Fred Morstatter, Jeffrey Brantingham, and Kristina Lerman. Measuring online emotional reactions to offline events. *arXiv preprint arXiv:2307.10245*, 2023.
 - [48] Siyi Guo, Keith Burghardt, Ashwin Rao, and Kristina Lerman. Emotion regulation and dynamics of moral concerns during the early covid-19 pandemic. *arXiv preprint arXiv:2203.03608*, 2022.
 - [49] Naga Vemprala, Paras Bhatt, Rohit Valecha, and HR Rao. Emotions during the covid-19 crisis: A health versus economy analysis of public responses. *American Behavioral Scientist*, 65(14):1972–1989, 2021.
 - [50] Jianghao Wang, Yichun Fan, Juan Palacios, Yuchen Chai, Nicolas Guetta-Jeanrenaud, Nick Obradovich, Chenghu Zhou, and Siqi Zheng. Global evidence of expressed sentiment alterations during the covid-19 pandemic. *Nature Human Behaviour*, 6(3):349–358, 2022.
 - [51] Xiangliang Zhang, Qiang Yang, Somayah Albaradei, Xiaoting Lyu, Hind Alamro, Adil Salhi, Changsheng Ma, Manal Alshehri, Inji Ibrahim Jaber, Faroug Tifratene, et al. Rise and fall of the global conversation and shifting sentiments during the covid-19 pandemic. *Humanities and social sciences communications*, 8(1):1–10, 2021.

- [52] Claire Wardle and Hossein Derakhshan. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg, 2017.
- [53] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining “fake news” a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.
- [54] Rasmus Nielsen and Lucas Graves. "news you don't believe": Audience perspectives on fake news. *Reuters Institute for the Study of Journalism*, 2017.
- [55] Andrew Dawson and Martin Innes. How russia's internet research agency built its disinformation campaign. *The Political Quarterly*, 90(2):245–256, 2019.
- [56] Giancarlo Ruffo, Alfonso Semeraro, Anastasia Giachanou, and Paolo Rosso. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer science review*, 47:100531, 2023.
- [57] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.
- [58] Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, 2022.
- [59] James H Kuklinski, Paul J Quirk, Jennifer Jerit, David Schwieder, and Robert F Rich. Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3):790–816, 2000.
- [60] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.
- [61] Fabiana Zollo and Walter Quattrociocchi. Misinformation spreading on facebook. In *Complex spreading phenomena in social systems: Influence and contagion in real-world social networks*, pages 177–196. Springer, 2018.
- [62] Stephan Lewandowsky and Sander Van Der Linden. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2):348–384, 2021.

- [63] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.
- [64] Michał Choraś, Konstantinos Demestichas, Agata Giełczyk, Álvaro Herrero, Paweł Ksieniewicz, Konstantina Remoundou, Daniel Urda, and Michał Woźniak. Advanced machine learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing*, 101:107050, 2021.
- [65] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.
- [66] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.
- [67] Eiman Alothali, Nazar Zaki, Elfadil A Mohamed, and Hany Alashwal. Detecting social bots on twitter: a literature review. In *2018 International conference on innovations in information technology (IIT)*, pages 175–180. IEEE, 2018.
- [68] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7, 2019.
- [69] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552, 2019.
- [70] Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(9):544, 2022.
- [71] John Zarocostas. How to fight an infodemic. *The lancet*, 395(10225):676, 2020.
- [72] David J Rothkopf. When the buzz bites back. *The Washington Post*, 11:B1–B5, 2003.
- [73] World Health Organization. Novel coronavirus(2019-ncov): Situation report, 13. 2 2020.
- [74] MediaBias FactCheck. Mediabiasfactcheck (2020). <https://mediabiasfactcheck.com/>.

- [75] Anna Bertani, Riccardo Gallotti, Stefano Menini, Pierluigi Sacco, and Manlio De Domenico. Large-scale digital signatures of emotional response to the covid-19 vaccination campaign. *EPJ Data Science*, 13(1):20, 2024.
- [76] Friederike Behrens and Mariska E Kret. The interplay between face-to-face contact and feedback on cooperation during real-life interactions. *Journal of Nonverbal Behavior*, 43(4):513–528, 2019.
- [77] Aimee L Drolet and Michael W Morris. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50, 2000.
- [78] Robert Kurzban. The social psychophysics of cooperation: Nonverbal communication in a public goods game. *Journal of Nonverbal Behavior*, 25:241–259, 2001.
- [79] Judee K Burgoon, Xinran Wang, Xunyu Chen, Steven J Pentland, and Norah E Dunbar. Nonverbal behaviors “speak” relational messages of dominance, trust, and composure. *Frontiers in psychology*, 12:624177, 2021.
- [80] Jing Jiang, Bohan Dai, Danling Peng, Chaozhe Zhu, Li Liu, and Chunming Lu. Neural synchronization during face-to-face communication. *Journal of Neuroscience*, 32(45):16064–16069, 2012.
- [81] Sebastian Feese, Bert Arnrich, Gerhard Tröster, Bertolt Meyer, and Klaus Jonas. Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 520–525. IEEE, 2012.
- [82] Joshua Baker. The empathic foundations of security dilemma de-escalation. *Political Psychology*, 40(6):1251–1266, 2019.
- [83] Robin Ian MacDonald Dunbar. *Grooming, gossip, and the evolution of language*. 1996.
- [84] Patricia Sariñana-González, Ángel Romero-Martínez, and Luis Moya-Albiol. Cooperation between strangers in face-to-face dyads produces more cardiovascular activation than competition or working alone. *Journal of Psychophysiology*, 2018.

- [85] James Halpin and Chris Wilson. How online interaction radicalises while group involvement restrains: a case study of action zealandia from 2019 to 2021. *Political Science*, 74(1):18–33, 2022.
- [86] Marjolijn L Antheunis, Alexander P Schouten, Patti M Valkenburg, and Jochen Peter. Interactive uncertainty reduction strategies and verbal affection in computer-mediated communication. *Communication Research*, 39(6):757–780, 2012.
- [87] Lyrisa Barnett Lidsky. Incendiary speech and social media. *Tex. Tech L. Rev.*, 44:147, 2011.
- [88] Alan Cowen, Disa Sauter, Jessica L Tracy, and Dacher Keltner. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90, 2019.
- [89] Kristen A Lindquist, Erika H Siegel, Karen S Quigley, and Lisa Feldman Barrett. The hundred-year emotion war: are emotions natural kinds or psychological constructions? comment on lench, flores, and bench (2011). 2013.
- [90] Vittorio Gallese, Christian Keysers, and Giacomo Rizzolatti. A unifying view of the basis of social cognition. *Trends in cognitive sciences*, 8(9):396–403, 2004.
- [91] David Garcia, Arvid Kappas, Dennis Küster, and Frank Schweitzer. The dynamics of emotions in online interaction. *Royal Society open science*, 3(8):160059, 2016.
- [92] Roser Beneito-Montagut. Encounters on the social web: Everyday life and emotions online. *Sociological Perspectives*, 58(4):537–553, 2015.
- [93] Roser Beneito-Montagut. Emotions, everyday life, and the social web: Age, gender, and social web engagement effects on online emotional expression. *Sociological Research Online*, 22(4):87–104, 2017.
- [94] Anna Chmiel, Julian Sienkiewicz, Mike Thelwall, Georgios Paltoglou, Kevan Buckley, Arvid Kappas, and Janusz A Hołyst. Collective emotions online and their influence on community life. *PloS one*, 6(7):e22207, 2011.
- [95] Jesse Fox and Jennifer J Moreland. The dark side of social networking sites: An exploration of the relational and psychological stressors associated with facebook use and affordances. *Computers in human behavior*, 45:168–176, 2015.

- [96] Jaehwan Jahng, Jerald D Kralik, Dong-Uk Hwang, and Jaeseung Jeong. Neural dynamics of two players when using nonverbal cues to gauge intentions to cooperate during the prisoner’s dilemma game. *NeuroImage*, 157:263–274, 2017.
- [97] Krystle Phirangee and Jim Hewitt. Loving this dialogue!!!!: Expressing emotion through the strategic manipulation of limited non-verbal cues in online learning environments. In *Emotions, technology, and learning*, pages 69–85. 2016.
- [98] Joseph B Walther, Tracy Loh, and Laura Granka. Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of language and social psychology*, 24(1):36–65, 2005.
- [99] Heung-Yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, 2018.
- [100] Monica A Riordan and Roger J Kreuz. Emotion encoding and interpretation in computer-mediated communication: Reasons for use. *Computers in human behavior*, 26(6):1667–1673, 2010.
- [101] Pawel Sobkowicz and Antoni Sobkowicz. Two-year study of emotion and communication patterns in a highly polarized political discussion forum. *Social Science Computer Review*, 30(4):448–469, 2012.
- [102] Qinfeng Zhu, Brian E Weeks, and Nojin Kwak. Implications of online incidental and selective exposure for political emotions: Affective polarization during elections. *New media & society*, page 14614448211061336, 2021.
- [103] Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports*, 11(1):22721, 2021.
- [104] Bjarke Mønsted and Sune Lehmann. Characterizing polarization in online vaccine discourse—a large-scale study. *PloS one*, 17(2):e0263746, 2022.
- [105] Kai Chen, Zihao He, Rong-Ching Chang, Jonathan May, and Kristina Lerman. Anger breeds controversy: Analyzing controversy and emotions on reddit. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 44–53. Springer, 2023.

- [106] Julia Jakob, Timo Dobbrick, Rainer Freudenthaler, Patrik Haffner, and Hartmut Wessler. Is constructive engagement online a lost cause? toxic outrage in online user comments across democratic political systems and discussion arenas. *Communication Research*, 50(4):508–531, 2023.
- [107] Helena Tomljenovic, Andreja Bubic, and Nikola Erceg. It just doesn’t feel right—the relevance of emotions and intuition for parental vaccine conspiracy beliefs and vaccination uptake. *Psychology & health*, 35(5):538–554, 2020.
- [108] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.
- [109] Jessica L Tracy and Daniel Randles. Four models of basic emotions: a review of ekman and cordaro, izard, levenson, and panksepp and watt. *Emotion review*, 3(4):397–405, 2011.
- [110] Bernd Lahno. Trust and emotion. *The Routledge handbook of trust and philosophy*, pages 147–159, 2020.
- [111] Zizi Papacharissi. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283, 2004.
- [112] David Asker and Elias Dinas. Thinking fast and furious: Emotional intensity and opinion polarization in online media. *Public Opinion Quarterly*, 83(3):487–509, 2019.
- [113] Edmond Ng. The pandemic of hate is giving covid-19 a helping hand. *The American Journal of Tropical Medicine and Hygiene*, 102(6):1158, 2020.
- [114] Pier Luigi Sacco, Riccardo Gallotti, Federico Pilati, Nicola Castaldo, and Manlio De Domenico. Emergence of knowledge communities and information centralization during the covid-19 pandemic. *Social Science & Medicine*, 285:114215, 2021.
- [115] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

- [116] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9, 2018.
- [117] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440, 2018.
- [118] Wen Shi, Diyi Liu, Jing Yang, Jing Zhang, Sanmei Wen, and Jing Su. Social bots’ sentiment engagement in health emergencies: A topic-based analysis of the covid-19 pandemic discussions on twitter. *International Journal of Environmental Research and Public Health*, 17(22):8701, 2020.
- [119] Jon-Patrick Allem and Emilio Ferrara. Could social bots pose a threat to public health? *American journal of public health*, 108(8):1005, 2018.
- [120] Sandra González-Bailón and Manlio De Domenico. Bots are less central than verified accounts during contentious political events. *Proceedings of the National Academy of Sciences*, 118(11):e2013443118, 2021.
- [121] Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *arXiv preprint arXiv:1707.00086*, 2017.
- [122] Massimo Stella, Marco Cristoforetti, and Manlio De Domenico. Influence of augmented humans in online interactions during voting events. *PloS one*, 14(5):e0214210, 2019.
- [123] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103, 2020.
- [124] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.
- [125] Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018.

- [126] Christos Baziotis, Nikos Pelekis, and Christos Doukeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.
- [127] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.
- [128] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.
- [129] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009.
- [130] Michael Bang Petersen, Mathias Osmundsen, and John Tooby. The evolutionary psychology of conflict and the functions of falsehood. *The Politics of Truth in Polarized America*, page 131, 2021.
- [131] Guillem Rico, Marc Guinjoan, and Eva Anduiza. The emotional underpinnings of populism: How anger and fear affect populist attitudes. *Swiss Political Science Review*, 23(4):444–461, 2017.
- [132] Ani Nalbandian, Kartik Sehgal, Aakriti Gupta, Mahesh V Madhavan, Claire McGroder, Jacob S Stevens, Joshua R Cook, Anna S Nordvig, Daniel Shalev, Tejasav S Sehrawat, et al. Post-acute covid-19 syndrome. *Nature medicine*, 27(4):601–615, 2021.
- [133] Hannah E Davis, Lisa McCorkell, Julia Moore Vogel, and Eric J Topol. Long covid: major findings, mechanisms and recommendations. *Nature Reviews Microbiology*, 21(3):133–146, 2023.
- [134] Daniel M Altmann, Emily M Whettlock, Siyi Liu, Deepa J Arachchillage, and Rosemary J Boyton. The immunology of long covid. *Nature Reviews Immunology*, 23(10):618–634, 2023.
- [135] Jon Klein, Jamie Wood, Jillian R Jaycox, Rahul M Dhodapkar, Peiwen Lu, Jeff R Gehlhausen, Alexandra Tabachnikova, Kerrie Greene, Laura Tabacof, Amyn A

- Malik, et al. Distinguishing features of long covid identified through immune profiling. *Nature*, 623(7985):139–148, 2023.
- [136] Alyson Haslam, Timothée Olivier, and Vinay Prasad. The definition of long covid used in interventional studies. *European Journal of Clinical Investigation*, 53(8):e13989, 2023.
- [137] Joan B Soriano, Srinivas Murthy, John C Marshall, Pryanka Relan, and Janet V Diaz. A clinical case definition of post-covid-19 condition by a delphi consensus. *The Lancet infectious diseases*, 22(4):e102–e107, 2022.
- [138] Ana Luiza Cabrera Martimbianco, Rafael Leite Pacheco, Ângela Maria Bagattini, and Rachel Riera. Frequency, signs and symptoms, and criteria adopted for long covid-19: A systematic review. *International journal of clinical practice*, 75(10):e14357, 2021.
- [139] Mirembe Woodrow, Charles Carey, Nida Ziauddeen, Rebecca Thomas, Athena Akrami, Vittoria Lutje, Darren C Greenwood, and Nisreen A Alwan. Systematic review of the prevalence of long covid. In *Open forum infectious diseases*, volume 10, page ofad233. Oxford University Press US, 2023.
- [140] Destin Groff, Ashley Sun, Anna E Ssentongo, Djibril M Ba, Nicholas Parsons, Govinda R Poudel, Alain Lekoubou, John S Oh, Jessica E Ericson, Paddy Ssentongo, et al. Short-term and long-term rates of postacute sequelae of sars-cov-2 infection: a systematic review. *JAMA network open*, 4(10):e2128568–e2128568, 2021.
- [141] Amy D Proal and Michael B VanElzakker. Long covid or post-acute sequelae of covid-19 (pasc): an overview of biological factors that may contribute to persistent symptoms. *Frontiers in microbiology*, 12:698169, 2021.
- [142] Yapeng Su, Dan Yuan, Daniel G Chen, Rachel H Ng, Kai Wang, Jongchan Choi, Sarah Li, Sunga Hong, Rongyu Zhang, Jingyi Xie, et al. Multiple early factors anticipate post-acute covid-19 sequelae. *Cell*, 185(5):881–895, 2022.
- [143] Kin Israel Notarte, Jesus Alfonso Catahay, Jacqueline Veronica Velasco, Adriel Pastrana, Abbygail Therese Ver, Flos Carmeli Pangilinan, Princess Juneire Peligro, Michael Casimiro, Jonathan Jaime Guerrero, Ma Margarita Leticia Gellaco, et al. Impact of covid-19 vaccination on the risk of developing long-covid and on existing long-covid symptoms: A systematic review. *EClinicalMedicine*, 53, 2022.

