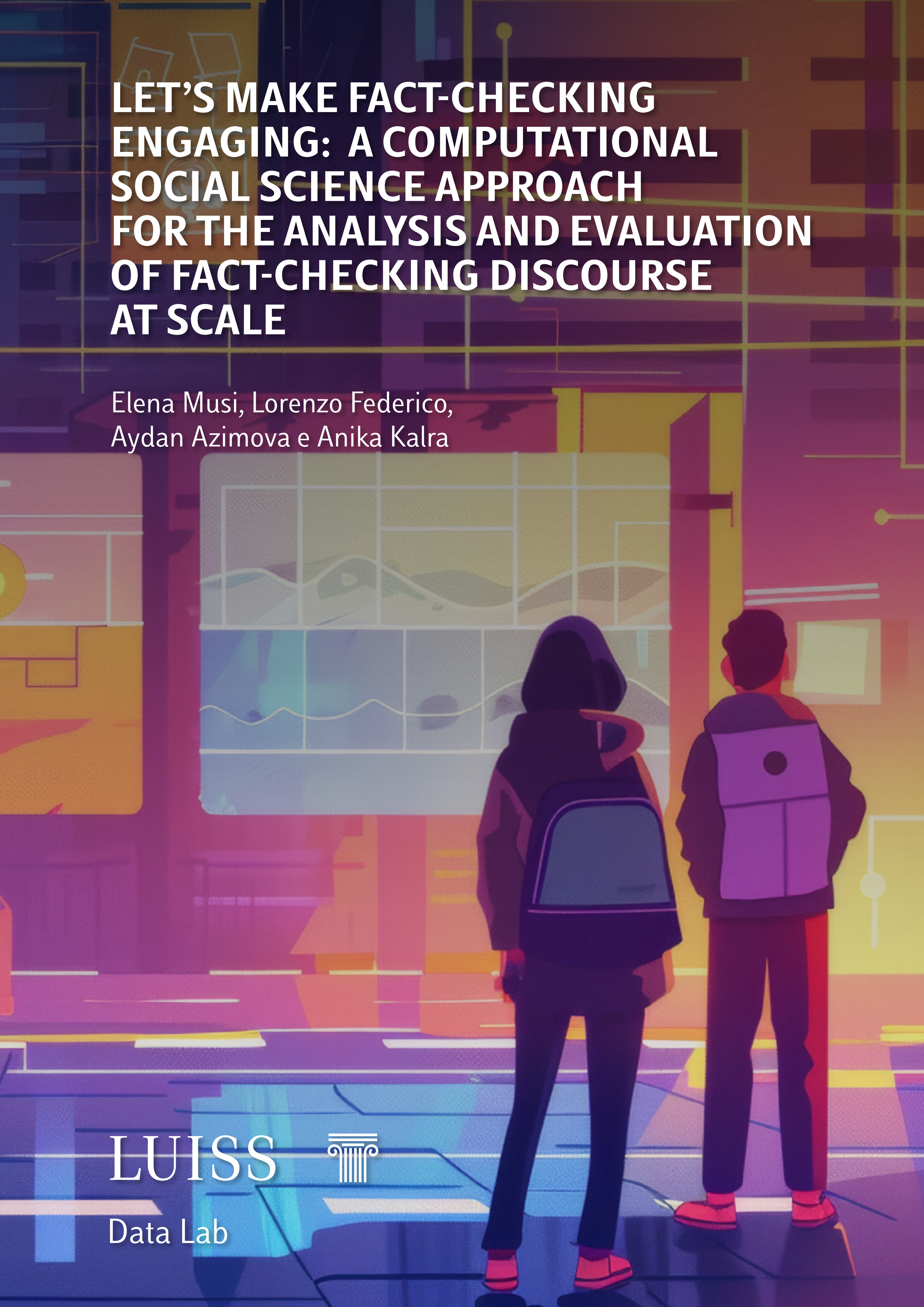
# LET'S MAKE FACT-CHECKING ENGAGING: A COMPUTATIONAL SOCIAL SCIENCE APPROACH FOR THE ANALYSIS AND EVALUATION OF FACT-CHECKING DISCOURSE AT SCALE

Elena Musi, Lorenzo Federico,
Aydan Azimova e Anika Kalra



LUISS
Data Lab

LUISS

Data Lab

To counter disinformation, a plethora of fact-checking organizations have arisen, but the fake news phenomenon is far from being solved. Whilst research in computational social science has widely focused on implementing automatic fact-checking, the analysis of the discourse features which make fact-checking impactful has been under-investigated. By combining social network analysis, emotion-based sentiment analysis and topic modeling, our study proposes a new methodology to compare and contrast content and audiences of fact-checks and disinformation messages, with a particular focus on engagement patterns. We apply our approach to a mixed corpus of Italian fact-checks and Italian disinformation accounts on Twitter. Drawing from the results of the analysis, we formulate a set of recommendations to make fact-checking discourse more engaging and impactful, leveraging rhetorical and content-based features.

**Keywords**
Fact-checking discourse,
Italian (mis)information ecosystem,
audience engagement,
computational social science

Italian Digital
Media Observatory

Co-financed by the Connecting Europe
Facility of the European Union

This report was written as part of the Italian Digital Media Observatory project under EC grant agreement INEA/CEF/ICT/A2020/2394428. The information, documentation and figures available in this deliverable were written by the Italian Digital Media Observatory project consortium and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

# 1. Introduction

The advent of the Network Society has radically changed the information ecosystem bringing at once to new participatory models of news production such as citizens' journalism, and to the fast proliferation of dis- and mis-information. The rise of the fake news phenomenon has triggered efforts by the journalism community to counter fakery through fact-checking, the verification of factual information to attain trustworthy news discourse. While the first official fact-checking organization was established in early 2000s to evaluate political claims appeared in the United States (Graves and Cherubini, 2016), the number of fact-checkers rapidly grew hand in hand with the term fake news becoming a buzz-word. According to **Duke Reporters' lab** the number of fact-checking organizations currently amounts to 353 at a global scale. Despite joint efforts such as the International Fact-Checking Network, the infodemic is far from being solved. As underlined by Temmerman et al. (2019) in relation to the political domain, the rise of perspectivism in the information landscape has caused a drop in trust in journalists as truth speakers, which might also cast doubt on the practice of fact-checking. In the post truth era, (fake) news making is a process ultimately aimed at gaining the acceptance of a certain interpretation of an event. It is thus a form of argumentation, intended as "a discourse aimed at convincing a reasonable critic of the acceptability of a standpoint by giving reasons that justify the standpoint" (Grootendorst and van Eemeren, 2004: 1). Thus, the way discourse is constructed in terms of  number and type of arguments affects its persuasiveness and outreach.