-
- [144] Christoph Schultheiß, Edith Willscher, Lisa Paschold, Cornelia Gottschick, Bianca Klee, Svenja-Sibylla Henkes, Lidia Bosurgi, Jochen Dutzmann, Daniel Sedding, Thomas Frese, et al. The il-1 β , il-6, and tnf cytokine triad is associated with post-acute sequelae of covid-19. *Cell Reports Medicine*, 3(6), 2022.
- [145] Amy D Proal, Michael B VanElzakker, Soo Aleman, Katie Bach, Brittany P Boribong, Marcus Buggert, Sara Cherry, Daniel S Chertow, Helen E Davies, Christopher L Dupont, et al. Sars-cov-2 reservoir in post-acute sequelae of covid-19 (pasc). *Nature Immunology*, 24(10):1616–1627, 2023.
- [146] Olalekan Lee Aiyegbusi, Sarah E Hughes, Grace Turner, Samantha Cruz Rivera, Christel McMullan, Joht Singh Chandan, Shamil Haroon, Gary Price, Elin Haf Davies, Krishnarajah Nirantharakumar, et al. Symptoms, complications and management of long covid: a review. *Journal of the Royal Society of Medicine*, 114(9):428–442, 2021.
- [147] Harry Crook, Sanara Raza, Joseph Nowell, Megan Young, and Paul Edison. Long covid—mechanisms, risk factors, and management. *bmj*, 374, 2021.
- [148] Vasileios Nittas, Manqi Gao, Erin A West, Tala Ballouz, Dominik Menges, Sarah Wulf Hanson, and Milo Alan Puhan. Long covid through a public health lens: an umbrella review. *Public health reviews*, 43:1604501, 2022.
- [149] Joseph Dumit. Illnesses you have to fight to get: Facts as forces in uncertain, emergent illnesses. *Social science & medicine*, 62(3):577–590, 2006.
- [150] Kristin K Barker. Electronic support groups, patient-consumers, and medicalization: The case of contested illness. *Journal of health and social behavior*, 49(1):20–36, 2008.
- [151] Michael Murphy, Nicholas Kontos, and Oliver Freudenreich. Electronic support groups: an open line of communication in contested illness. *Psychosomatics*, 57(6):547–555, 2016.
- [152] Felicity Callard and Elisa Perego. How and why patients made long covid. *Social science & medicine*, 268:113426, 2021.
- [153] Ikhwan Yuda Kusuma and Suherman Suherman. The pulse of long covid on twitter: A social network analysis. *Archives of Iranian Medicine*, 27(1):36, 2024.

- [154] Esperanza Miyake and Sam Martin. Long covid: Online patient narratives, public health communication and vaccine hesitancy. *Digital health*, 7:20552076211059649, 2021.
- [155] David Russell, Naomi J Spence, Jo-Ana D Chase, Tatum Schwartz, Christa M Tumminello, and Erin Bouldin. Support amid uncertainty: Long covid illness experiences and the role of online communities. *SSM-Qualitative Research in Health*, 2:100177, 2022.
- [156] Paolo Biancone, Silvana Secinaro, Roberto Marseglia, and Davide Calandra. E-health for the future. managerial perspectives using a multiple case study approach. *Technovation*, 120:102406, 2023.
- [157] Bayou Tilahun Assaye, Muluken Belachew, Aynadis Worku, Sefefe Birhanu, Ayenew Sisay, Mitiku Kassaw, and Habtamu Mekonen. Perception towards the implementation of telemedicine during covid-19 pandemic: a cross-sectional study. *BMC Health Services Research*, 23(1):967, 2023.
- [158] Chen Chen, Spencer R Hauptert, Lauren Zimmermann, Xu Shi, Lars G Fritsche, and Bhramar Mukherjee. Global prevalence of post-coronavirus disease 2019 (covid-19) condition or long covid: a meta-analysis and systematic review. *The Journal of infectious diseases*, 226(9):1593–1607, 2022.
- [159] Manlio De Domenico. Prevalence of long covid decreases for increasing covid-19 vaccine uptake. *PLOS Global Public Health*, 3(6):e0001917, 2023.
- [160] Tara Kirk Sell, Kelsey Lane Warmbrod, Crystal Watson, Marc Trotochaud, Elena Martin, Sanjana J Ravi, Maurice Balick, and Emile Servan-Schreiber. Using prediction polling to harness collective intelligence for disease forecasting. *BMC public health*, 21:1–9, 2021.
- [161] Emile Servan-Schreiber and Camille Larmanou. Crowd forecasting infectious disease outbreaks. In *The Routledge Handbook of Collective Intelligence for Democracy and Governance*, pages 401–406. Routledge.
- [162] Ray-Ming Chen. Quantifying collective intelligence and behaviours of sars-cov-2 via environmental resources from virus’ perspectives. *Environmental Research*, 198:111278, 2021.

-
- [163] Sarah Wulf Hanson, Cristiana Abbafati, Joachim G Aerts, Ziyad Al-Aly, Charlie Ashbaugh, Tala Ballouz, Oleg Blyuss, Polina Bobkova, Gouke Bonsel, Svetlana Borzakova, et al. A global systematic analysis of the occurrence, severity, and recovery pattern of long covid in 2020 and 2021. *MedRxiv*, 2022.
- [164] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- [165] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.
- [166] Tiago P Peixoto. *Descriptive vs. inferential community detection in networks: Pitfalls, myths and half-truths*. Cambridge University Press, 2023.
- [167] Tiago P Peixoto. The graph-tool python library. (*No Title*), 2017.
- [168] Arsham Ghavasieh and Manlio De Domenico. Diversity of information pathways drives sparsity in real-world networks. *Nature Physics*, 20(3):512–519, 2024.
- [169] Pascal Boyer and Nora Parren. Threat-related information suggests competence: A possible factor in the spread of rumors. *PloS one*, 10(6):e0128421, 2015.
- [170] Mahdi Jalili and Matjaž Perc. Information cascades in complex networks. *Journal of Complex Networks*, 5(5):665–693, 2017.
- [171] Tanushree Mitra, Graham Wright, and Eric Gilbert. Credibility and the dynamics of collective attention. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–17, 2017.
- [172] Pietro Battiston, Ridhi Kashyap, and Valentina Rotondi. Reliance on scientists and experts during an epidemic: Evidence from the covid-19 outbreak in italy. *SSM-population health*, 13:100721, 2021.
- [173] Francesco Pierri, Brea L Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. Online misinformation is linked to early covid-19 vaccination hesitancy and refusal. *Scientific reports*, 12(1):5966, 2022.
- [174] Sebastian Funk, Erez Gilad, Chris Watkins, and Vincent AA Jansen. The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences*, 106(16):6872–6877, 2009.

- [175] Clara Granell, Sergio Gómez, and Alex Arenas. Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Physical review letters*, 111(12):128701, 2013.
- [176] Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. Disease containment strategies based on mobility and information dissemination. *Scientific reports*, 5(1):10650, 2015.
- [177] Joshua S Weitz, Sang Woo Park, Ceyhun Eksin, and Jonathan Dushoff. Awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations. *Proceedings of the National Academy of Sciences*, 117(51):32764–32771, 2020.
- [178] Valeria d’Andrea, Riccardo Gallotti, Nicola Castaldo, and Manlio De Domenico. Individual risk perception and empirical social structures shape the dynamics of infectious disease outbreaks. *PLOS Computational Biology*, 18(2):e1009760, 2022.
- [179] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitam, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3), 2020.
- [180] Guido Caldarelli, Rocco De Nicola, Marinella Petrocchi, Manuel Pratelli, and Fabio Saracco. Flow of online misinformation during the peak of the covid-19 pandemic in italy. *EPJ data science*, 10(1):34, 2021.
- [181] Peter R Harris, Elizabeth Sillence, and Pam Briggs. Perceived threat and corroboration: key factors that improve a predictive model of trust in internet-based health information and advice. *Journal of medical Internet research*, 13(3):e51, 2011.
- [182] Elizabeth Sillence, John Matthew Blythe, Pam Briggs, and Mark Moss. A revised model of trust in internet-based health information and advice: cross-sectional questionnaire study. *Journal of medical Internet research*, 21(11):e11125, 2019.
- [183] Margherita Schiavi, Stefania Fugazzaro, Anna Bertolini, Monica Denti, Carlotta Mainini, Monia Allisen Accogli, Ginevra Bedogni, Daniele Ghizzoni, Otmen Esseroukh, Cecilia Gualdi, et al. “like before, but not exactly”: the qualy-react qualitative inquiry into the lived experience of long covid. *BMC Public Health*, 22(1):599, 2022.

- [184] Hannah Metzler, Bernard Rimé, Max Pellert, Thomas Niederkrotenthaler, Anna Di Natale, and David Garcia. Collective emotions during the covid-19 outbreak. *Emotion*, 23(3):844, 2023.
- [185] Isabelle Freiling, Nicole M Krause, Dietram A Scheufele, and Dominique Brossard. Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during covid-19. *New media & society*, 25(1):141–162, 2023.
- [186] Federica Biassoni, Stefania Balzarotti, Daniela Abati, Alice Salducco, and Martina Gnerre. Narratives on the present and the future in the time of covid-19 pandemic: Uncertainty, subjective feeling and the role of positive anticipatory states. *Frontiers in Communication*, 8:1165219, 2023.
- [187] Kaixin Ji, Danula Hettiachchi, Flora D Salim, Falk Scholer, and Damiano Spina. Characterizing information seeking processes with multiple physiological signals. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1006–1017, 2024.
- [188] Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258, 2023.
- [189] Colton Margus, Natasha Brown, Attila J Hertelendy, Michelle R Safferman, Alexander Hart, and Gregory R Ciottone. Emergency physician twitter use in the covid-19 pandemic as a potential predictor of impending surge: Retrospective observational study. *Journal of medical Internet research*, 23(7):e28615, 2021.
- [190] Fengrui Jing, Zhenlong Li, Shan Qiao, Jiajia Zhang, Banky Olatosi, and Xiaoming Li. Using geospatial social media data for infectious disease studies: a systematic review. *International Journal of Digital Earth*, 16(1):130–157, 2023.
- [191] Lauren E Charles-Smith, Tera L Reynolds, Mark A Cameron, Mike Conway, Eric HY Lau, Jennifer M Olsen, Julie A Pavlin, Mika Shigematsu, Laura C Streichert, Katie J Suda, et al. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10):e0139701, 2015.

- [192] Patrick Pilipiec, Isak Samsten, and András Bota. Surveillance of communicable diseases using social media: A systematic review. *PLoS One*, 18(2):e0282101, 2023.
- [193] Eric Afful-Dadzie, Anthony Afful-Dadzie, and Sulemana Bankuoru Egala. Social media in health communication: A literature review of information quality. *Health Information Management Journal*, 52(1):3–17, 2023.
- [194] Sangwon Lee, Edson C Tandoc Jr, and Edmund WJ Lee. Social media may hinder learning about science; social media’s role in learning about covid-19. *Computers in Human Behavior*, 138:107487, 2023.
- [195] Babajide Osatuyi. Information sharing on social media sites. *Computers in Human Behavior*, 29(6):2622–2631, 2013.
- [196] Oscar Westlund and Mathias A Färdigh. Accessing the news in an age of mobile media: Tracing displacing and complementary effects of mobile news on newspapers and online news. *Mobile media & communication*, 3(1):53–74, 2015.
- [197] Rasmus Kleis Nielsen and Kim Christian Schrøder. The relative importance of social media for accessing, finding, and engaging with news: An eight-country cross-media comparison. *Digital journalism*, 2(4):472–489, 2014.
- [198] David Westerman, Patric R Spence, and Brandon Van Der Heide. Social media as information source: Recency of updates and credibility of information. *Journal of computer-mediated communication*, 19(2):171–183, 2014.
- [199] Rashi Renjith. The effect of information overload in digital media news content. *Communication and Media Studies*, 6(1):73–85, 2017.
- [200] Manuel Gomez Rodriguez, Krishna Gummadi, and Bernhard Schoelkopf. Quantifying information overload in social media and its impact on social contagions. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 170–179, 2014.
- [201] Ksenia Koroleva, Hanna Krasnova, and Oliver Günther. ‘stop spamming me!’-exploring information overload on facebook. 2010.
- [202] Peter Gordon Roetzel. Information overload in the information age: a review of the literature from business administration, business psychology, and related

- disciplines with a bibliometric approach and framework development. *Business research*, 12(2):479–522, 2019.
- [203] Usha M Rodrigues and Jian Xu. Regulation of covid-19 fake news infodemic in china and india. *Media International Australia*, 177(1):125–131, 2020.
- [204] Valeria Mazzeo and Andrea Rapisarda. Investigating fake and reliable news sources using complex networks analysis. *Frontiers in Physics*, 10:886544, 2022.
- [205] Ramón Salaverría, Nataly Buslón, Fernando López-Pan, Bienvenido León, Ignacio López-Goñi, and María-Carmen Erviti. Desinformación en tiempos de pandemia: tipología de los bulos sobre la covid-19. *Profesional de la Información*, 29(3), 2020.
- [206] Morgan Marietta and David C Barker. *One nation, two realities: Dueling facts in American democracy*. Oxford University Press, 2019.
- [207] Jon Roozenbeek and Sander Van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1):1–10, 2019.
- [208] J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. *Types, sources, and claims of COVID-19 misinformation*. PhD thesis, University of Oxford, 2020.
- [209] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22:100104, 2021.
- [210] Mohan J Dutta-Bergman. Complementarity in consumption of news types across traditional and new media. *Journal of broadcasting & electronic media*, 48(1):41–60, 2004.
- [211] CoMuNe. Covid-19 infodemics observatory. <https://covid19obs.fbk.eu/#/>, COVID-19 Infodemics Observatory, 2 2020. [Accessed 13-Jan-2024].
- [212] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- [213] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, 23(3):337–343, 1977.

- [214] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [215] Aaron D Wyner and Jacob Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258, 1989.
- [216] Yun Gao, Ioannis Kontoyiannis, and Elie Bienenstock. Lempel-ziv and ctw entropy estimators for spike trains. In *Estimation of entropy Workshop, Neural Information Processing Systems Conference (NIPS)*, 2003.
- [217] Rich Ling. Confirmation bias in the era of mobile news consumption: the social and psychological dimensions. *Digital Journalism*, 8(5):596–604, 2020.
- [218] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, 13(9):e0203958, 2018.
- [219] Distribution of twitter users worldwide as of january 2023, by gender (statista, 2023). <https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/>. [Accessed 25-Feb-2024].
- [220] Distribution of twitter users worldwide as of april 2021, by age group (statista, 2021). <https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/>. [Accessed 25-Feb-2024].
- [221] Seth C Kalichman. *Denying AIDS: Conspiracy theories, pseudoscience, and human tragedy*. Springer, 2009.
- [222] Yla Tausczik, Kate Faasse, James W Pennebaker, and Keith J Petrie. Public anxiety and information seeking following the h1n1 outbreak: blogs, newspaper articles, and wikipedia visits. *Health communication*, 27(2):179–185, 2012.
- [223] Joachim Allgaier and Anna Lydia Svalastog. The communication aspects of the ebola virus disease outbreak in western africa—do we need to counter one, two, or many epidemics? *Croatian medical journal*, 56(5):496, 2015.
- [224] Rasmus Nielsen, Richard Fletcher, Nic Newman, J Brennen, and Philip Howard. *Navigating the ‘infodemic’: How people in six countries access and rate news and information about coronavirus*. Reuters Institute for the Study of Journalism, 2020.

- [225] Aengus Bridgman, Eric Merkley, Peter John Loewen, Taylor Owen, Derek Ruths, Lisa Teichmann, and Oleg Zhilin. The causes and consequences of covid-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, 1(3), 2020.
- [226] Sahil Loomba, Alexandre De Figueiredo, Simon J Piatek, Kristen De Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.
- [227] Daniel Romer and Kathleen Hall Jamieson. Conspiracy theories as barriers to controlling the spread of covid-19 in the us. *Social science & medicine*, 263:113356, 2020.
- [228] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. Covid-19-related infodemic and its impact on public health: A global social media analysis. *The American journal of tropical medicine and hygiene*, 103(4):1621, 2020.
- [229] Laura Burbach, Patrick Halbach, Martina Ziefle, and André Calero Valdez. Bubble trouble: strategies against filter bubbles in online social networks. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Healthcare Applications: 10th International Conference, DHM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21*, pages 441–456. Springer, 2019.
- [230] Leticia Bode and Emily K Vraga. See something, say something: Correction of global health misinformation on social media. *Health communication*, 33(9):1131–1140, 2018.
- [231] Kate Starbird. Disinformation’s spread: bots, trolls and all of us. *Nature*, 571(7766):449–450, 2019.
- [232] W Lance Bennett and Steven Livingston. The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication*, 33(2):122–139, 2018.

- [233] Edda Humprecht, Frank Esser, and Peter Van Aelst. Resilience to online disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics*, 25(3):493–516, 2020.
- [234] Jay J Van Bavel, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman, et al. Using social and behavioural science to support covid-19 pandemic response. *Nature human behaviour*, 4(5):460–471, 2020.
- [235] Maria da Penha de Andrade Abi Harb, Lena Veiga E Silva, Nandamudi Lankalpalli Vijaykumar, Marcelino Silva da Silva, and Carlos Renato Lisboa Francês. An analysis of the deleterious impact of the infodemic during the covid-19 pandemic in brazil: a case study considering possible correlations with socioeconomic aspects of brazilian demography. *International Journal of Environmental Research and Public Health*, 19(6):3208, 2022.
- [236] Gordon Pennycook and David G Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019.
- [237] Michael V Bronstein, Gordon Pennycook, Adam Bear, David G Rand, and Tyrone D Cannon. Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of applied research in memory and cognition*, 8(1):108–117, 2019.
- [238] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10, 2020.
- [239] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*, 2020.
- [240] Albert KM Chan, Chris P Nickson, Jenny W Rudolph, Anna Lee, and Gavin M Joynt. Social media for rapid knowledge dissemination: early experience from the covid-19 pandemic. *Anaesthesia*, 75(12):1579, 2020.

-
- [241] Philipp Wicke and Marianna M Bolognesi. Framing covid-19: How we conceptualize and discuss the pandemic on twitter. *PloS one*, 15(9):e0240010, 2020.
- [242] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2):e19273, 2020.
- [243] Amir Masoud Forati and Rina Ghose. Geospatial analysis of misinformation in covid-19 related tweets. *Applied Geography*, 133:102473, 2021.
- [244] Ali Unlu, Sophie Truong, Nitin Sawhney, Jonas Sivelä, and Tuukka Tammi. Tracing the dynamics of misinformation and vaccine stance in finland amid covid-19. *Information, Communication & Society*, 28(2):193–217, 2025.
- [245] Victor Suarez-Lledo, Esther Ortega-Martin, Jesus Carretero-Bravo, Begoña Ramos-Fiol, Javier Alvarez-Galvez, et al. Unraveling the use of disinformation hashtags by social bots during the covid-19 pandemic: Social networks analysis. *JMIR infodemiology*, 5(1):e50021, 2025.
- [246] Ali Unlu, Sophie Truong, Nitin Sawhney, and Tuukka Tammi. Unveiling the veiled threat: the impact of bots on covid-19 health communication. *Social Science Computer Review*, page 08944393241275641, 2024.
- [247] Veda C Storey and Daniel E O’Leary. Text analysis of evolving emotions and sentiments in covid-19 twitter communication. *Cognitive Computation*, 16(4):1834–1857, 2024.
- [248] Chao Shen, Zhenyu Song, Pengyu He, Limin Liu, and Zhenyu Xiong. Online rumors during the covid-19 pandemic: co-evolution of themes and emotions. *Frontiers in Public Health*, 12:1375731, 2024.
- [249] Amreen Batool and Yung-Cheol Byun. Enhanced sentiment analysis and topic modeling during the pandemic using automated latent dirichlet allocation. *IEEE Access*, 12:81206–81220, 2024.
- [250] Farah Alshanik, Rawand Khasawneh, Alaa Dalky, Ethar Qawasmeh, et al. Unveiling topics and emotions in arabic tweets surrounding the covid-19 pandemic: Topic modeling and sentiment analysis approach. *JMIR infodemiology*, 5(1):e53434, 2025.

- [251] Alp Aydınlar, Arda Mavi, Ece Kütükçü, Elçim Elgün Kırmılı, Deniz Alış, Ata Akın, and Levent Altıntaş. Awareness and level of digital literacy among students receiving health-based education. *BMC Medical Education*, 24(1):38, 2024.
- [252] Raffael Heiss, Leticia Bode, Zar Motik Adisuryo, Livia Brito, Ana Cuadra, Peng Gao, Yi Han, Megan Hearst, Kexin Huang, Andrea Kinyua, et al. Debunking mental health misperceptions in short-form social media videos: An experimental test of scientific credibility cues. *Health communication*, 39(13):3059–3071, 2024.
- [253] Jay J Van Bavel and Andrea Pereira. The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, 22(3):213–224, 2018.
- [254] Jay J Van Bavel, Shana Kushner Gadarian, Eric Knowles, and Kai Ruggeri. Political polarization and health. *Nature Medicine*, pages 1–9, 2024.
- [255] Pekka Räsänen, Aki Koivula, and Teo Keipi. Online information seeking patterns and social inequality in a digital economy. *The International Review of Retail, Distribution and Consumer Research*, 31(2):211–228, 2021.
- [256] Mohamed Mostagir and James Siderius. Social inequality and the spread of misinformation. *Management Science*, 69(2):968–995, 2023.
- [257] Daniel C Hallin and Paolo Mancini. *Comparing media systems: Three models of media and politics*. Cambridge university press, 2004.
- [258] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [259] Etienne Becht, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Evaluation of umap as an alternative to t-sne for single-cell data. *BioRxiv*, page 298430, 2018.
- [260] Anna Bertani, Valeria Mazzeo, and Riccardo Gallotti. Decoding the news media diet of disinformation spreaders. *Entropy*, 26(3):270, 2024.
- [261] Douglass C North. *Institutions, institutional change and economic performance*. Cambridge university press, 1990.
- [262] Daron Acemoglu, Simon Johnson, and James A Robinson. Institutions as a fundamental cause of long-run growth. *Handbook of economic growth*, 1:385–472, 2005.

- [263] Peter A Hall and David Soskice. Varieties of capitalism. *Emerging trends in the social and behavioral sciences*, 2009.
- [264] Christian Welzel and Ronald Inglehart. Democratization as the growth of freedom: The human development perspective. *Japanese Journal of Political Science*, 6(3):313–343, 2005.
- [265] Jennifer McCoy, Tahmina Rahman, and Murat Somer. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American behavioral scientist*, 62(1):16–42, 2018.
- [266] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. Cross-country trends in affective polarization. *Review of Economics and Statistics*, 106(2):557–565, 2024.
- [267] Kwanho Kim, Chul-joo Lee, Jennifer Ihm, and Yunjin Kim. A comprehensive examination of association between belief in vaccine misinformation and vaccination intention in the covid-19 context. *Journal of Health Communication*, 27(7):495–509, 2022.
- [268] Sun Kyong Lee, Juhyung Sun, Seulki Jang, and Shane Connelly. Misinformation of covid-19 vaccines and vaccine hesitancy. *Scientific reports*, 12(1):13681, 2022.
- [269] Qinfeng Zhu, Tai-Quan Peng, and Xinzhi Zhang. How do individual and societal factors shape news authentication? comparing misinformation resilience across hong kong, the netherlands, and the united states. *The International Journal of Press/Politics*, page 19401612251318838, 2025.
- [270] Toni GLA Van Der Meer and Michael Hameleers. Fighting biased news diets: Using news media literacy interventions to stimulate online cross-cutting media exposure patterns. *New Media & Society*, 23(11):3156–3178, 2021.
- [271] S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. Does media literacy help identification of fake news? information literacy helps, but other literacies don’t. *American behavioral scientist*, 65(2):371–388, 2021.
- [272] Carlos Rodríguez-Pérez and María José Canel. Exploring european citizens’ resilience to misinformation: Media legitimacy and media trust as predictive variables. *Media and Communication*, 11(2):30–41, 2023.
- [273] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

- [274] Andrew M Guess. (almost) everything in moderation: new evidence on americans' online media diets. *American Journal of Political Science*, 65(4):1007–1022, 2021.
- [275] Diana C Mutz. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press, 2006.
- [276] Jenna Bednar. Polarization, diversity, and democratic robustness. *Proceedings of the National Academy of Sciences*, 118(50):e2113843118, 2021.
- [277] Daniel P Aldrich. *Building resilience: Social capital in post-disaster recovery*. University of Chicago Press, 2012.
- [278] Jonathan Mellon and Christopher Prosser. Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3):2053168017720008, 2017.
- [279] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
- [280] Taylor Jim. The woman who founded the 'incel' movement. <https://www.bbc.com/news/world-us-canada-45284455>, 8 2018. [Accessed 20-Sept-2025].
- [281] Debbie Ging. Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and masculinities*, 22(4):638–657, 2019.
- [282] Roberta Liggett O'Malley, Karen Holt, and Thomas J Holt. An exploration of the involuntary celibate (incel) subculture online. *Journal of interpersonal violence*, 37(7-8):NP4981–NP5008, 2022.
- [283] Alyssa Maryn, Jordan Keough, Ceilidh McConnell, and Deinera Exner-Cortens. Identifying pathways to the incel community and where to intervene: A qualitative study with former incels. *Sex Roles*, 90(7):910–922, 2024.
- [284] Sylvia Jaki, Tom De Smedt, Maja Gwózdź, Rudresh Panchal, Alexander Rossa, and Guy De Pauw. Online hatred of women in the incels. me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2):240–268, 2019.
- [285] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM conference on web science*, pages 87–96, 2019.

-
- [286] Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3197–3207, 2022.
- [287] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020.
- [288] Emily Klein and Jennifer Golbeck. A lexicon for studying radicalization in incel communities. In *Proceedings of the 16th ACM Web Science Conference*, pages 262–267, 2024.
- [289] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.
- [290] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017.
- [291] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013.
- [292] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.
- [293] Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545, 2020.
- [294] Claire Wardle. Understanding information disorder. <https://firstdraftnews.org/long-form-article/understanding-information-disorder/>, First Draft News, 9 2020. [Accessed 10-Oct-2023].
- [295] Economist Intelligence Unit. Democracy index 2020. <https://www.eiu.com/n/campaigns/democracy-index-2020/>, 2020. [Accessed 22-May-2023].

- [296] United Nations. World population prospects 2022. <https://population.un.org/wpp/Download/Standard/MostUsed/>, 2022. [Accessed 22-May-2023].
- [297] Freedom House. Freedom on the net 2020: The pandemic’s digital shadow. <https://freedomhouse.org/report/freedom-net/2020/pandemics-digital-shadow>, 2020. [Accessed 22-May-2023].
- [298] The global disinformation index. <https://www.disinformationindex.org/about>. Accessed 13-May-2023.
- [299] Institute for Economics and Peace. Global terrorism index 2020: Measuring the impact of terrorism. <https://visionofhumanity.org/wp-content/uploads/2020/11/GTI-2020-web-1.pdf>, 2020. [Accessed 22-May-2023].
- [300] International Monetary Fund. World economic outlook database. <https://www.imf.org/en/Publications/WE0/weo-database/2022/October>, 2022. [Accessed 22-May-2023].
- [301] OECD. Tertiary graduation rate (indicator). <https://data.oecd.org/students/tertiary-graduation-rate.htm#indicator-chart>, 2023. [Accessed 22-May-2023].
- [302] OECD. Programme for international student assessment - 2018 results. <https://www.oecd.org/pisa/publications/pisa-2018-results.htm>, 2019. [Accessed 22-May-2023].
- [303] OECD. Social security contributions (indicator). <https://data.oecd.org/tax/social-security-contributions.htm>, 2023. [Accessed 22-May-2023].
- [304] OECD. Social spending (indicator). <https://data.oecd.org/social-exp/social-spending.htm>, 2023. [Accessed 22-May-2023].
- [305] OECD. Trust in government (indicator). <https://data.oecd.org/gga/trust-in-government.htm>, 2023. [Accessed 22-May-2023].
- [306] Reporters Without Borders. 2020 world press freedom index: “entering a decisive decade for journalism, exacerbated by coronavirus. <https://rsf.org/en/2020-world-press-freedom-index-entering-decisive-decade-journalism-exacerbated-coronavirus>, 2020. [Accessed 22-May-2023].

- [307] United Nations. E-government survey 2020. digital government in the decade of action for sustainable development. <https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2020>, 2020. [Accessed 22-May-2023].
- [308] V-Dem Institute. The v-dem dataset. <https://www.v-dem.net/data/the-v-dem-dataset/>, 2022. [Accessed 22-May-2023].
- [309] World Bank. Individuals using the internet (% of population) (indicator). <https://data.worldbank.org/indicator/IT.NET.USER.ZS>, 2023. [Accessed 22-May-2023].
- [310] World Bank. Gini index (indicator). <https://data.worldbank.org/indicator/SI.POV.GINI>, 2020. [Accessed 22-May-2023].

Appendix A

Supplementary Information Chapter 3

A.1 Data

We collected social media data about the large massive Vaccination Campaign against Covid-19 through special Twitter’s endpoint dedicated to COVID-19 research between August 31, 2020 and July 15, 2021, from the announcement of the availability of the first Covid-19 vaccines up to the peak of the vaccine campaign in Europe and in the United States. Our dataset consists of 9.6 million of messages posted on Twitter from 3 million of unique users. We focused our attention on data captured by the filter of the Twitter Firehose on Covid-19 in 18 among the most represented languages on Twitter. More specifically, we focused on terms related to the vaccination, anti-vax campaigns but also to the most known vaccine brands, such as Pfizer, Astrazeneca, Johnson&Johnson, Moderna, Sputnik V.

We collected online discussions where non specific contentious topic dominates consisting of more than 10 million of messages posted on Twitter from 6 million of unique users. This dataset covers only few days, from the 21/4/2021 to the 24/4/2021, since our interest was to capture any message on Twitter without filtering for specific keywords. In this way, we collected social media posts that represent a sort of baseline sample of the ordinary discussion found on Twitter. Thus, we have compared our analysis about a specific polarizing issue, such as the massive pharmaceutical intervention campaign against Covid-19 with a baseline sample taken as a benchmark of online discussions.

A.2. Emotional Analysis

Language	Subjects		Verbs	Participes	Root	Anti-vax
English	vaccine(s)	vaccination(s)	vaccine	vaccinated	vaccin	anti-vax
Italian	vaccino(i)	vaccinazione(i)	vaccinare	vaccinato	vaccin	no-vax
Spanish	vacuna(s)	vacunación/vacunas	vacunar	vacunado	vacun	anti-vacunas
Catalan	vacuna(s)	vacunació/vacunes	vacunar	vacunat	vacun	
Portuguese	vacina(s)	vacinação/ões	vacinar	vacinado	vacin	anti-vacinas
French	vaccin (s)	vaccination(s)	vacciner	vacciné	vaccin	anti-vaccins
German	Impfstoff (e)	Impfung(en)	impfen	geimpft/ geimpften	impf	anti-impf
Russian	вакцина	Вакцинация/прививки	вакцинировать	вакцинированный	вакцина	антивакс
Polish	szczepionka/i	szczepienie	szczepić	szczepione	szczepi	Anty-szczepionki
Danish	vaccine (r)	vaccination (er)	vaccinere	vaccineret	vaccin	
Swedish	vaccin(er)	vaccinering	vaccinera	vaccinerad	vacciner	Vaccinationsmotstånd
Turkish	aşı(lar)	aşılama	aşılamak	aşılanmış	aşı	aşı karşıtı
Indonesian	vaksin	vaksinasi	divaksinasi	divaksinasi	vaksin	
Japanese	ワクチン	予防接種	予防接種	予防接種		
Thai	วัคซีน	การฉีดวัคซีน	ฉีดวัคซีน	ฉีดวัคซีน		
Korean	백신	백신 접종	예방 접종	예방 접종		안티 백신
Arabic		مصل	تلقيح	تلح	تطعيم	
Hindi	टीका (टीके)	टीका (टीकाकरण)	टीका लगाना	टीका		एंटी टीका लगाना

Figure A.1: **Terms about vaccine discourse.** Terms from the 18 most used language on Twitter, used to filter the Twitter Firehose on Covid-19 for this analysis. The terms are related to vaccination. Additionally, we considered also the following terms related to the vaccine names: astrazeneca, moderna, johnson & johnson, sputnik, pfizer, biontech.

A.2 Emotional Analysis

We wanted to investigate how emotional responses change during the large-scale Covid-19 vaccination campaign with respect to a baseline sample where no contentious topic dominates. For this reason, we tested different algorithms able to capture the emotional responses of the messages posted on Twitter. Initially, we focused on the well-known Valence- Arousal-Dominance model [128]. This algorithm does not perform particularly well on our dataset, since the three dimension of emotions are highly correlated among them. We decided to adopt another algorithm that classifies the emotions in eight categories; four of which with a positive valence (Joy, Trust, Anticipation, Surprise) and four with a negative valence (Fear, Disgust, Anger, Sadness). This allows us to better capture the variability and differences between each emotion in our case study. In some of our analysis, we decided to show the emotional distribution normalized in a range between -1 (completely negative valence) and +1 (completely positive valence). In order to do so, we found the total emotional value of the messages posted on Twitter by summing each singular emotion. Then, we divided each emotion by the total emotional value found and we grouped the positive emotions (Anticipation, Surprise,

Joy and Trust) and the negative emotions (Sadness, Anger, Fear, Disgust) into two different categories (positive vs negative emotions) and finally we normalized the data in a range between -1 and +1.