Extant scholarly research has highlighted various factors which hinder the effectiveness of fact-checking. One set of issues lies in the phenomenon of fake news itself: through digital media fake news spreads at an incremental pace and encompasses both blatantly false (disinformation) and (un)intentionally misleading information (misinformation). As a result, human fact checking struggles to keep up with the abundance of information, while automatic fact checking (Zhou and Zafarani, 2020), relying on binary classification systems, cannot keep up with the gray area of misinformation. A second facet has to do with the audience's attitudes towards fact-checking. Results from studies evaluating fact-checking show that the effectiveness in correcting misperceptions is not agreed upon: for instance, while Nyhan and Reifler (2015) reveal the presence of a backfire effect where the correction of false information has the corollary of radicalizing individual views, Wood and Porter (2019) report on the elusive nature of the backfire effect. Acknowledging the difference between behaviors in lab and in the wild, Jiang and Wilson (2018) have analyzed linguistic signals marking users' comments in presence of misinformation and fact-checks, revealing significant variations in terms of misinformation-awareness signals, emojis and swear words.

Rather than addressing cognitive aspects which crucially depend on the population sample and the domain (e.g. pandemic vs politics) or users' language,  this study aims at investigating at scale differences in the construction of discourse between fact-checks and fake news and the resulting audience engagement. It does so by taking as a case study the Italian (dis)information ecosystem. The underlying assumption is that disinformation and fact-checking share as felicity condition *visibility*, which, at least partially, depends on the construction of news discourse: a fake news is effective if it manages to persuade a wide audience about its truth, the same way as fact-checks reach their communication goal if they manage to correct false beliefs across publics.

With the final goal of making fact-checking discourse engaging and enhance its outreach, we tackle three preliminary research questions:

- How can we monitor at scale whether fact-checkers' audiences and discourses are (dis)similar to those of disinformation accounts?
- How can we map engagement levels driven by fact-checks and fake news in terms of audiences' size and activities?
- What presentational and topical features make fake and fact-checked news discourse engaging?

To address such questions, we adopt a computational social science approach combining network analysis, emotion-based sentiment analysis and topic modeling. We take as a case study the activities of Twitter accounts of the 7 Italian Fact-Checkers affiliated to the International Fact-Checking Network and of 10 disinformation outlets with similar number of followers from 1st November 2020 to 1st November 2021. The study is organized as follows: after having introduced state of the art computational social science approaches to evaluate the impact of fact-checking (section 2), we describe the process of data collection (section 3.1) and the methodological approaches leveraged for the analysis (section 3.2). We then report on the results of the social network analysis (section 4), comparing fact-checkers and disinformation accounts in terms of total number of interactions (section 4.1), distribution of number of retweets per account respectively making up the two audiences (section 4.2), and dissasortativity (section 4.3). Section 5 is devoted to the results of the sentiment analysis (polarity and emotion) (section 5.1) and the topic modeling (section 5.2) with respect to engagement metrics, complemented with the qualitative analysis of identified trends. In section 6 we propose a set of data-informed recommendations on how to boost fact-checks' engagement through discourse features.

# 2. Computational social science to evaluate fact-checking

So far, research has cast doubt mainly on the i) epistemology of the fact-checking process and ii) the assumption that information directly affects attitudes. A to i), Uscinski and Butler (2013) for instance, identify five methodological criticisms which make fact-checking naive (e.g. inexplicit selection criteria of facts), while Marietta et al. (2015) point to the confusion caused by different fact-checkers providing different ratings over the same news. Focusing on political ads, Amazeen (2016) comments on the lack of impartiality in what claims are selected and evaluated. Adopting a more holistic approach, Nieminem and Sankari (2018) come up with 25 criteria to evaluate the fact-checking process claiming that complex propositions containing multiple facts shall not be treated as a single one, and that claims whose truthfulness cannot be defined in practice shall not be targeted by fact-checks. As to ii) various studies have revealed the presence of a backfire effect when fact-checkers repeat false information to correct it (Cohen et al., 2007): repetitions even though in the scope of a warning have often as a side-effect increased misbelief.

To our knowledge Brandtzaeg et al. (2018), in the frame of the three-year European Union project REVEAL, address for the first time the use and impact of fact-checking services from an audience perspective, combining interviews with journalists with content analysis of social media users' online conversation about fact-checks. The results show that usefulness is the main reason underlying positive attitudes, while trustworthiness issues such as partisanship back negative attitudes. The considered audience in these studies is by default engaged by fact-checkers, either professionally (journalists), or as digital media users which are sensitive to the matter to the point of discussing it. As pointed out by Raves, Nyhan and Reifler (2016)'s field study among reporters, messages promoting the high status and journalistic values of fact-checking are more effective in increasing fact-checking coverage compared to messages highlighting the audience's demand for fact-checking. This shows that the practice of fact-checking responds first of all to professional motives within journalism. These studies leave un-investigated the actual outreach of fact-checkers in the wild and its correlation with discursive elements in comparison with disinformation accounts.

Studies stemming from computational social science mostly employ natural language processing techniques to understand the spread of fake news and create systems for their automatic detection (Alonso et al., 2021). Context-based approaches (Antonakaki et al., 2021) leverage network analysis: for instance, Shu, Bernard and Liu (2019) show how different types of networks can be used to represent and model fake news propagation.

Adopting a comparative perspective, Burel et al. (2020) analyze how misinformation and fact-checking information about COVID-19 dilagate over Twitter, combining spread variance analysis, impulse response modeling and causal analysis. The results show similarities in the way the information diffuses over time, even if fact-checks happen to be far less shared and in a shorter term, hindering the positive impact against misinformation. The authors advocate for a content analysis of language features which might make fact-checks more appealing, which we tackle in the present study. At the content-level, sentiment analysis (Anoop et al., 2021) is used as a feature to build classifiers for the automatic identification of disinformation both at the polarity and emotional levels.

However neither social network analysis nor sentiment analysis have so far been used to examine discursive features of fact-checks and match them with their outreach. Rather than using computational techniques to replace the role played by human fact-checkers, we leverage them to shed light on key issues and empower the fact-checking process.

# 3. Data and Methods

## 3.1. Corpus Collection

To collect our corpus we focus on Twitter for two main reasons. First, fact-checking websites have a Twitter official account which allows for multiple ways of interaction (retweets, replies, likes) that can be used to develop users' networks. Second, Academic Research Access to the Twitter API enables the collection of large amounts of public tweets (10 million per month) coupled with interaction content. As case study, we consider the accounts of the 7 Italian fact-checkers part of the **International Fact-checking Network**: Pagella Politica, BUTAC, Bufale.net, Facta, La Voce, OpenFactCheck and Blasting News. We then discard the last two: the Open Fact Check twitter account has been created towards the end of the considered time span, while Blasting News Italia has a very low amount of interactions and it publishes mostly content unrelated to fact-checking. As disinformation outlets, we pick the 10 Italian accounts whose name are anonymized for privacy, we will refer to them as Disinfo 1,...,10.. For the selection of the accounts we take as a benchmark those identified in the joint report by Università Luiss Guido Carli, Harvard Kennedy School and School of Information dell'Università del Michigan about the **Italian disinformation ecosystem** published in 2021. To allow for comparison with the fact-checkers accounts, we retain those that have a similar distribution in terms of number of followers, ranging from 1000 to 100000; finally, we retrieve further accounts from those suggested under the section "pagine correlate" of these three accounts, filtering them as to number of followers. We select 10 instead of 7 accounts to guarantee balance in political orientations. The time span used for the collection is the month 20/11/2020 -- 20/11/2021. Overall, the composition of our corpus is represented in Figure 1:

Table 1: Number of tweets and maximum number of followers of each account in our corpus

| **Fact-checkers** | **Number of tweets** | **Maximum Number of followers** |
|---|---|---|
| Pagella Politica | 2156 | 20904 |
| Butac | 1929 | 30565 |
| Bufale.net | 6416 | 11526 |
| La Voce | 1061 | 75740 |
| Facta | 750 | 3544 |
| **TOT** | **12312** | **142279** |
| **Disinformation accounts** | **Number of tweets** | **Maximum Number of followers** |
| Disinfo 1 | 7401 | 29807 |
| Disinfo 2 | 13922 | 25154 |
| Disinfo 3 | 13301 | 18874 |
| Disinfo 4 | 2670 | 99599 |
| Disinfo 5 | 4379 | 1195 |
| Disinfo 6 | 407 | 8347 |
| Disinfo 7 | 323 | 6348 |
| Disinfo 8 | 2005 | 5232 |
| Disinfo 9 | 2469 | 4000 |
| Disinfo 10 | 736 | 1377 |
| **TOT** | **40212** | **199933** |

## 3.2. Methods

Our analytic pipeline follows a three-tiered approach, combining social network analysis with topic modeling (Negara et al., 2019) and sentiment analysis (Liu, 2012), and observing their interrelation to disclose potential patterns of overlap. Network analysis has proved to be an advantageous methodology to surface patterns of information flow, as well as attention giving and receiving across social media. In this context, the approach generally followed is that of building sociocentric networks, which include nodes and links formed about a particular set of topics or accounts that share topical interests (Himelboim, 2017). The input of our propagation networks are hop-based news cascades (Castillo et al., 2011), tree-like structures that capture the propagation of  a news article across a social network in a step (hop) by step manner.

More specifically, the networks are built around a topic which coincides with the content published by an outlet. For each such news outlet S, we build an associated network that represents the interactions between users and the content produced by the outlet. This network is bipartite: it is made of two types of nodes, and connections are drawn only between nodes of different types. More formally, for each S, we create the network G(S) = (T (S), U(S), R(S)) where:

- T(S) is the first part of the network, representing the tweets of the official account of the outlet S.
- U(S) is the second part of the network, representing the users that retweeted at least one of the tweets.
- R(S) is the edge set, where each edge connects a user with a tweet they retweeted.

This representation allows us to obtain information on the frequency and the patterns of repeated interactions between each of the 7 fact-checkers and 10 disinformation outlets, and users who constitute the audience of their content.

Topic modeling is a natural language processing statistical approach based on the analysis of lexical clusters. We use as an algorithm the Latent Dirichlet Allocation (LDA). LDA has proved to be a successful tool for a quantitative analysis of newspaper articles (Jacobi et al., 2016). As to sentiment analysis, we utilize the open source Python Library with a model trained on FEEL-IT (Bianchi, Nozza and Hovy, 2021). We adopt this library since FEEL-IT is the main benchmark, non-domain specific, corpus of tweets available for sentiment analysis in Italian. Furthermore, it is annotated as to four emotions, anger, fear, sadness and joy, allowing at once emotion detection as well as binary polar sentiment analysis (positive vs. negative), collapsing positive and negative emotions. We, in fact, test whether specific emotions spread faster than others showing viral behaviors.

# 4. Results

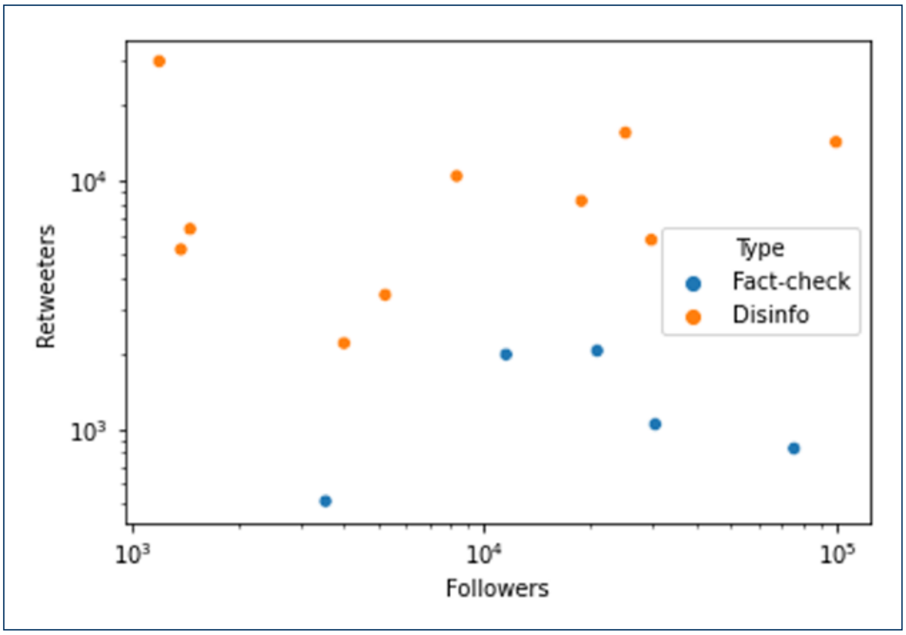
## 4.1. Social network analysis

### 4.1.1. Comparison of social networks: active or passive audience?

As the first level of analysis, we analyze the size and type of the networks represented by the number of followers, the overall number of retweets and the number of accounts retweeting content from the fact-checkers and influencers accounts. It turns out that the size of the audience (number of followers) which accesses the content is similar, but the overall number of retweets is higher for the disinformation accounts across the board (see Fig. 2), suggesting a more proactive behavior.

To check whether this trend is due to a restricted set of "hyper-active" users, we look at the number of active users, users that retweeted at least one tweet from the following accounts. The number of individual retweeters is consistently higher for each of the dis influencers accounts:

Figure 1 Scatterplot of the number of followers and of retweeters for all the accounts



Each point represents one of the 15 news outlets with different colors for fact-checks and disinformation accounts, and its coordinates are the number of followers and retweeters, both on logarithmic scale.