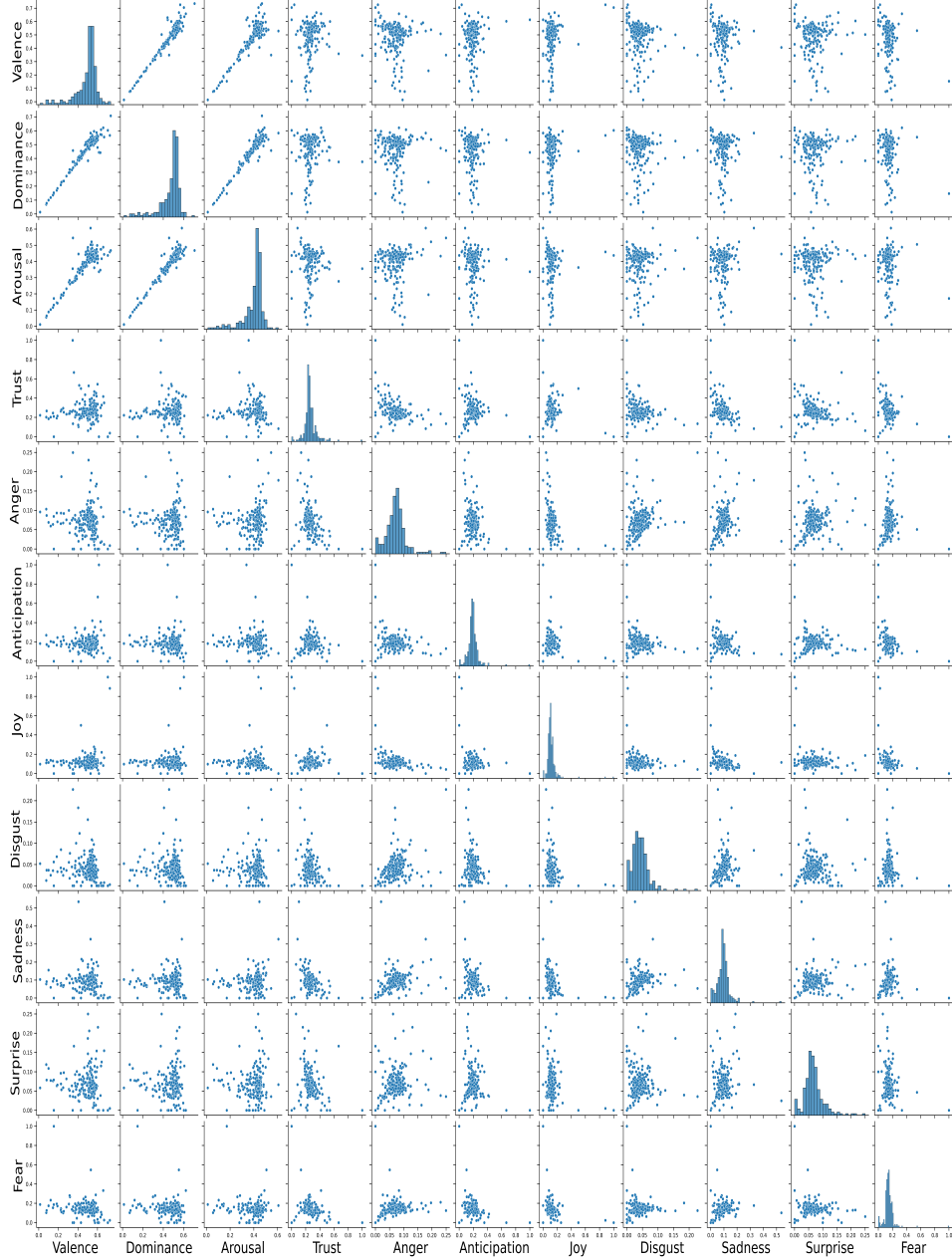


Figure A.2: Correlation among Valence-Arousal-Dominance emotional dimension and the eight emotions. For each emotions, we calculated the Spearman correlation, finding that the three VAD emotions are highly correlated among them, while the four positive and four negative emotions are not correlated with respect to each singular emotion.

A.3. Type of Users Accounts

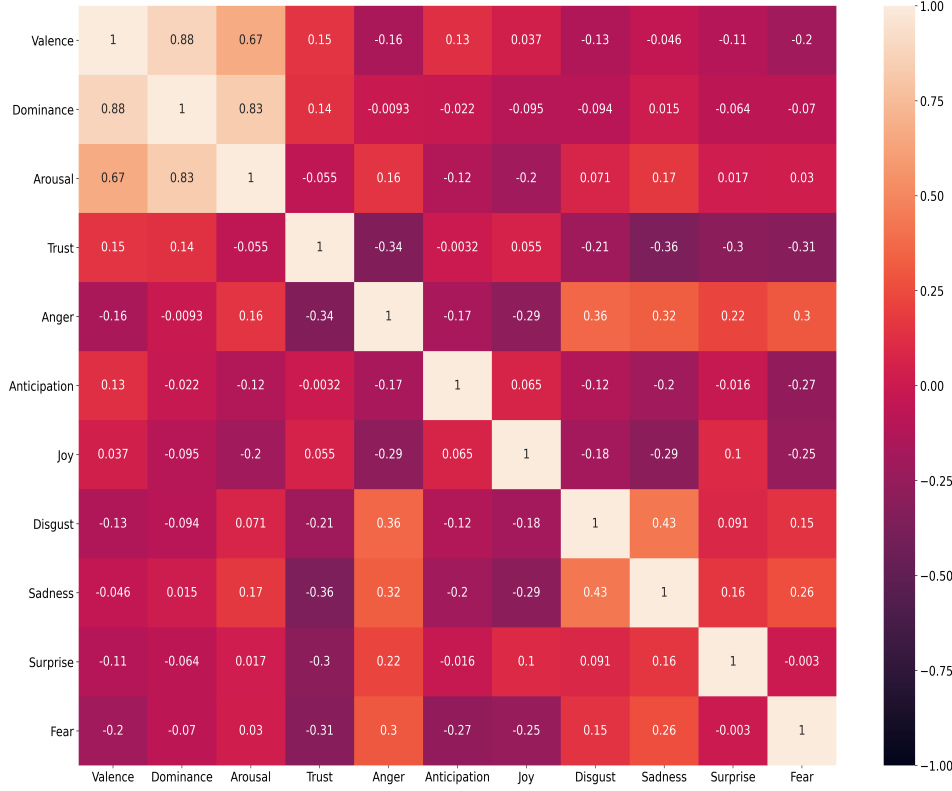


Figure A.3: **Heatmap showing the Spearman Correlation among Valence-Arousal-Dominance emotional dimension and the eight emotions.** We show the Spearman Correlation against each singular emotions (Joy, Surprise, Trust, Anticipation, Fear, Anger, Disgust, Sadness) and the three VAD emotions. We find that the the VAD emotions are highly correlated among them with values of 0.88 among Dominance and Valence, 0.83 among Dominance and Arousal, 0.67 among Valence and Arousal.

A.3 Type of Users Accounts

In our research, we were also interested in the emotional response compared to the baseline, especially when we take into account characteristics of the users and the type of information shared in the online platform. In particular, a dimension that might elicit different emotional response is the public status of the accounts (verified vs unverified accounts). A.6 shows the emotional distribution of verified and unverified accounts (A) and the emotional responses of the verified and unverified users as distinguished between bots and humans. Discerning bots from humans is usually carried out by observing the social behaviour of the users in online social systems. In particular, to distinguish a bot

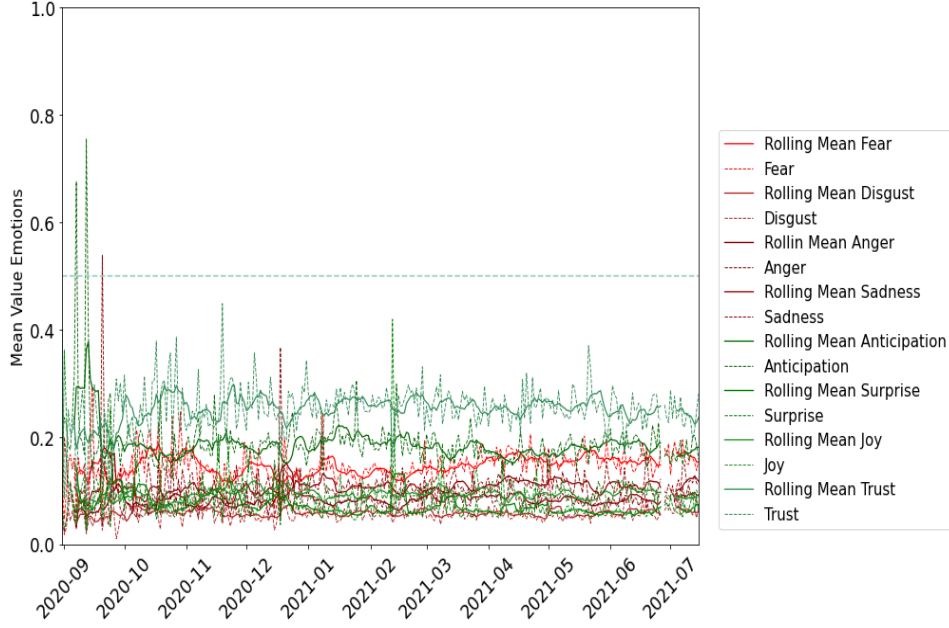


Figure A.4: **Distribution of positive and negative emotions during the entire phase of Covid-19 Vaccination Campaign.** We show the average distribution per day of each emotions. In particular, the emotions with a positive valence (Joy, Trust, Anticipation, Surprise) are represented by different scale of green, whereas the emotions with a negative valence (Fear, Anger, Disgust, Sadness) are drawn by different scale of red. Based on this plot, we decided to normalize the data by summing the value of each emotion and then divided each singular emotion by the total emotional value. Separately, we summed the positive and the negative emotions and then we normalized each new category in a range between -1 to +1 in order to have a more comparable emotional value for each message posted on Twitter.

from a human, we chose the criteria proposed by [294]. More specifically, automated accounts are usually associated to some forms of unusual behavior, either for the great amount of contents created or the frequency at which it created, as discussed in the supplementary material of [114] and also in [117] and [120]. However, in this case study, we are not mainly interested in the differences between bot and human behaviours, but rather in the differences between verified and unverified users within each category.

A.4 Statistical Analysis

For this work, we used several statistical Python libraries in order to perform our analysis. In particular, we used the Scipy python library that provides several algorithms for

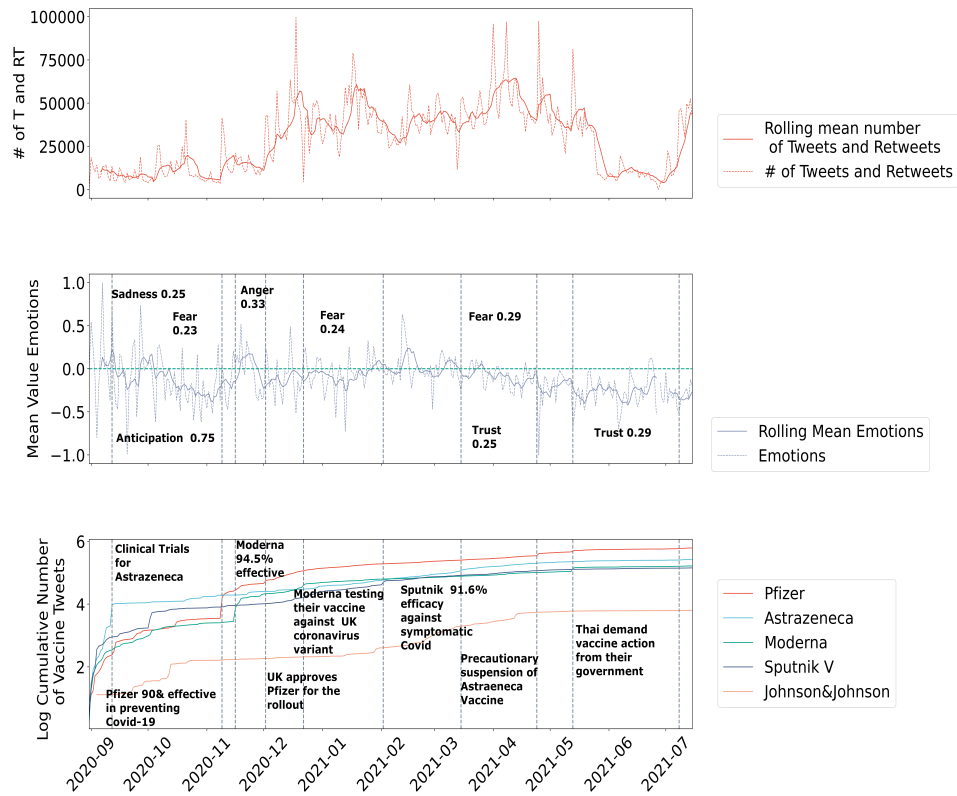


Figure A.5: Evolution of the messages, emotions and events during the Covid-19 Vaccination Campaign. In the first panel, we quantified the distribution of tweets and retweets posted on Twitter during the entire phase of Covid-19 Vaccination Campaign. The second panel shows the distribution of emotions in a range between -1 (completely negative valence) and +1 (completely positive valence). We calculated the mean emotion for those emotional dimensions that have reach values higher than 0.25 in particular days. The third panel indicates the cumulative logarithmic number of messages with reference to particular vaccines names. Labels refer to manually inspected events about the massive COVID-19 vaccination campaign.

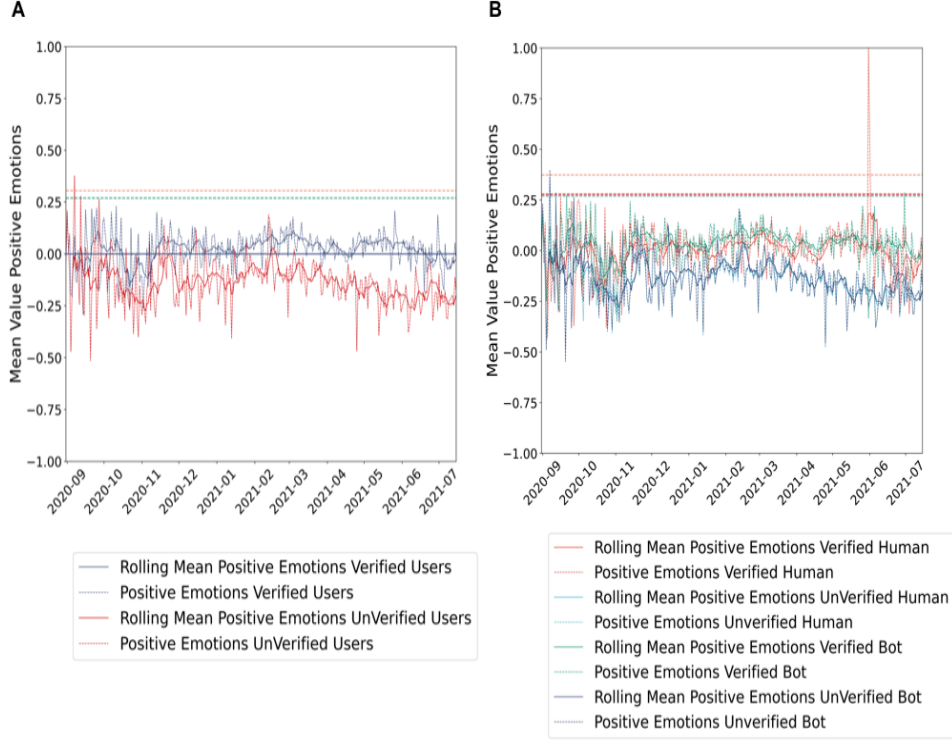


Figure A.6: **Emotional Analysis for different type of users accounts.** We quantified (A) emotional trends from -1 to +1 for verified/unverified users, (B) for verified and unverified human, verified and unverified bot. We have shown both the mean distribution per day and the rolling mean emotions in order to better visualize the frequent variation of the emotional distribution with respect to each type of users accounts. Especially in the second panel, we observe that the differences between the four categories of users tend to stabilize as the vaccination campaign unfolds, whereas in the very first months the responses are less predictable.

optimization, integration, interpolation, statistics and many other classes of problems. Among all the analysis that we performed, the Figure 4 might need a more precise and deeper research material. To find the distribution of messages across the emotional range, we use a joint probability distribution function, in order to quantify the distribution of tweets across the emotional range (-1, +1), when we consider the political leaning associated with the link shared and the type of users (verified, unverified). In particular, we grouped the data into 11 bins, which represent the emotional range (-1, +1). For the purpose of simplicity, we also multiplied each value for 100 in the panels related to our two samples in order to round our results, but we have shown the original values in each colorbar. The last panel shows the percentage difference between the vaccine and baseline sample.

Appendix B

Supplementary Information for Chapter 4

Data Collection

We collected data through Twitter API by retrieving all the messages that contained at least one relevant keywords related to the long Covid Discussion, consisting in total of 12 million of tweets and 1.7 million of unique users. The list of the keywords used to retrieve the data are the following: "long covid" and #longCovid.

Emotional Analysis

As explained in the main text, we used an algorithm to detect eight emotions, four of which with a negative valence (Sadness, Fear, Disgust and Anger) and four with a positive valence (Trust, Anticipation, Surprise, and Joy). We quantified the distribution of each emotion both for the long Covid dataset and the baseline sample. The Long Covid sample is characterized by having a higher value of Anticipation compared to all the other emotions and with respect to the baseline sample, as shown by Figure B.1.

We quantified the emotional responses with respect of different types of news (see Figure B.2), classified as reliable, unreliable and news not containing any web-domains. Our result suggests that unreliable news media source are more skewed towards the positive valence for long Covid sample, while the different news (reliable and unreliable) as well as the tweets not containing any urls show the same distribution with the median 0 for all the categories considered.

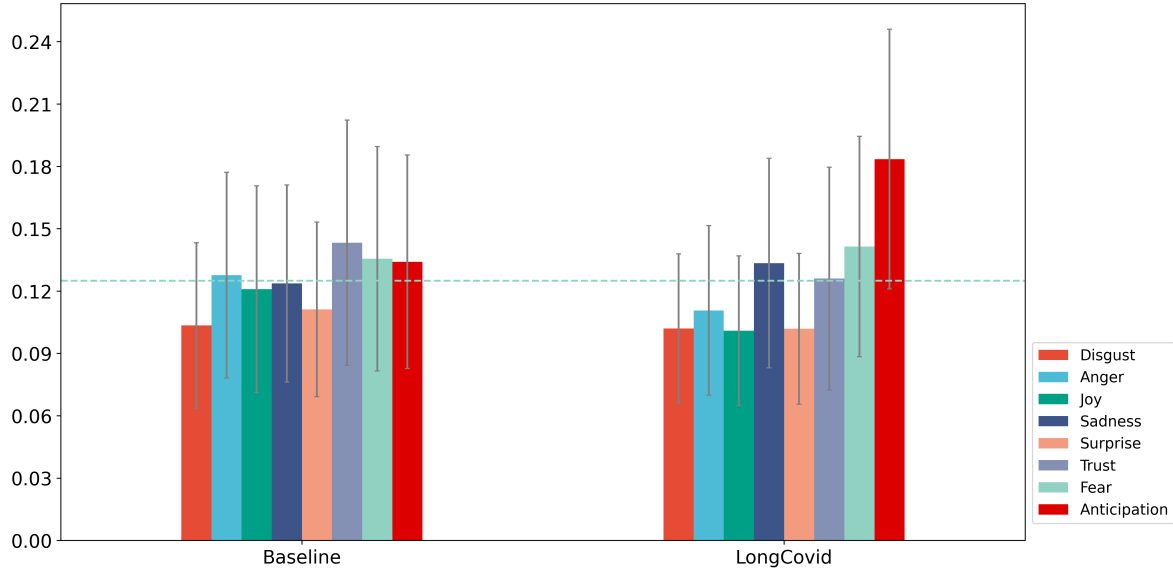


Figure B.1: **Emotion Mean Value.** For each sample of data (LongCovid and baseline), we quantified the mean value of each emotion: Anger, Sadness, Surprise, Fear, Anticipation, Trust, Disgust, Joy are considered. The horizontal line is placed on the 0.125 value to better visualize the difference between the two samples of data.

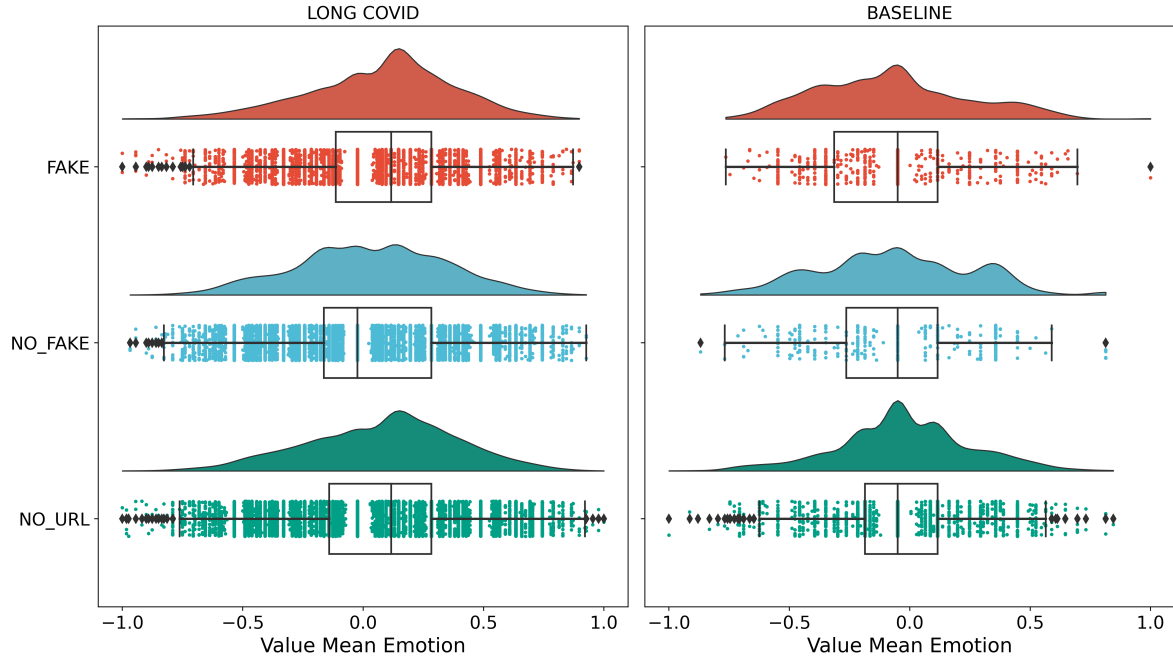


Figure B.2: **Emotion Distribution for different type of news.** For each sample (LongCovid vs baseline), we quantified the distribution of emotional values from -1 to 1, respectively for tweets containing fake news, with no fake news and without a link URL.

Emotional Analysis along the Information Cascades

We also used another algorithm which gives a sentiment score between -1 (completely negative valence) to + 1 (positive valence) in order to gain insights into the emotional reactions arise throughout the information cascades. We compared our result with the emotional reactions in the baseline sample, showing that the latter is indeed more skewed towards the positive valence, while the emotions for the long Covid discussion is equally distributed between the positive and negative lines, highlighting how this topic is highly polarized in relation to the emotions aroused.

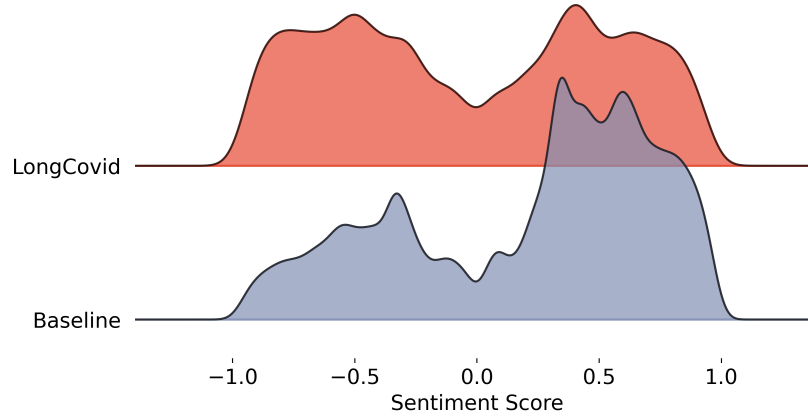


Figure B.3: **Distribution of Sentiment Scores in Information Cascades.** We quantified the sentiment scores both for LongCovid Discussions and Baseline sample.

Metadata Analysis: characteristics of the online communication networks

Similarly to the analysis shown in the main text, our interest was to characterize the online communication by quantified the percentage of the different news categories over time and with respect to the baseline sample. Figure B.4 shows the percentage distribution of news classified as reliable (Mainstream Media and Science) and unreliable sources (Satire, Political, Clickbait, Fake/Hoax, Conspiracy/Junkscience). In this context, we observe that long Covid discussions are characterized by having a higher volume of reliable sources with respect to the baseline sample, as highlighted by the Panel C, showing the cumulative deviation from the baseline.

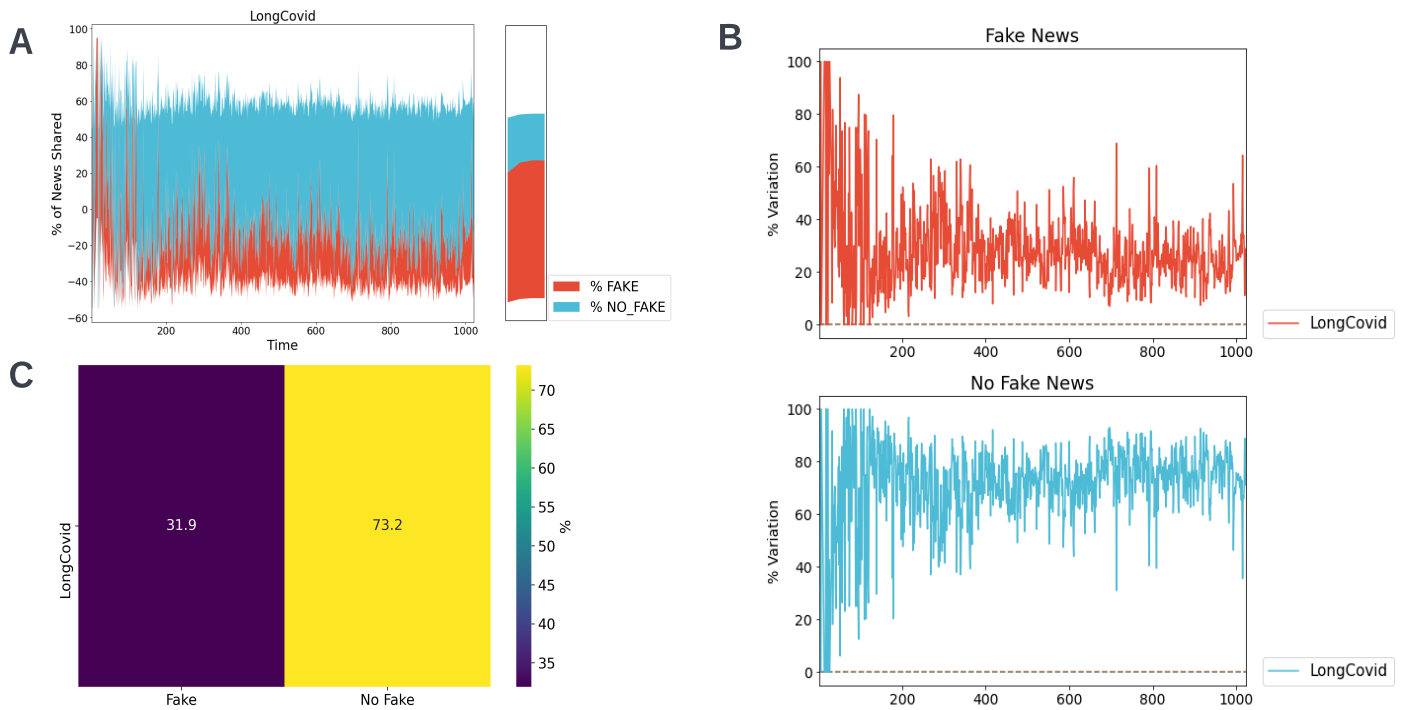


Figure B.4: **Distribution of Reliable and Unreliable News for Long Covid Discussion and Baseline Sample.** We quantified the percentage of reliable and unreliable news for LongCovid Discussion and a baseline sample (panel A). Panel B highlights the percentage variation of the news media categories with respect of the baseline. The Panel C displays the cumulative deviation of the different news with respect to the baseline sample.

We also consider the distribution of political leaning over time, showing that news media categories are characterized by the center leaning, which is one of the most relevant in the long Covid dataset. Interestingly, the right and right-center leanings show spikes in specific moment of the conversation, especially in the first months, as shown by Figure B.5.

Community Analysis

The online communication networks have been analyzed by inspecting and characterizing the different communities founded through the Louvain method. Figure B.6 (panel A) shows the distribution of news media sources for each community considered, while panel B reveals the distribution of political leaning (least-biased, right, right-center, left-center and left) for each community considered.

As mentioned in the main text, we quantified how many users are responsible for how

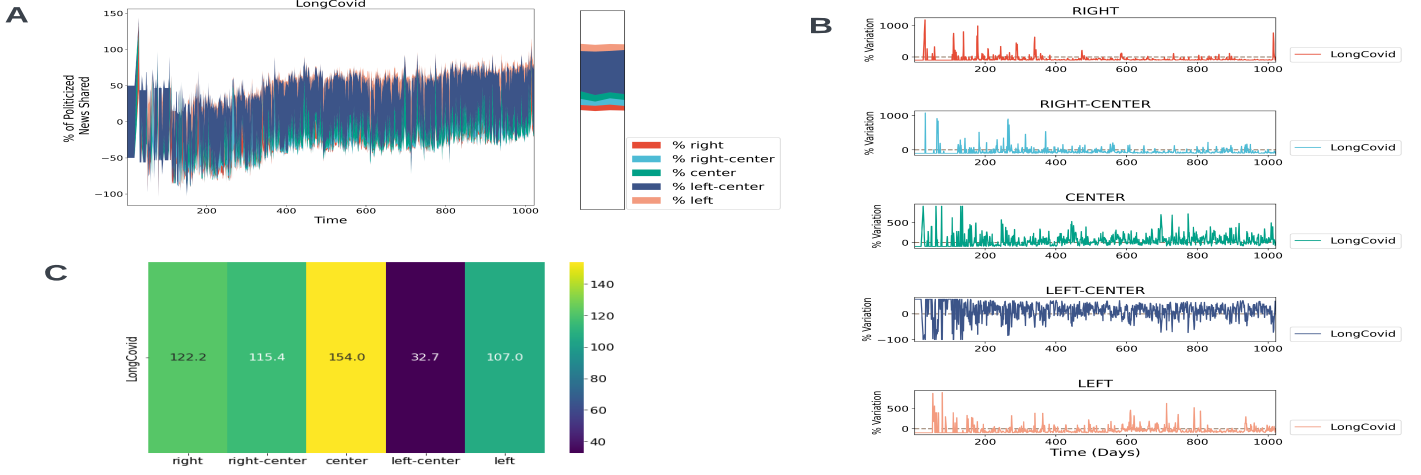


Figure B.5: Distribution of Political Leaning for Long Covid Discussion and Baseline Sample. We quantified the percentage of politicized content for LongCovid Discussion and a baseline sample. Panel B highlights the percentage variation of the politicized content from the baseline. The Panel C displays the cumulative deviation of the different political leaning with respect to the baseline sample.

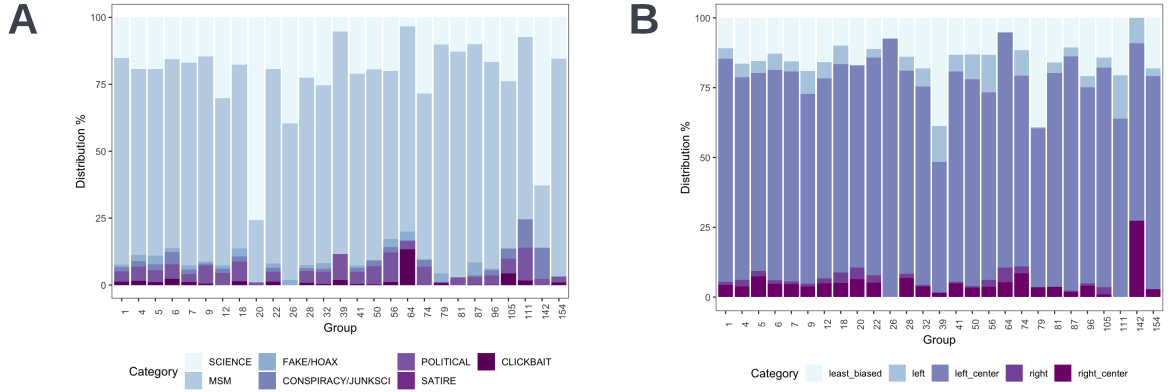


Figure B.6: Communities-level news and political content We identify the largest online communities and, for each community, we show the estimated news distribution and political content.

many news media shared, social activities and politicized content during the evolution of Long Covid discussion (see Figure B.7). The results of our analysis highlight that there is an evident centralization in all the sphere considered: social activities, news but also political centralization, quantified by the Gini Index.