Table 2: Total number of retweets, individual retweeters and followers for all the accounts

| Fact-checkers | Number of retweets | Number of active users | Maximum Number of followers |
|---|---|---|---|
| Pagella Politica | 6044 | 2056 | 20904 |
| Butac | 5144 | 1041 | 30565 |
| Bufale.net | 7320 | 1985 | 11526 |
| La Voce | 1823 | 832 | 75740 |
| Facta | 1505 | 511 | 3544 |

| Dis Influencers | Number of retweets | Number of active users | Maximum Number of followers |
|---|---|---|---|
| Disinfo 1 | 33622 | 5738 | 29807 |
| Disinfo 2 | 319282 | 15469 | 25154 |
| Disinfo 3 | 60847 | 8234 | 18874 |
| Disinfo 4 | 188315 | 14184 | 99599 |
| Disinfo 5 | 3788 | 1195 | 1195 |
| Disinfo 6 | 62505 | 8347 | 8347 |
| Disinfo 7 | 4460 | 1462 | 6348 |
| Disinfo 8 | 9732 | 3450 | 5232 |
| Disinfo 9 | 9526 | 2205 | 4000 |
| Disinfo 10 | 11048 | 5255 | 1377 |

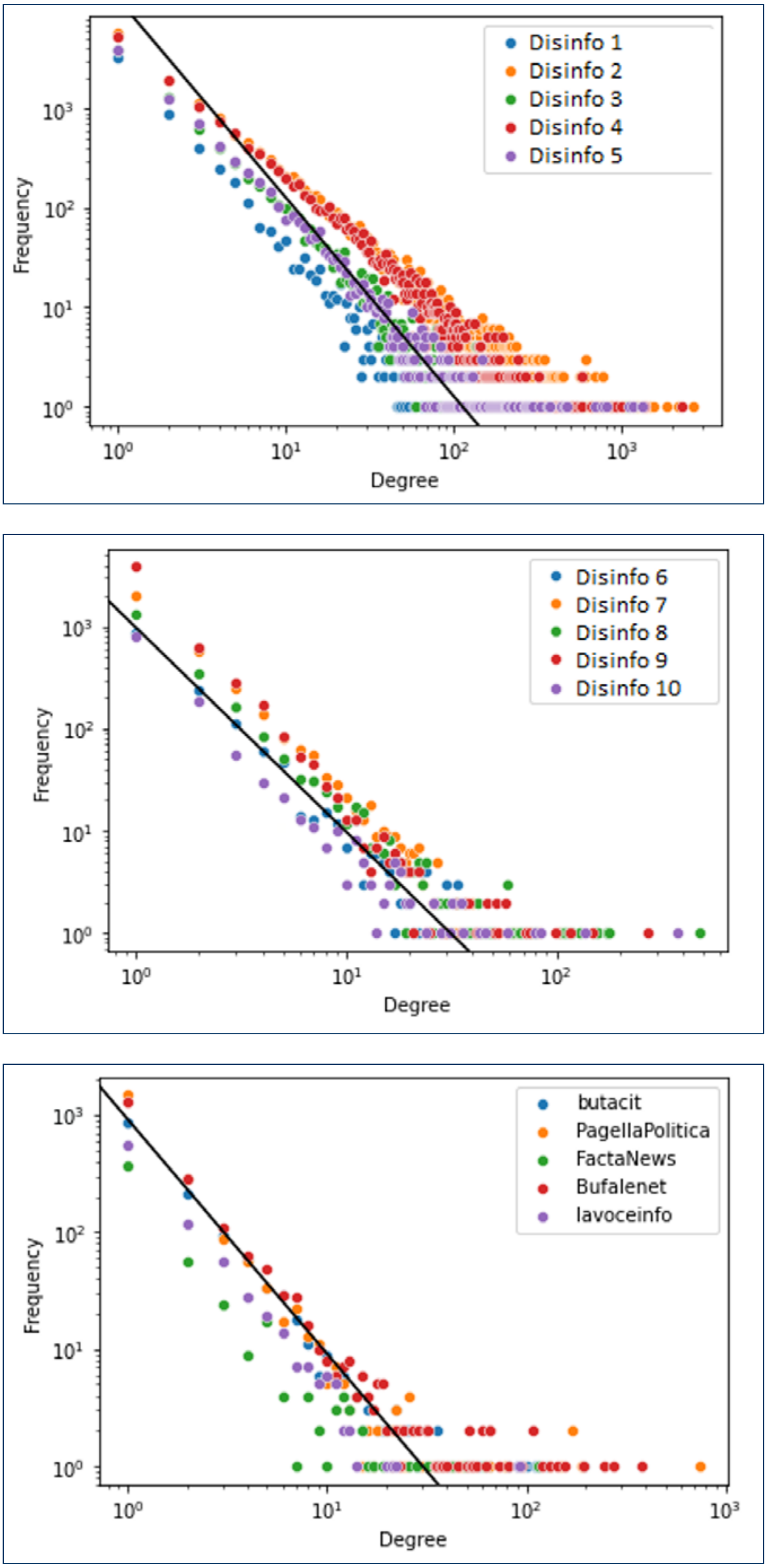
### 4.1.2. Distribution of the number of retweets per user

We study the distribution of the number of retweets for each active retweeter, showing that it is extremely inhomogeneous. While the vast majority of users retweet just a few tweets, for each fact-checker and disinformation account there is a small number of very active followers that retweet an unusually large amount of tweets, often in the hundreds or even thousands.

We observe that the distribution of the number of tweets retweeted by each active user, that is, the degree of a uniformly chosen vertex u in U(S), follows quite perfectly a power law for both the disinformation accounts and fact-checkers. A discrete probability distribution is called a power law if the probability to draw d is proportional to  for some fixed number a>1. In our study we notice that in most cases, the exponent of the power law is close to 2. This means that for each outlet S, whether it is a fact-checker or a source of disinformation, the number of accounts which have retweeted d tweets from the official account of S, is proportional to . A power law with exponent a<3 is called a scale-free distribution. Scale-free distributions have been extensively observed in multiple cases in network structures describing social interactions, among many other fields. Scale-free degree distributions are characterized by the presence of a majority of vertices of low degree (1 or 2) and a small minority of vertices of very high degree, sometimes of the same order of magnitude as the total size of the graph. Lu et al. (2014) and Kwak et al. (2010) show that on Twitter both the total number of retweets a tweet receives and the number of followers and followed for each user follow a power law. This is usually explained as a result of what is known as preferential attachment (Albert and Barabasi, 2002). Under the hypothesis of preferential attachment, a user or a tweet receives new followers or retweets respectively at a rate that is proportional to the number of followers or retweets they already have.

To show that the distribution of the number of retweet per user follows indeed a power law, in Figures 2-4 we plotted the logarithm of d on the horizontal axis against the logarithm of the number of vertices that have degree d in each network on the vertical axis. The fact that the plots resemble straight lines shows that the distributions follow a power law and the slope of the line corresponds to the exponent -a. For reference, the black lines indicate the theoretical degree distribution that corresponds to a power law with exponent 2. It, thus, emerges that the distribution of the number of retweets a user gives to any of the outlets over a full year timespan also follows a power law behavior. It is possible that the same preferential attachment phenomenon is in action here. Further investigation on the time-evolution across the months of the number of retweets between users would be necessary to confirm this hypothesis.