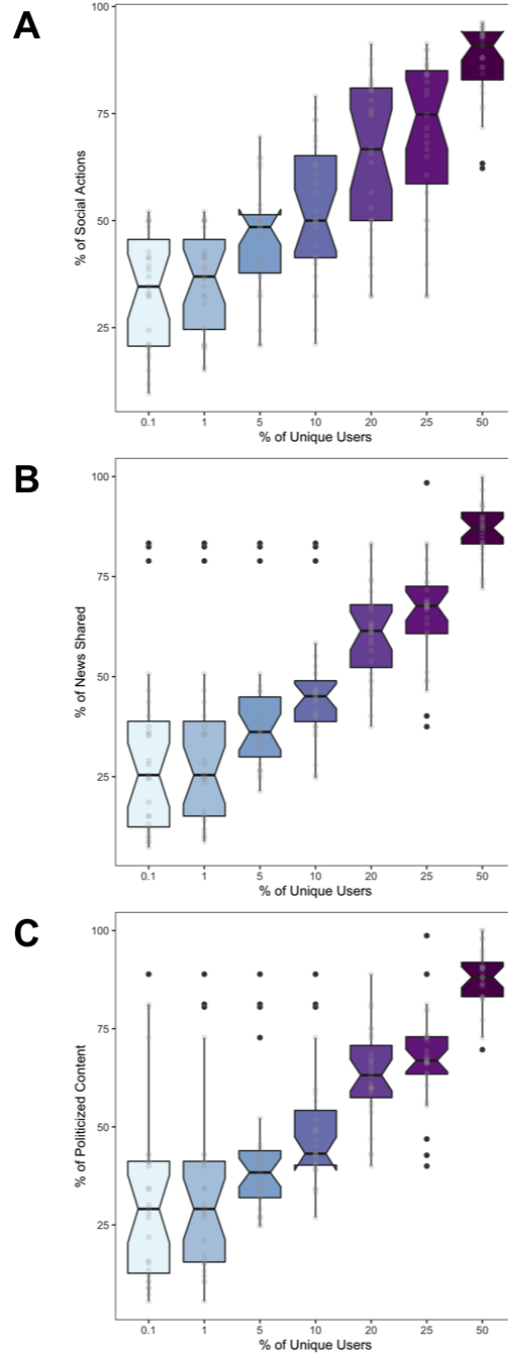


Figure B.7: **Communities Centralization** For each panel (A-C), we quantified (A) the percentage of social activities, (B) the percentage of news shared, (C) the percentage of political content posting on Twitter concerning discussions about Long Covid.

News Media Categorization

As mentioned in the main text, we studied the percentage of the different news media sources over time, the variation compared to the baseline sample, demonstrating that

the news media outlet sharing more information was science B.8.

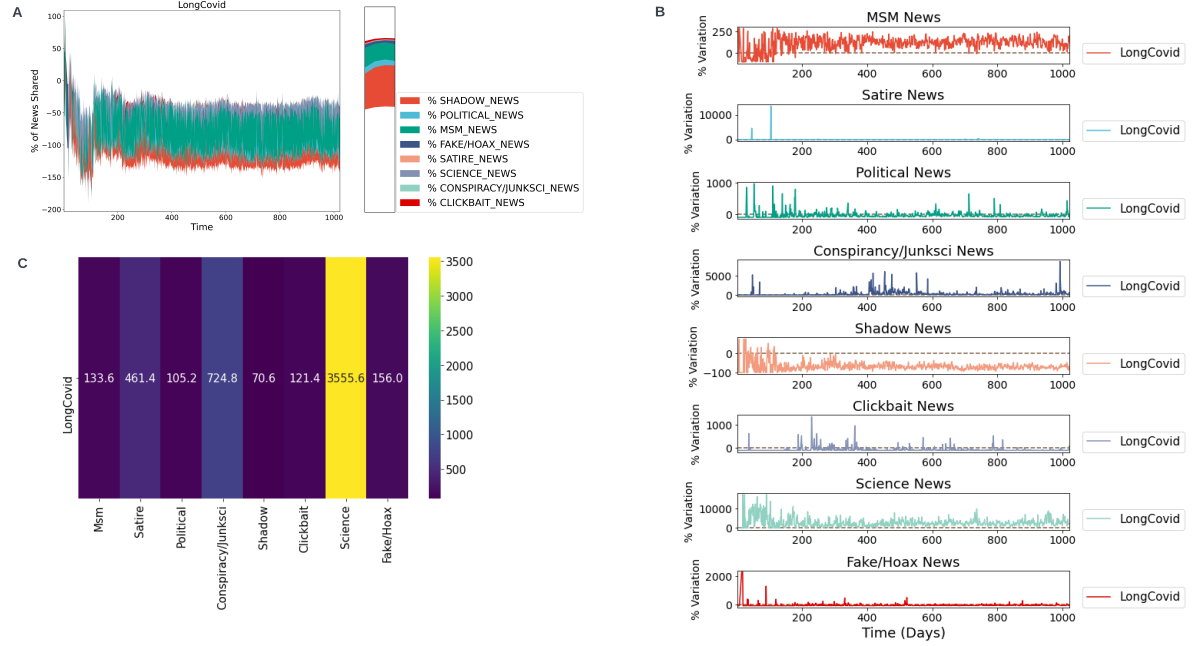


Figure B.8: **Distribution of the different news media sources.** We quantified the distribution of news media sources. We quantified the variation between the volumes of the eight news media classification posted and the baseline sample.

Mainstream Media and Scientific News media outlets

We also investigated several web-domains classified according two of the most relevant news categories: Mainstream Media and Scientific News. We extracted the five domains most frequent for each category, inspecting the distribution of these over time and calculating the variation from the baseline dataset. We observed that both the categories shows spikes of certain web-domains over time.

Dynamic Iri

To better quantify the prevalence of infodemic risk, we applied the Dynamic Infodemic Risk index, which gives us the likelihood that a user endorses or engages with on-line messages pointing to potentially misleading sources, as discussed in [1]. This index quantifies if and how much users interact with information potentially misleading, ranging from 0 to 1. As shown by B.10, we found that the Dynamic Infodemic Risk index

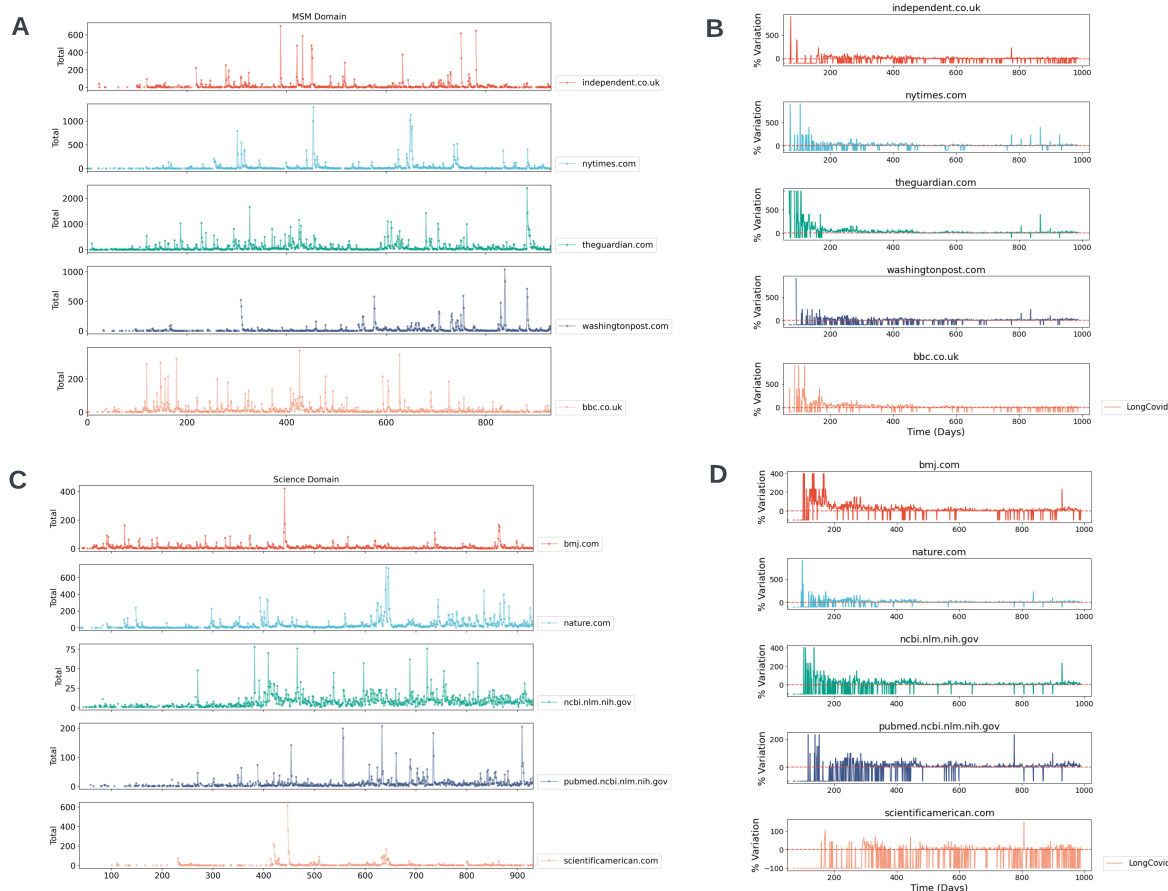


Figure B.9: **Distribution of the most frequent web-domains.** We quantified the distribution of the most frequent web-domains for news classified as science and mainstream media outlets. We quantified the variation between the volumes of those domains posted and the baseline sample.

is relatively low for each year considered. This highlights that, in the case of Long COVID discussions, the type of content users interact with is mainly reliable.

Reliability of the Mainstream and Scientific Sources

To better assess the reliability of the scientific and mainstream sources (used mainly for the reliability index score) we decided to investigate such a classification more deeply. To have a better understanding of the sources, we prompted a zero-shot classification through openAi (using gpt-3.5-turbo) to better classify the different sources within these broad categories. Then, we manually inspected the results. Within the scientific category, domains have been classified as belonging to peer-reviewed journals, pre-print, agencies and scientific communications channels. Examples of scientific jour-

nals are nature.com, sciencedirect.com, plos.org, journals.sagepub.com, jstor.org, while we found only one preprint shared corresponding to researchgate.net. Agencies are mainly represented by reputable health-related organizations, either governmental agencies, international health organizations, or recognized research/medical institutions, examples are who.it, cdc.gov, nih.gov, paho.org as well as apa.org, aap.org. Scientific communication are instead represented by well known scientific magazines such as scientificamerican.com, newscientist.com, nationalgeographic.com, discovermagazine.com, by sites sharing latest news about science such as sciencedaily.com, sciencenews.org, sciencealert.com, other health focused media such as medicalnewstoday.com, medscape.com, webmd.com, healthline.com, and other critical site specialized in debunking pseudoscience and health misinformation such as sciencebasedmedicine.org, skeptical-raptor.com, skepticscience.com. By manually inspecting the different domains in scientific sources, we found that the majority of unique sources are scientific communication journals (67,50 %), followed by peer-reviewed journals (17.50%), agencies (13,25%) and preprint, which represents only 1.25 % of our dataset. With respect to how many times each subcategory has been mentioned, the most prevalent source is peer-reviewed journals (45.15%), followed by scientific communication journals (36,68%), agencies (17,28%) and preprints that have been mentioned only 0.87 % times in our dataset. For this reason, we argue that the scientific sources found in our data are reliable sources coming from peer reviewed journals, websites sharing reliable and scientific information and governmental and international agencies, with a small and negligible portion of preprints. In addition, by using MediaBiasFactChecking data, we also looked at the political leaning of such sources, finding that any of our scientific sources has been classified towards a specific leaning. We conducted the same analysis for the mainstream sources in order to investigate potential biases towards a specific political leaning and we decided to classify our mainstream sources in a more grained way. In particular, we found that the majority of mainstream journals have been classified as center-left, followed by center and center-right content. No one of them has been classified at the extreme levels of political leaning such as left and/or right, as shown in the Figure 2.

With respect to a more detailed classification of the sources, we used a prompt resulting in a zero-classification label for each source. The domains have been classified in one of these subcategories: Mainstream journals (MSM), Think Tank / Policy, Blogs, University / Official News, Agency / Official, Fact-Check / Research. Examples of Mainstream Journals are nytimes.com, voanews.com, bloomberg.com, washingtonpost.com, theguardian.com, globalnews.ca, bbc.in, independent.co.uk, usnews.com, timesofindia.indiatimes.com, npr.org, cbc.ca, cbsnews.com, economist.com. Think Tank

/ Policy are represented by policy and research-oriented sources, such as brookings.edu, urban.org, rand.org, cato.org, cgdev.org, eurasiagroup.net. The category ‘Blogs’ includes general digital media such as vice.com, motherjones.com, spectator.co.uk, prospect-magazine.co.uk, but also websites providing polls such as fivethirtyeight.com, morningconsult.com, and opinion blogs such as medium.com. University and official news encompass education and research-oriented outlets, a mix of academic journalism, university communications, and education-focused media such as nber.org, theconversation.com, news.harvard.edu, daily.jstor.org. insidehighered.com. Agency represents official government websites, major international news agencies such as gouvernement.fr, whitehouse.gov, gao.gov, ansa.it, afp.com, dpa-international.com. Fact-Check / Research represents fact-checking, data transparency, and research communication organizations such as factcheck.org, fullfact.org, mediabiasfactcheck.com, ourworldindata.com, consumerreports.org. As shown by Table 2, the majority of unique sources fall into the category of mainstream journals. The percentage of mentions received by these journals is pretty high, consisting of 95,64. By carefully scrutinizing both scientific and mainstream sources, and using the Media Bias/Fact Check classification along with our granular subcategory labeling, we are confident that our reliability index accurately reflects the proportion of reliable sources shared by each user. This index takes into account only sources that have been correctly classified, ensuring that users’ reliability scores are based on verified, credible information. We acknowledge that this approach does not capture all aspects of amplification, interpretive bias, or narrative framing. However, we explicitly checked for preprints, low-quality, and potentially unreliable sources within the scientific category, finding them to be negligible (preprints accounted for only 1.25% of unique sources and 0.87% of mentions). Similarly, within the mainstream category, sources classified as blogs, opinion sites, or extreme political outlets represent a very small fraction of mentions, confirming that the vast majority of MSM content shared is established, centrist, and reliable.

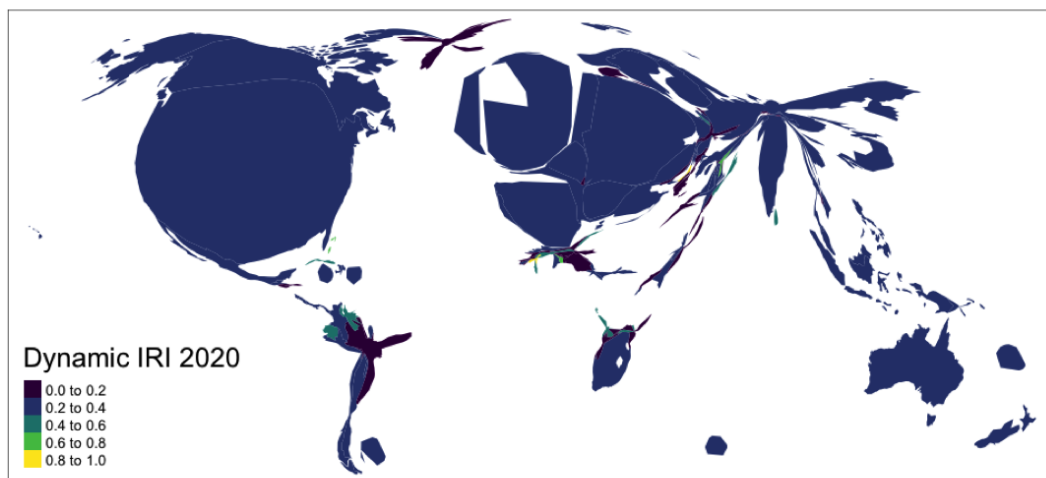
Table B.1: Scientific sources by category. We show the absolute and percentage values of scientific sources classified into one of the following categories: peer-reviewed journal, science communication, agency, preprint.

Category	% Unique Sources (abs.)	% of Mentions (abs.)
Peer-reviewed Journal	17.50% (14 journals)	45.15% (93,929)
Preprint	1.25% (1 journal)	0.87% (1,827)
Agency	13.75% (11 journals)	17.28% (35,949)
Science Communication	67.50% (54 journals)	36.68% (76,298)

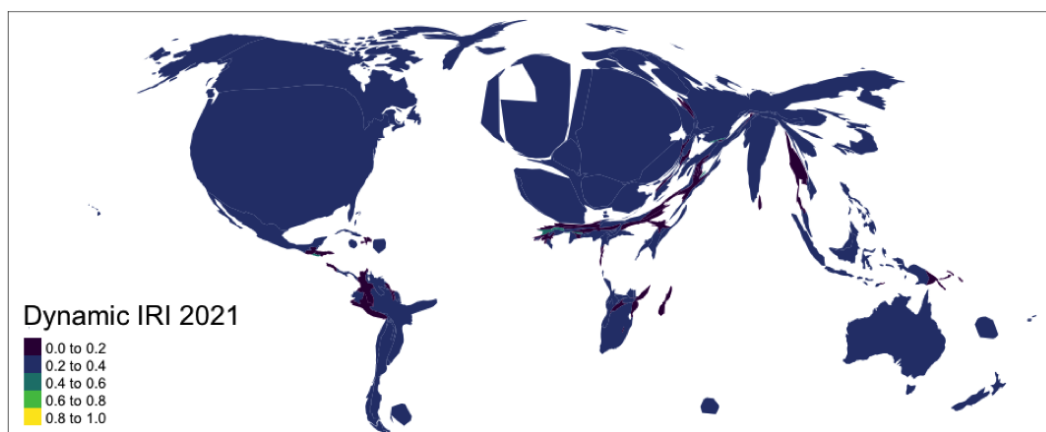
Table B.2: Mainstream Source by Category. We show the distribution of unique sources and mentions according to six subcategories.

Category	% Unique Sources (abs.)	% Sources of Mentions (abs.)
Mainstream (MSM)	75.10% (549)	95.64% (844,384)
Think Tank / Policy	4.50% (33)	1.70% (15,023)
Blogs	15.04% (110)	1.32% (11,701)
University / Official News	1.36% (10)	0.822% (7,258)
Fact Check / Research	1.77% (13)	0.24% (2,173)
Agencies	2.18% (16)	0.25% (2,263)

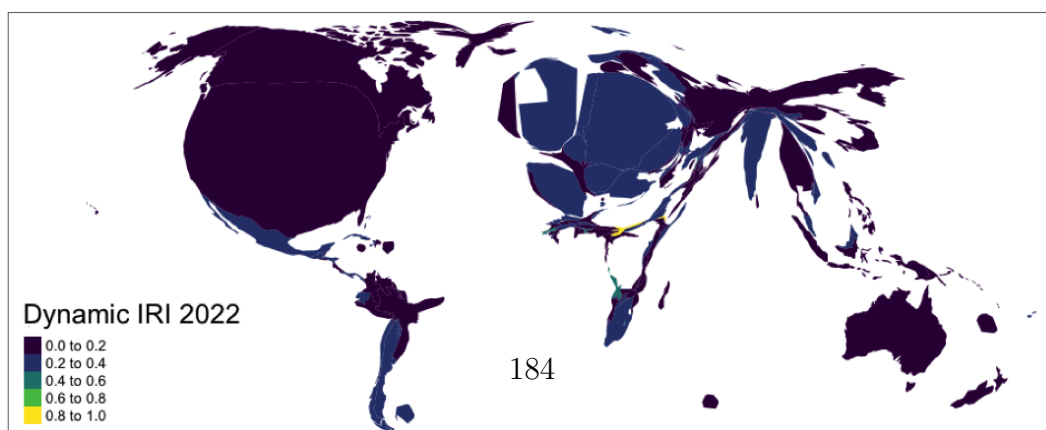
A



B



C



Appendix C

Supplementary Information Chapter 5

C.1 Levels of High-Risk users

We therefore choose to distinguish users between those only sharing reliable (Mainstream Media, Science) and Low Risk news (Satire, Other, Political, Clickbait), and those sharing high risk content (Fake/Hoax, Conspiracy/Junkscience), and then differentiating further based on the increasing engagement with this type of high risk content. For this reason, we decided to classify the users into four categories, highlighting their increasing interaction with high risk sources. For the purpose of comparison, we decided to perform the same analysis, taking into account other levels of interactions with high risk news media sources in order to understand if the same pattern is recognizable in such a scenario. Our initial result is confirmed with the ones found when considering several levels of high risk sources as shown in Figure C.1.

C.2 Verified accounts

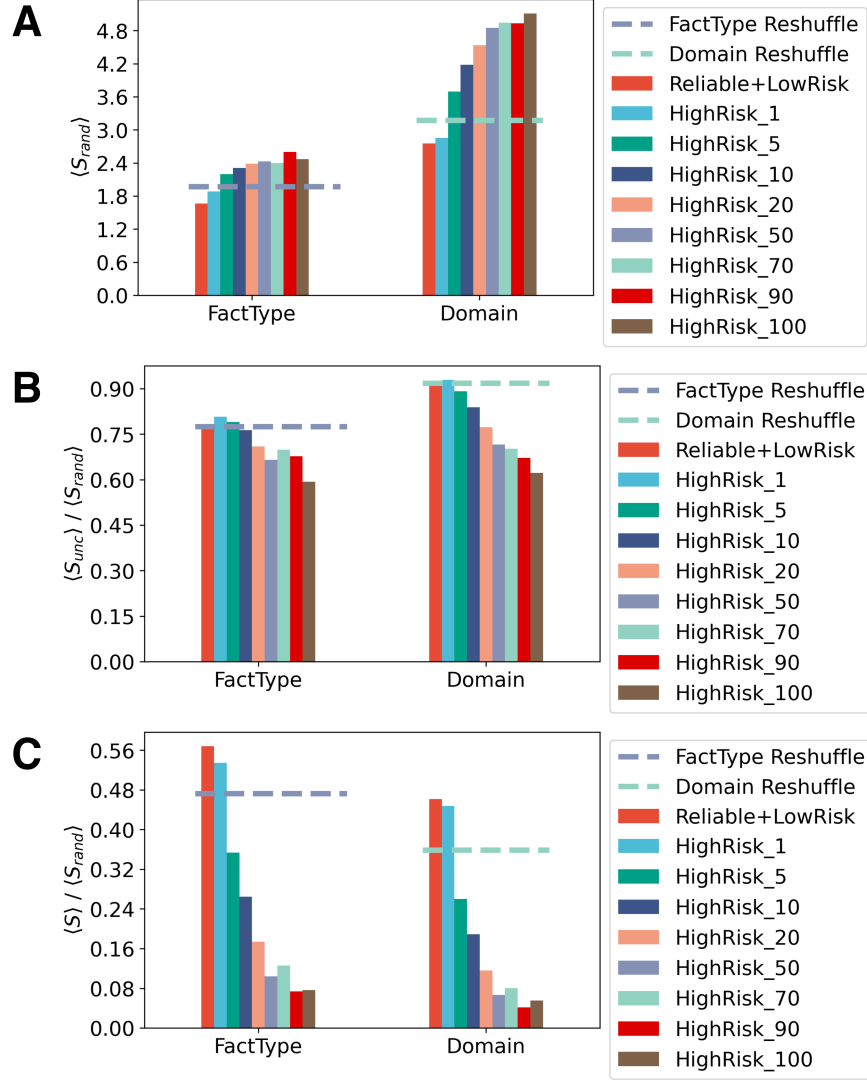
In order to differentiate between single users accounts and accounts managed by multiple users, we decided to perform the same analysis by comparing users classified as verified and those classified as unverified. In this context, Verified users are those having the blue verified badge, a blue check mark, that defines those accounts that are of public interest because they are considered authentic, notable and active on Twitter. The verification process of users is given by the blue check mark that can be found next to the username, while unverified accounts do not have this distinctive signal. However, this definition pertains to the period before Musk's takeover, during which accounts were required to undergo request verification. Starting from April 1st, 2023, there has been a change in

the rule. For users to acquire the verification badge now, they are required to subscribe to Twitter Blue. Regarding this research, we adhere to the initial definition since the new rule was not in effect during the considered time frame. However, we found that the verified accounts are almost 2% of our dataset, sharing mainly reliable information, few accounts engage with high-risk news media content. Also in this case, we observe that both verified users and unverified users show a similar pattern especially when comparing the random entropy and the actual entropy, demonstrating again that users sharing unreliable content have a more varied diet but more regular compared with users sharing only reliable and low-risk content.

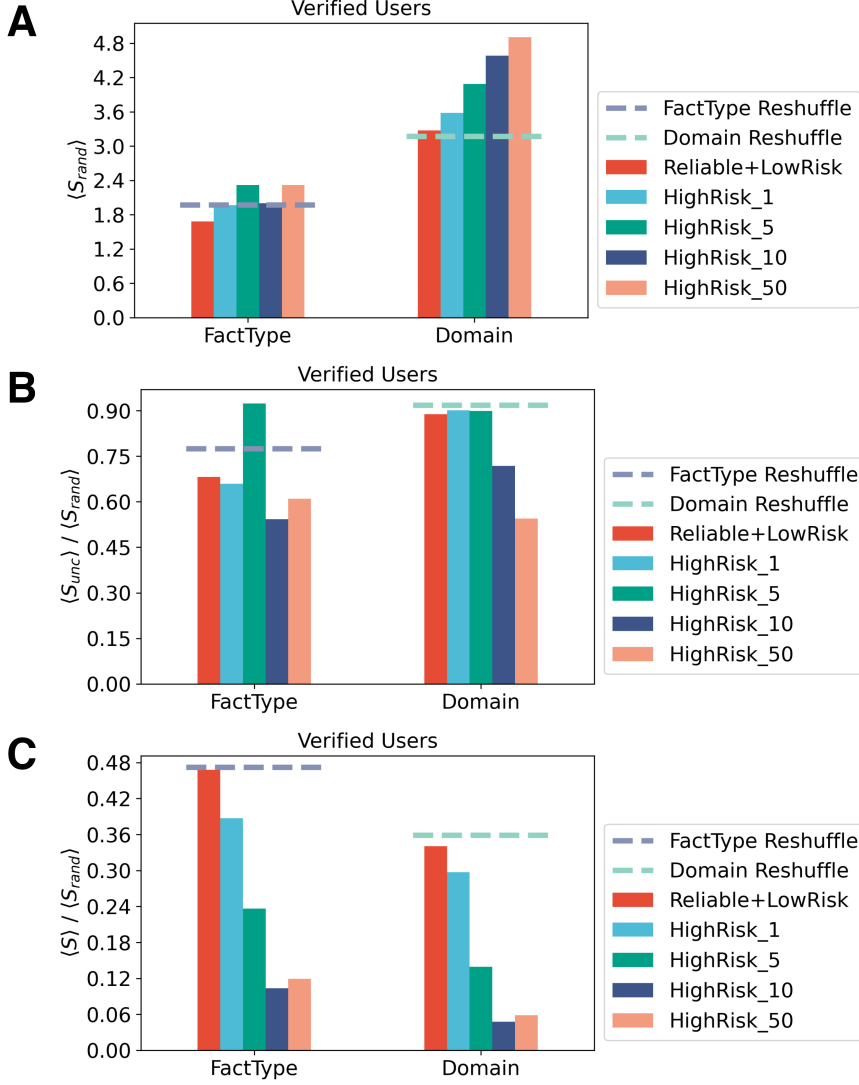
C.3 Cross-countries Analysis

In our analyses, we consider an aggregated sample of 25 countries, having at least 500 tweets on average per day, highlighting a sort of collective pattern across countries. For the purpose of comparison, we decided to take into consideration two different countries to explore the differences or similarities among them with respect to the results found in our analysis. Drawing inspiration from World Press Freedom¹ Index, an index measuring the freedom of the press in 180 countries world- wide, we compared Colombia with a lower value (45,23) against the Netherlands having a much bigger value (87). We found that both countries show a similar pattern, taking into consideration also the values observed when aggregating all the countries considered in the study. Indeed, the news media environment (inter-relations among different news media categories) is characterized by having a strong relationship with mainstream media journals and political news, when we look at the value not normalized. On the contrary, the relations between fake and conspiracy news is much stronger when comparing our results with the null model. Interestingly, we found that this relation is much stronger for Colombia rather than the Netherlands. Indeed, the Netherlands show a much stronger relation between Fake and Political News when analyzing our results normalized. We investigated also the news media diet by comparing the three different measures of entropy (Random, Uncorrelated and the actual entropy), demonstrating that the results are consistent with the ones found in the aggregated analysis and among the two different countries.

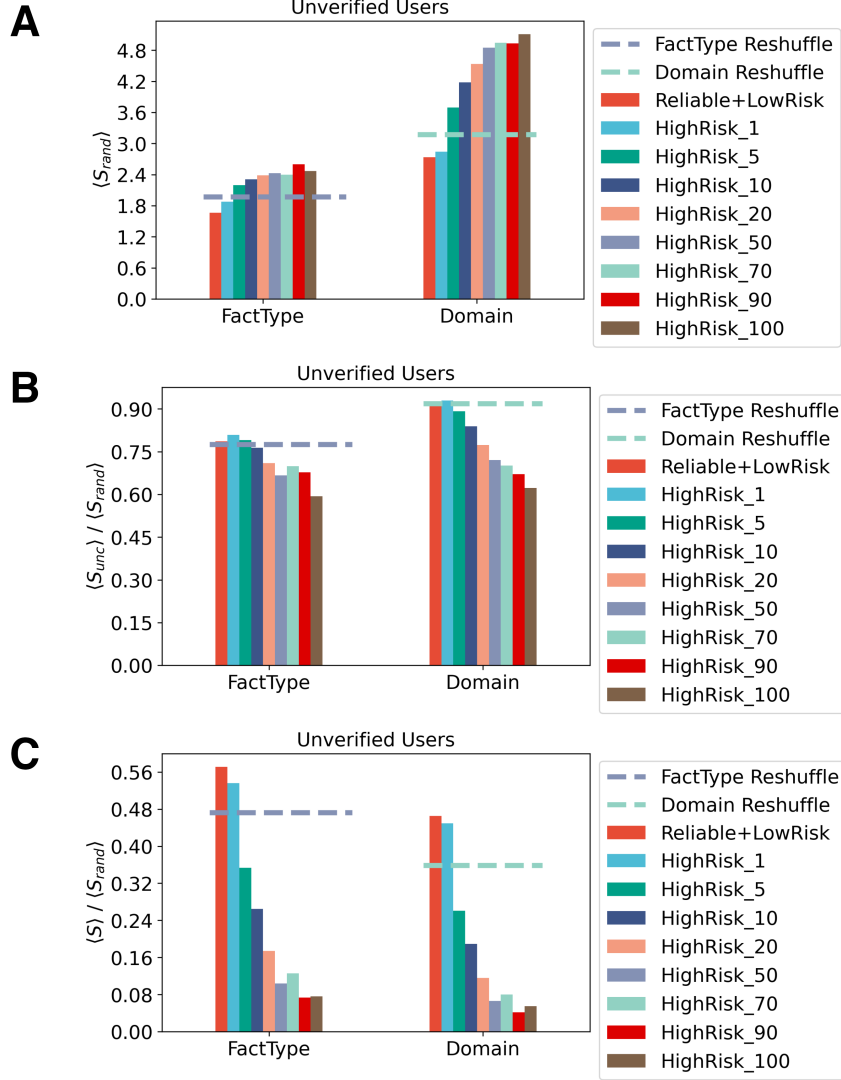
¹<https://rsf.org/en/index>



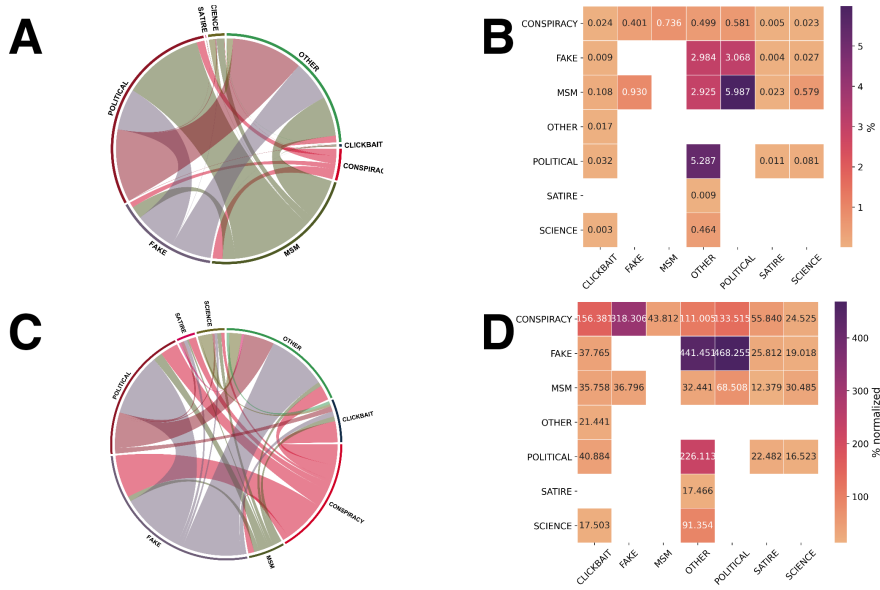
Supplementary Figure C.1: **News Media Diet for different type of users (9 categories of users).** (A) Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.