Figures 2: Plots of the degree distributions in the 15 analyzed networks



### 4.1.3. Disassortativity of user-content interaction

As a second step to understand what are the reasons underlying users' activity, we calculate the coefficient of assortativity (van der Hoorn and Litvak, 2015) of the networks, finding all of them to be disassortative. A disassortative network is one in which nodes are more likely to be connected with other nodes which have characteristics opposed to theirs. In this case, it encapsulates the tendency for users (nodes) with low degree (i.e. which interact sporadically with that specifical news outlet) to connect more often with tweets of high degree (i.e. that receive a high number of retweets) and vice versa. We recall that the degree of a tweet in our network G(S) is the number of users that have retweeted it, while the degree of a user is the number of different tweets from the account S they have retweeted. There is no standard measure of disassortativity for bipartite networks in the literature; we have thus adapted those developed by van der Hoorn and Litvak 2015 for directed networks, networks where each edge has a starting and ending node instead of two symmetrical adjacent nodes. For the purpose of computing its assortativity coefficient, we treated the bipartite user-tweet network as a directed network with every edge pointing from the user to the tweet they retweeted. We computed the Spearman rho and the Kendall tau correlation coefficients between the degrees of the starting and end vertex of each edge. We observe that all the networks have negative assortativity coefficients, as shown in Figure 2.

In a randomly generated graph with no significant assortative or disassortative properties, it is expected that they are both close to 0 (van der Hoorn and Litvak, 2015). We can conclude that all the 15 networks show a very evident disassortative structure.

Figure 3: Assortativity coefficients for all the 15 networks

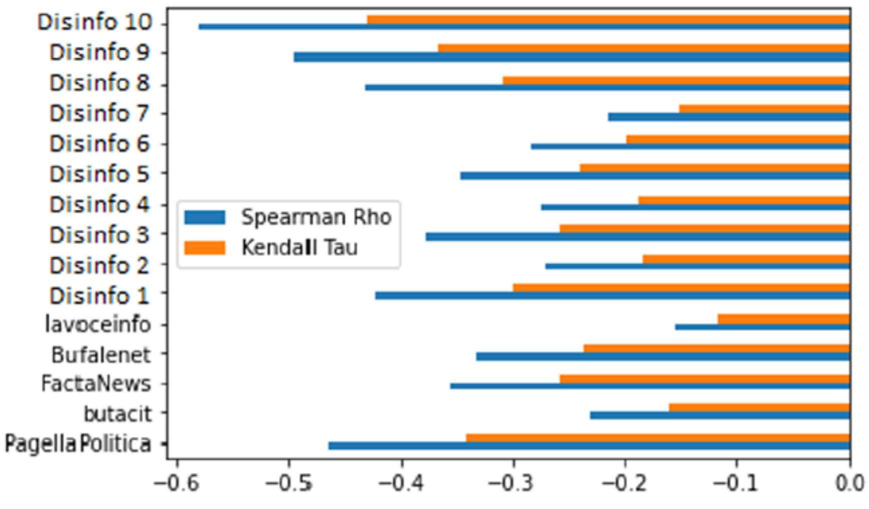


Table 3: Assortativity coefficients and the p-values of their test against the null hypothesis (the network has 0 assortativity).

| | Spearman Rho | rho_p | Kendall Tau | tau_p |
|---|---|---|---|---|
| Pagella Politica | -0.46421458224216633 | 0.0 | -0.3419139140453798 | 6.76844478955568*10$^{-303)}$ |
| butacit | -0.23130510597921522 | 1.7236427167844637$^{*10-63}$ | -0.16130864338343115 | 1.518325735739206$^{*10-61}$ |

| | Spearman Rho | rho_p | Kendall Tau | tau_p |
|---|---|---|---|---|
| Facta News | -0.35611505271332083 | 2.9546532490926023*$^{10-46}$ | -0.2593296645426564 | 5.26788371148615*$^{10-44}$ |
| Bufalenet | -0.3315165007650904 | 1.8359450965145596*$^{10-193}$ | -0.23743242238512902 | 2.3532252911714362*$^{10-188}$ |
| lavoceinfo | -0.15624166493626468 | 1.8336505712225093*$^{10-11}$ | -0.11692432790731076 | 1.998540433461791*$^{10-11}$ |
| Disinfo 1 | -0.4221289549183348 | 0.0 | -0.29935135366837623 | 0.0 |
| Disinfo 2 | -0.27103395187003193 | 0.0 | -0.18466332165025043 | 0.0 |
| Disinfo 3 | -0.37664826622271014 | 0.0 | -0.25998220838386166 | 0.0 |
| Disinfo 4 | -0.2751367520096186 | 0.0 | -0.18801872890760907 | 0.0 |
| Disinfo 5 | -0.3469117159825592 | 0.0 | -0.24008055346890522 | 0.0 |
| Disinfo 6 | -0.28215560531324985 | 2.0758997534726*$^{10-82}$ | -0.1996443036203274 | 1.9451323141451076*$^{10-81}$ |
| Disinfo 7 | -0.2149742568548442 | 4.013859829899452*$^{10-102}$ | -0.15118065118377771 | 1.820315072771605*$^{10-101}$ |
| Disinfo 8 | -0.4323185305735767 | 0.0 | -0.30768026663752895 | 0.0 |
| Disinfo 9 | -0.4947006382347482 | 0.0 | -0.3657658363705887 | 0.0 |
| Disinfo 10 | -0.5802926978904156 | 0.0 | -0.4308882305964561 | 7.762630426473147*$^{10-304}$ |

We interpret this as the effect of a core-periphery structure (Borgatti and Everett, 2000), in which the core users, that is the ones that retweet the content of the fact-checker or the disinformation account very often, tend to share equally both the most retweeted tweets and the least retweeted ones; the periphery users, instead, the ones that engage sporadically with the content, share the most retweeted tweets only.

# 5. Sentiment analysis and topic modeling

## 5.1. Emotions and engagement

Through the sentiment analysis we verified whether engagement levels are correlated with the expression of positive vs. negative polarity and emotion types within. As a first step, we classified all the tweets by polarity and by four types of emotion (fear, anger, sadness and joy). The engagement scores amount to the sum of likes and retweets, as per definition of engagement rate according to the official social media analytics. To include active engagement in our analysis, we have collected the replies to the fact-checkers' and disinformation tweets taking the same period used to gather the original tweets as time span. We ran the sentiment analysis algorithm on a total of 7743 replies to the fact-checkers and 144639 replies to the disinformation messages, after filtering out replies in other languages or that did not contain actual words (e.g., URLs as the sole content).

To check on significant correlations, we have resorted to the Kruskal-Wallis test ("one-way ANOVA on ranks"), taking as dependent variable the engagement rate (likes or retweets) and as independent variable polarity and emotion (divided into the four independent groups joy, fear, anger and sadness). We have chosen Kruskal-Wallis since the engagement rate of a tweet does not have a normal distribution, calling for a non parametric test. The results show that engagement differs based on emotion type and overall valence (positive or negative), for each of the fact-checkers and the disinformation accounts. Both emotion and polarity p values are far lower than the standard significance level of 0.05 for both types of outlets.

Table 4: Kruskal-Wallis test results for distribution of engagement across polarity and emotion

| | Fact-Checkers | Disinformation accounts |
|---|---|---|
| Polarity vs likes stat | 65.5850125375346 | 82.58280270828494 |
| Polarity vs. Likes p | 5.566043961536826 $^{*10-16}$ | 1.0133489691540844$^{*10-19}$ |
| Polarity vs. Retweets stat | 72.19572230896163 | 283.9782668002984 |
| Polarity vs. Retweets p | 1.9487773839990402 $^{*10-17}$ | 1.0201975957189145$^{*10-63}$ |
| Polarity vs replies stat | 19.83290936136614 | 572.3666922458817 |
| Polarity vs replies p | 8.451498941519831 $^{*10-6}$ | 1.715929211177166$^{*10-126}$ |
| Emotion vs likes stat | 75.89142549322602 | 412.0472290580368 |
| Emotion vs likes p | 2.333653083418208$^{*10-16}$ | 5.439322516647739$^{*10-89}$ |
| Emotion vs retweets stat | 107.9765331762197 | 447.2776546558097 |
| Emotion vs retweets p | 2.990653953478151$^{*10-23}$ | 1.2679080292693671$^{*10-96}$ |
| Emotion vs replies stat | 138.12824917278863 | 752.8357742331785 |
| Emotion vs replies p | 9.572465218183718$^{*10-30}$ | 7.322420315929276$^{*10-163}$ |

Focusing on valence (positive and negative), negative polarity triggers more engagement than positive polarity. The graph of the quantile function of engagement for the two categories clearly shows that negative tweets dominate the distribution of engagement and they are targets of more replies. The same behavior applies both to the fact-checkers (Figure 3) and the disinformation accounts, with an even starker divide for the latter compared to the former (Figure 4). This result does not align with studies that attested viral patterns associated more to positive rather than negative content in New York Times news (Berger and Milkman, 2012) or in social interactions among Firm's consumers (Godes et al. 2005). On the other side, it confirms Nanath and Joy's (2021) and Jiménez-Zafra et al.'s (2021) findings that negative emotions, instead of positive ones, are significant predictors of retweets, respectively in a dataset of COVID-19 related tweets and in the political domain. These mismatches show that both social media type and context need to be taken into account to understand engagement patterns across communities: while sharing positive emotions might serve a self-presentation purpose in the context of e-commerce (having positive experiencers increases ethos) or general news sharing, it is not as relevant in a pandemic context where topics associated to positive feelings (e.g. food, travel) are not easily available, or in the political arena, where social media are used to advocate for changes, serving propaganda purposes.

Similarly, a community of fact-checkers' followers is inclined to share debunked content which points to the dangerous consequences of disinformation (e.g "#violenzasulledonne: nella maggior parte dei casi le violenze non vengono denunciate. Tra marzo e maggio 2020 i casi di donne che non hanno denunciato sono in aumento rispetto al 2019. I numeri sulla #violenzacontroledonne in tempi di pandemia, #Violence over women: in most cases  violence is not reported. Between March and May 2020, the cases of women who do not report [violence] are increasing compared to 2019. The numbers on #violenceagainstwomen in times of pandemic).  On the other side, followers of disinformation accounts are good candidates to share negative content since feeling the social burden of disclosing harmful conspiracies (e.g "La scrittrice Sonia Savioli racconta su #Byoblu24 i retroscena inquietanti di questa "pandemia". Autrice del libro "Il Giallo del Coronavirus", spiega che la vera minaccia non è il virus ma "un progetto disumanizzante e autocratic"", 'The writer Sonia Savioli tells the disturbing background of this "pandemic" on #Byoblu24. Author of the book "Il Giallo del Coronavirus", explains that the real threat is not the virus but "a dehumanizing and autocratic project""). In both communities, apprehension about negative outcomes of events constitutes a key driver in the (mis)information ecosystem.

As far as emotions are concerned, tweets classified as joy trigger less retweets and likes in both datasets (fact-checks ---FT--- and disinformation ---DS---); however, anger plays a more prominent role in the FT dataset, while fear is the primary engagement-drive in the DS dataset. The expression of sadness is, among negative emotions, the one bringing less retweets and likes. This trend confirms the role played by arousal in prompting transmission (Berger, 2011): while both anger and fear are emotions characterized by high arousal, sadness and joy are characterized by low arousal.

Zooming into active engagement, the number of replies to emotion types differs from 'passive' engagement and varies across datasets: while in the FT dataset anger is the emotion firing up more replies, followed by sadness, fear and joy, in the DS dataset anger is still on top, but fear outperforms sadness, followed by joy.

Table 5: Average engagement scores per type of emotion

| | Average engagement fact checkers | Average engagement disinformation accounts |
|---|---|---|
| Anger | 10.628940092165898 | 72.94480435345945 |
| Fear | 8.80192926045016 | 89.45438824038045 |
| Sadness | 8.09787018255578 | 64.75536220153784 |
| Joy | 8.360907271514343 | 38.88088137839629 |
| Positive | 7.024793388429752 | 36.833483024968565 |
| Negative | 9.962329147683075 | 74.0126823946543 |

Figure 4: Quantile plots of engagement (likes+retweets) and replies for the tweets of the fact-checkers, divided by emotion and polarity.

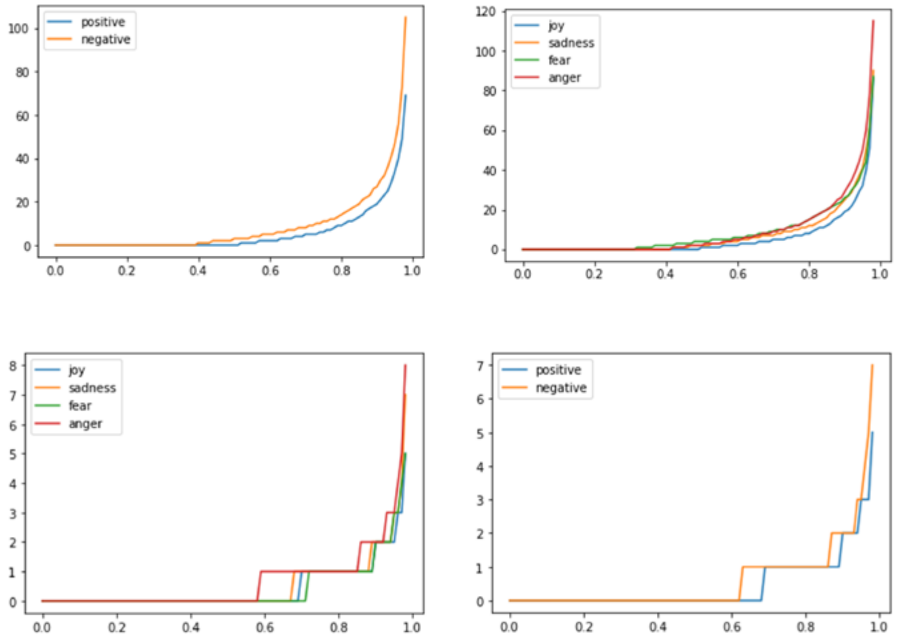
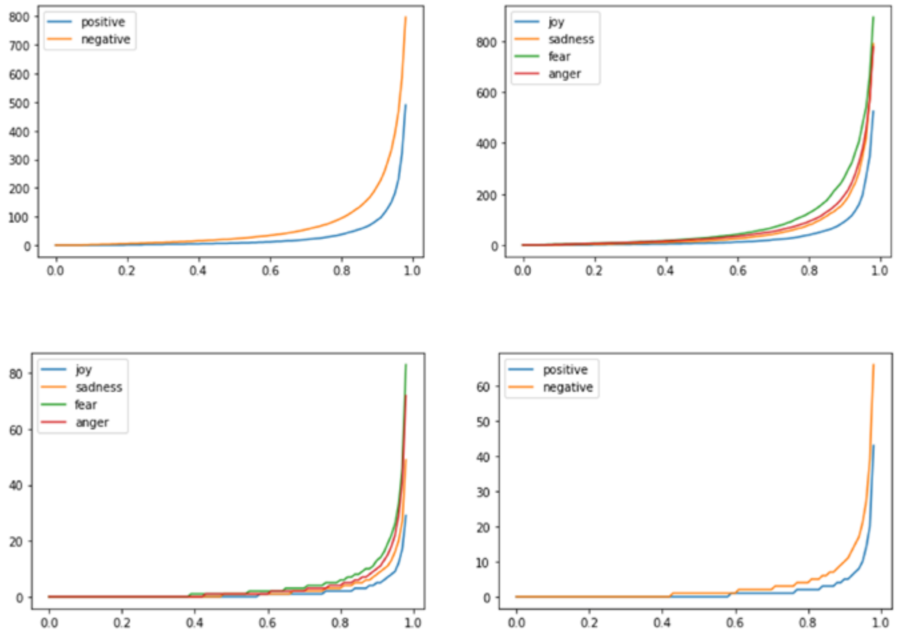