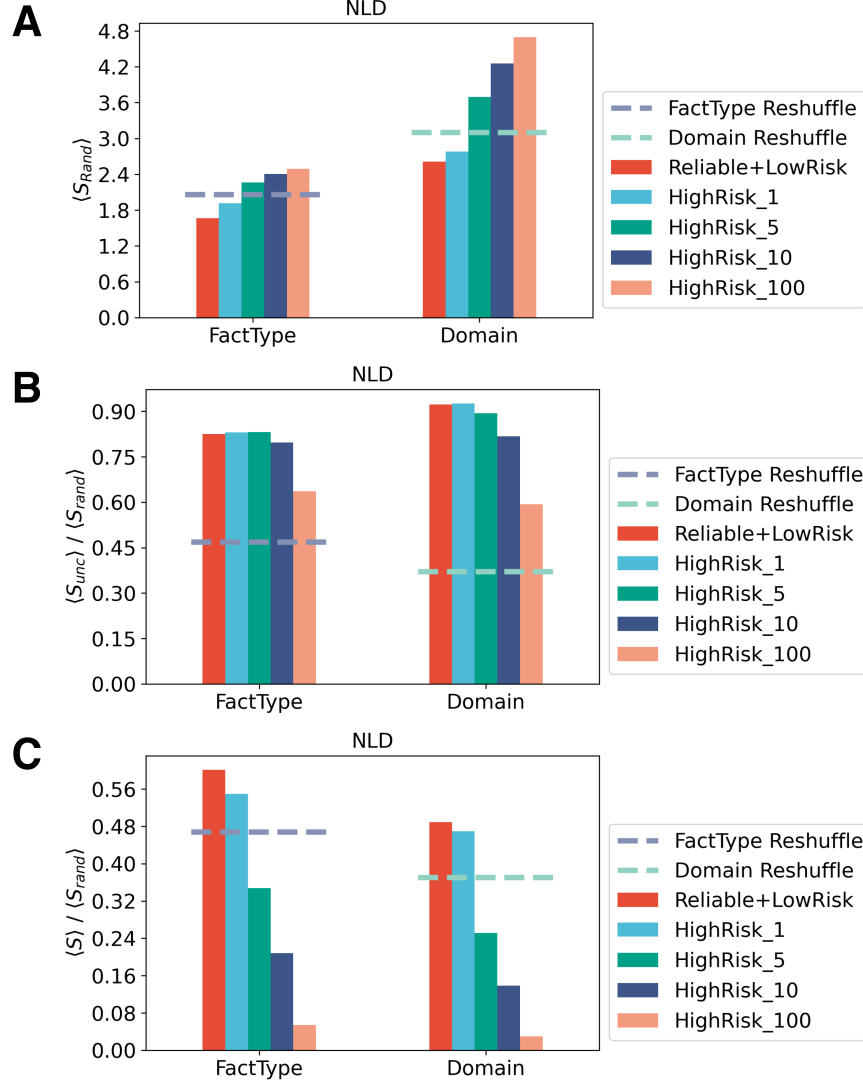
Supplementary Figure C.2: **News Media Diet for verified users.** (A) Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.



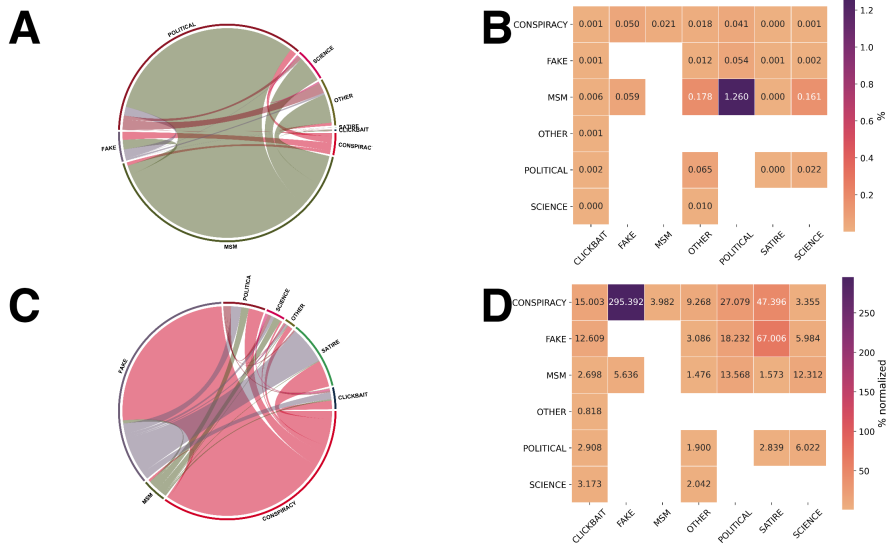
Supplementary Figure C.3: **News Media Diet for users classified as unverified.** (A) Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (B) Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. (C) The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.



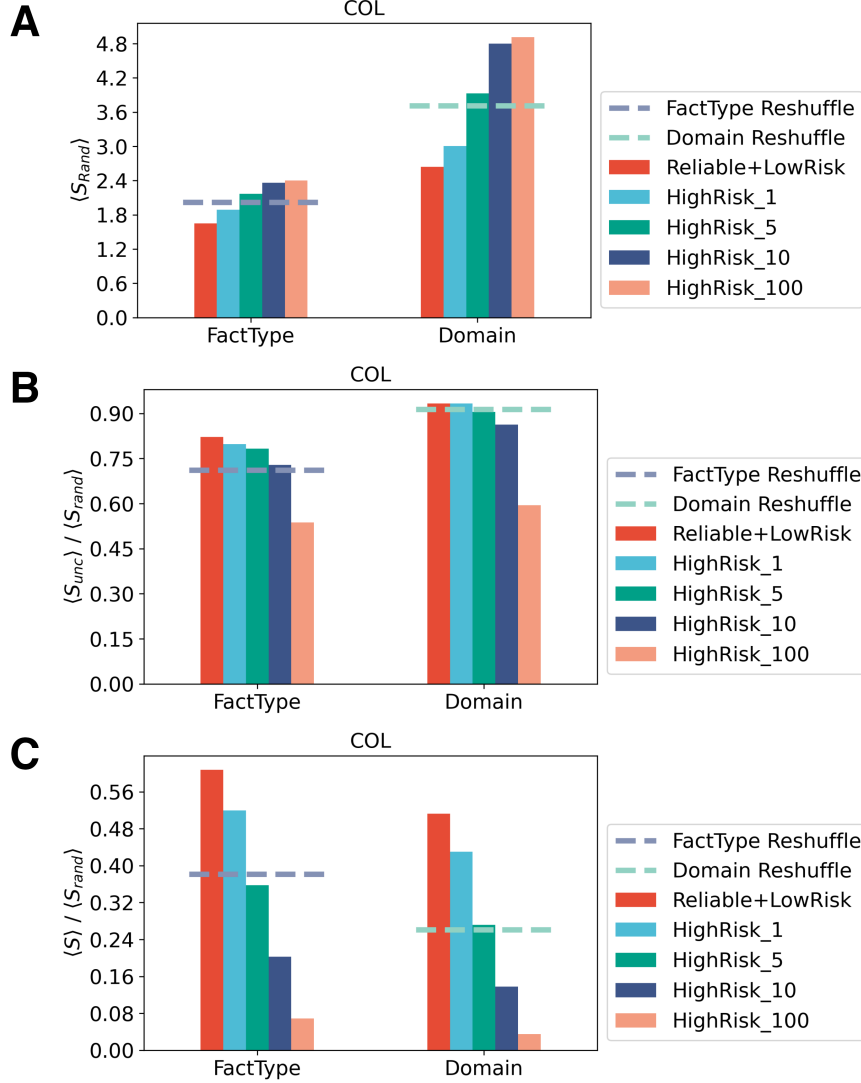
Supplementary Figure C.4: **The News Media Environment (The Netherlands).** (A) Weighted networks of the interactions among different news media types (also known as intra-relations) for 25 countries in 2020. (B) Heatmap showing the corresponding value of the inter-relations among the eight categories. (C) Weighted networks normalized comparing with a null model accounting for the proportion of the number of news belonging to each category. The chances of sharing conspiracy and fake news by the same users is much higher than the strong relation observed between mainstream media and political news (Panel A). (D) Heatmap showing the corresponding normalized value of the inter-relations among the eight categories.



Supplementary Figure C.5: **News Media Diet for different type of users (The Netherlands).** **(A)** Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. **(B)** Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. **(C)** The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.



Supplementary Figure C.6: **The News Media Environment (Colombia).** (A) Weighted networks of the interactions among different news media types (also known as intra-relations) for 25 countries in 2020. (B) Heatmap showing the corresponding value of the inter-relations among the eight categories. (C) Weighted networks normalized comparing with a null model accounting for the proportion of the number of news belonging to each category. The chances of sharing conspiracy and fake news by the same users is much higher than the strong relation observed between mainstream media and political news (Panel A). (D) Heatmap showing the corresponding normalized value of the inter-relations among the eight categories.



Supplementary Figure C.7: **News Media Diet for different type of users (Colombia)**. **(A)** Random Entropy S_{rand} calculated for different type of accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. **(B)** Shannon Entropy S_{unc} calculated on the domain and the type of news shared by different type of users accounts: users posting reliable or low-risk content and users posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content. **(C)** The actual entropy S calculated on the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high risk (conspiracy/junk science and/or fake/hoax) content.

Appendix D

Supplementary Information Chapter 6

Table D.1: List of the socioeconomic indicators considered for analysis. Table of the socioeconomic indices, the organization and sources, and a brief description.

Variable	Description
eiudemocracyindex	It measures the level of democracy in a given country, based on 60 different indicators over 5 main categories: electoral processes and pluralism, functioning of government, political participation, democratic political culture, and civil liberties. It is produced by the Economist Intelligent Unit (EIU) of the Economist group. The data come from experts' assessments and public opinion surveys, mainly from the World Values Survey. The index ranges from 0 (highly undemocratic) to 10 (highly democratic) [295].
eiupoliticalparticipation	The EIU Political Participation indicator [296] measures the degree of citizen engagement in political processes. It corresponds to one of the 5 main categories of indicators used to compute the Economist Democracy Index. It examines factors such as voter turnout, citizens' engagement with politics, party membership and participation in civil society organizations.

Variable	Description
fh freedom index	The Freedom Index [297], produced by the US NGO Freedom House, assesses the political freedoms and civil liberties of citizens in a given country. It ranges from 0 to 100 and is composed of 10 political rights indicators and 15 civil liberties indicators which assess several dimensions of freedom, including electoral processes, political pluralism, freedom of expression and associational rights.
iep global peace index	The Global Peace Index [298], compiled by the Institute for Economics and Peace (IEP), quantifies the relative peacefulness of countries and regions. It incorporates various indicators, including levels of violence, militarization, and political instability, to assess the overall peacefulness and stability of a nation. Originally, the index was expressed on a scale from 1 to 5, where 1 signifies most peaceful and 5 least peaceful. In order to align its interpretation to that of the other indices, it was reversed.
iep global terrorism index	The Global Terrorism Index [299], also produced by the Institute for Economics and Peace (IEP), provides a comprehensive measure of terrorism-related activities worldwide. It analyzes factors such as the number of terrorist incidents, casualties, and the economic impact of terrorism, enabling a comparative assessment of terrorism levels across countries.
imf gdp percap current prices	This variable corresponds to the Gross Domestic Product per capita at current prices, as reported by the International Monetary Fund [300]. It is calculated by dividing the GDP by the population size and serves as a measure of the economic prosperity and standard of living within a nation.
oecd perc adults tertiary edu	This variable represents the percentage of adults (aged 25-64) with tertiary education in a given OECD country [301].
oecd perc foreign students	This variable measures the percentage of foreign students enrolled in educational institutions in a given OECD country. It offers insights into a nation's attractiveness as an educational destination and reflects its openness to international students.

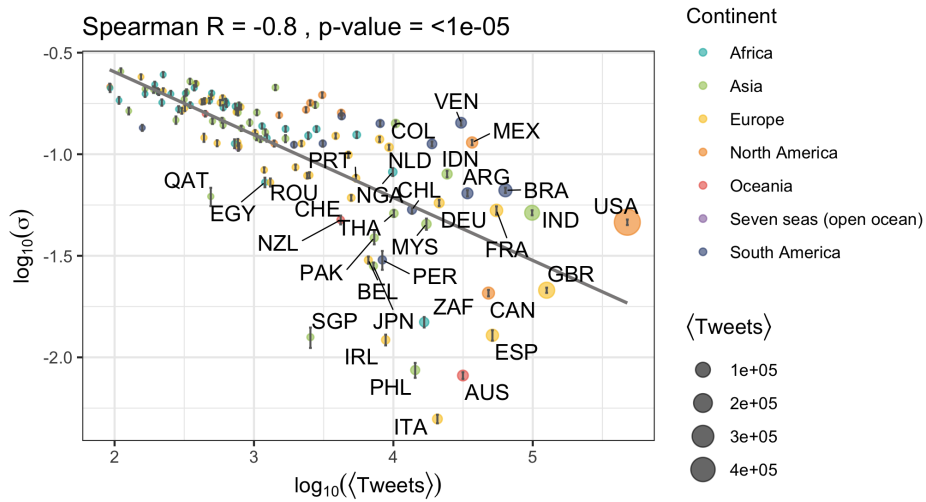
Variable	Description
oecd_pisa_mathematics	This variable represents the performance of 15-year-old students in mathematics, as assessed by the 2018 Programme for International Student Assessment (PISA). Provides a measure of educational results in each of the 78 countries surveyed [302].
oecd_pisa_science	This variable represents the scientific knowledge and skills of 15-year-old students in a given country, as assessed by the 2018 Programme for International Student Assessment (PISA) [302].
oecd_soc_security	This indicator measures the extent of social security contributions within a given OECD country as a percentage of its GDP. These are “compulsory payments paid to general government that confer entitlement to receive a (contingent) future social benefit” and provide insights into the level of social protection within a country, its economic development and the level of social inequality [303].
oecd_social_spending	This variable quantifies the share of a given OECD country’s GDP allocated to social expenditures, including welfare, healthcare, and family support programs. It serves as an indicator of the country’s level of social development and its commitment to redistributing resources for the well-being of its population [304].
oecd_trust_government	This variable assesses the share of citizens of a given country that report to trust their own country’s government. It is calculated using the number of people answering positively to the survey question “In this country, do you have confidence in the national government?” [305].
rsf_freedom_press	The Press Freedom Index evaluates the level of freedom enjoyed by journalists and the media in a given country. It is produced by the France-based NGO Reporters Without Borders. It assesses factors such as media independence, access to information, and protection of journalists, providing insights into the state of press freedom and freedom of expression [306].

Variable	Description
un egov e-participation index	The E-Participation Index measures the level of citizen participation in online government processes and decision-making. It is calculated in the United Nations E-Government Survey carried out by the UN Department of Economic and Social Affairs. It assesses the effectiveness of IT services that facilitate information exchange and interaction between government and citizens. Moreover, it gives a measure of citizen involvement in policy and decision-making. The index ranges from 0 to 1 [307].
un median age	This variable measures the median age of a country, namely the age that divides a population into two equal halves, with half of the population being older and half being younger than that. The estimates are provided by the United Nations [296].
vdem pol polarization	This indicator measures “the extent to which political differences affect social relationships beyond political discussions” [308]. When supporters of contrasting political factions are hesitant to participate in friendly interactions, which could happen in a variety of contexts such as in their workplace or in family functions, it indicates a high level of political polarization within societies. This variable is derived from the V-Dem (Varieties of Democracy) dataset, a large-scale political science dataset which aggregates country expert judgments in order to produce valid estimates of several concepts related to democracy. In particular, this indicator ranges from 0 to 4, where 0 signifies no polarization (i.e., supporters of opposing political factions tend to engage with each other in a cordial and friendly manner) and 4 represents the highest amount of polarization, with a generally hostile social environment among supporters of different political factions.

Variable	Description
vdem soc polarization	This variable measures the level of polarization in a society. Unlike the previous variable, which focused on the nature of interactions between supporters of opposing political camps, this indicator provides information on whether 'there is general agreement on the general direction this society should develop' [308]. More specifically, this variable comes from the aggregation of country experts answering the question "How would you characterize the differences of opinions on major political issues in this society?". The possible values originally span from 0 to 4, with 0 signifying serious polarization in society, resulting in major clashes of views, and 4 meaning that, despite the presence of differing opinions, there is a general consensus on the direction for key political issues. However, similarly to the Global Peace Index, the values of this index were flipped to achieve easier interpretation. Consequently, higher values of the metric in the final dataset correspond to higher levels of polarization.
wb perc internet use	This indicator measures the level of Internet use in a given country, according to World Bank estimates [309]. In particular, it represents the percentage of individuals who reported in a survey to have used the Internet in the last 3 months. It provides insights into the level of internet penetration, digital connectivity and technological development in a country.
wb gini coefficient	The Gini coefficient is a measure of income inequality. It ranges from 0 to 1, where higher values represent greater income inequality. Data come from the World Bank, except in the case of New Zealand, where the estimate is produced by the OECD [310].

Appendix E

Supplementary Information for Chapter 7



Supplementary Figure E.1: **Correlation between average number of tweets and volatility of Infodemic Risk Index** We calculated the average number of tweets per country and the volatility, defined as the standard deviation of IRI every two days.