Figure 5: Quantile plots of engagement (likes+retweets) and replies for the tweets of the dis influencers, divided by emotion and polarity.



The popularity of sadness content in the FT dataset does not confirm the hypothesis that arousal emotion are predictors of engagement. A possible explanation is that retweets and likes represent different types of 'actions' with respect to 'replies': the former presuppose agreement with the content of the original tweet (and thus a sharing of the emotion it expresses), while the latter can mark the presence of disagreement, or anyways misalignment with the original tweet (in terms of polarity or intensity of the emotion). The replies do not, in fact, always express the same emotions that are expressed in the tweet they are reacting to:

Table 6: Contingency table emotion FT * emotion replies.

| emotion | anger | fear | joy | sadness | TOT |
|---|---|---|---|---|---|
| anger | 2797 | 442 | 484 | 676 | **4399** |
| fear | 285 | 110 | 54 | 78 | **527** |
| joy | 601 | 98 | 181 | 164 | **1044** |
| sadness | 639 | 138 | 134 | 214 | **1125** |
| **TOT** | **4322** | **788** | **853** | **1132** | **7095** |

On the contrary, the replies tend to be of the anger type across the board, regardless of the emotion type expressed by the original tweet. This is not surprising, if we consider fact-checking as a service that users rely on to be well informed and to share truthful information. Users' comments can be conceived as reviews of the service offered: as underlined by research in e-WOM (Sen and Lerman, 2017), there is a widespread tendency for users to write negative rather than positive reviews since they are perceived as more informative. Adopting such as frame, it is understandable why users tend to reply to Fact-checkers' tweets which deliver sadness content, and might thus hinder public attitudes spreading depressive feelings, with comments that express (for the most) disagreement with high tones (e.g. "Che schifo dar spazio alle intenzioni d'un #fascista! [...]" http://twitter.com/enzo2308/, 'How disgusting to give space to the intentions of a #fascist!'; "Mi spiegate come mai un portale che si occupa di bufale (non l'animale) tenga in considerazione i sondaggi? Sono uno strumento di un'inattendibilità abissale, inaffidabili e fallaci", https://twitter.com/Luigi90397177/, 'Can you explain to me why a portal that deals with buffaloes (not the animal) takes polls into consideration? They are an instrument of strong unreliability, unreliable and fallacious').

Turning to the DS dataset, the ratio of replies is higher overall, confirming that the communities which follow disinformation accounts are more engaged both passively and actively. This is partially due to the widely documented strategy (Guo et al., 2019; Zhang et al., 2019), which applies also to our dataset, used by disinformation accounts of crafting messages with contain inflammatory language (classified as anger comments) to arouse a like-minded crowd:

Table 7: Contingency table emotion DS * emotion replies

| emotion | anger | fear | joy | sadness | TOT |
|---|---|---|---|---|---|
| anger | 64340 | 16481 | 7769 | 10834 | **99424** |
| fear | 4098 | 2588 | 608 | 973 | **8267** |
| joy | 9671 | 2943 | 2231 | 1929 | **16774** |
| sadness | 10005 | 3056 | 1685 | 2517 | **17263** |
| **TOT** | **88114** | **25068** | **12293** | **16253** | **141728** |

Differently from the FT datasets, most replies classified as *anger* do not express disagreement with the original tweets, but rather amplify and intensify their sentiment through a resonance effect (e.g. "Esatto! Chi nega è un Padrone, padroncino, servo dei Padroni", 'Exactly! Who denies is a Master, little master, servant of the Master'). Observing the replies commenting on original tweets classified as *sad*, it becomes apparent that a similar bandwagon mechanism is in place and even furthered. The replies tend, in fact, to provide arguments in favor of the interpretation advanced in the original tweet:

### 5.1.1. Example

Original tweet

*"Guardiamo i dati di Australia e Argentina: in entrambe le realtà il confinamento non solo non ha prodotto benefici medico-sanitari, ha addirittura generato conseguenze disastrose".* #RadioAttività *Con* @DiegoFusaro

*"Let's look at the data from Australia and Argentina: in both realities the lockdown has not only produced no medical-health benefits, it has even generated disastrous consequences".*

*User:*
*A conferma... La curva delle morti in Svezia*
*"To confirm... the death curve in Sweden"*

In this example, the reply strengthens the claim that the lockdown has only negative effects, anticipating the potential counterargument that there is no evidence the absence of a lockdown would have not made things worse, presenting the low death curve in Sweden, where lockdown was not imposed, as an argument.

## 5.2. Topics and engagement

To investigate what topics trigger highest engagement across the DS and the FT corpora, we have selected all the tweets within 10% highest engagement score per corpus. We have, then, carried out STM (Structural Topic Modeling) to pull out topics. As a result, six main topics have emerged for the FT dataset and four for the DT dataset.

As a generative model of word counts, STM allows visualizing the topical contents, words used within a topic. The topical potential of the two datasets is respectively visualized in the two diagrams below:

Fig. 6 Word cloud topical content FT dataset with top engagement

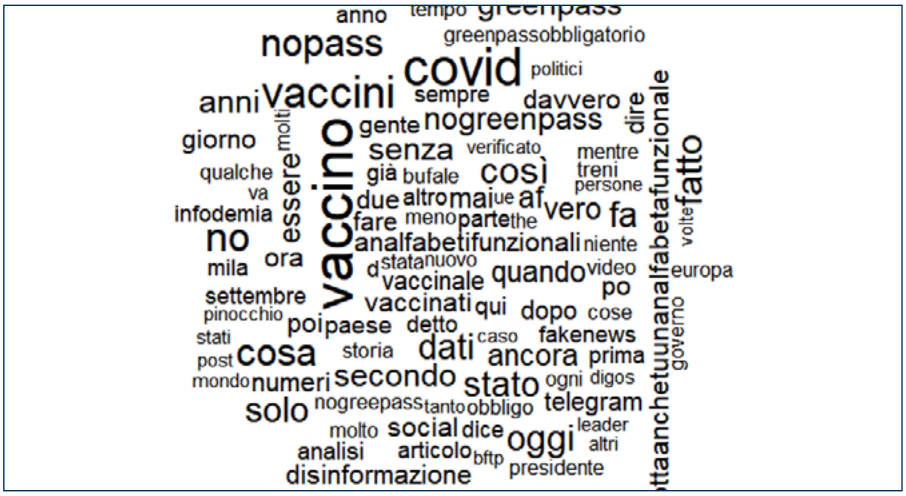
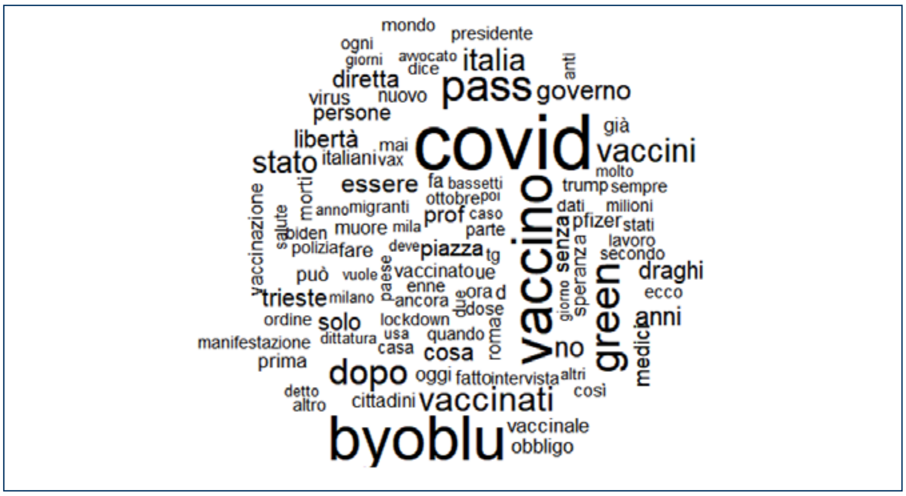


Fig 7. Word cloud topical content DS dataset with top engagement



From the comparison of the two word clouds, it emerges that covid and vaccino ('vaccine') are the most prominent themes in both datasets. Focusing on the entities referred to by the words in the two clouds, the DS dataset includes a few words expressing personal names (e.g. 'draghi', 'trump', 'biden') of political figures as well a common name designating categories of people (e.g. 'medici', 'migranti', 'polizia'). This is not surprising since blame culture constitutes a kernel of disinformation: the tendency of accounting for issues by finding fault with individuals or groups is a sense-making strategy at the very core of conspiracy theories (Locke, 2009).

### 5.2.1. Example

*"con l'introduzione dell obbligovaccinale con l estensione del greenpass con la discriminazione dei bambini non vaccinati a scuola che draghi ha annunciato oggi in conferenzastampa la deriva totalitaria è completa la lega continuerà a sostenere questo vile affarista"*

*"It is with the introduction of the vaccine obligation, with the extension of the greenpass, with the discrimination of unvaccinated children at school announced today by Draghi in the press conference, that the totalitarian drift is complete -- the party la lega will continue to support this cowardly businessman"*

In this tweet, for instance, a direct causal link is created between the measures announced by Draghi and the start of a totalitarian regime, portraying the politician as the responsible of a detrimental situation.

On the other side, the FT dataset is characterized by highly polarizing terms such as "nogreenpass', "nopass" and "antivax" which reduce the vaccination controversy to a binary choice. Such framing process risks discouraging that portion of the audience who does not believe in such a black and white scenario, and who is looking for scientific arguments to shape personal decision making processes. Moreso, when fact-checks contain satirical mottos aimed at gathering visibility, which might sound offensive: it is the case of the phrase "analfabeta funzionale" (lit. 'functional illiterate'), that has even been conventionalized as an hashtag (#adottaunanalfabetafunzionale. lit. '#adopta functionalilliterate') used by bufale.net to make reference to the antivax community:

### 5.2.2. Example

*"C'è una pagina, Dea, che in un post ad alto contenuto cringe e che affolla gli ambienti degli analfabeti funzionali mostra un foglio stampato e fotografato proveniente da Israele. #complottismo #covid19 #greenpass #israele #seconda-fase #vaccino"* http://twitter.com/Bufalenet/status/1353326921002573824

*"There is a page, Goddess, that in a post with a high cringe content and which crowds the environments of functional illiterates, it shows a printed and photographed sheet from Israel.' #complottism # covid19 #greenpass #israel #second-phase #vaccine"*

In this tweet published by *Bufale.net* the phrase is used to make reference to the spread of a news article according to which the vaccine in Israel will have to be repeated every six months for the foreseeable future as a means of surveillance. Even though the news constitutes an instance of disinformation, name calling its readers as *functional illiterates* might not be the most effective tactic to change their view. Face-threatening acts (Brown et al., 1987) are, in fact, bound to inhibit an audience which feels misperceived by the interlocutor (the fact-checker in our case).

# 6. Conclusion: recommendations for fact-checkers

The multi-layer analysis of the FT and the DS corpora shows commonalities and differences both in the tweets' content and the communities who access and share them, which suggests a roadmap to improve the efficacy of the fact-checking process.

Drawing form the analysis we identified a set of 5 recommendations for fact-checkers to counter the spread of disinformation:

- **Primary focus on making the audience more active rather than wider**: the social network analysis reveals that audiences (followers) of fact-checkers and disinformation accounts do not significantly differ in size, but rather in behavior with a much more active community spreading misleading information. Followers of disinformation accounts tend on average to retweet more frequently than fact-checkers' followers. This suggests that fact-checkers' shall, in the first place, try to make their audience more active rather than larger, to guarantee wider visibility of their content: when retweeted, fact-checked content becomes accessible to communities who do not directly follow fact-checkers' accounts. These communities are more infodemically vulnerable compared to those who deliberately choose to engage with fact-checked information.
- **Prompt messages likeability to advance message popularity**: the network analysis has also revealed two trends which dominate the flows of audiences' activity: first, the distribution of retweets per user follows preferential attachment, with users retweeting content from an outlet at a rate proportional to the number of interactions which feature that outlet; second, disassortative patterns show that, regardless the dataset, most active followers equally share most popular (number of likes + retweets) and least popular tweets in the same vein, while passive followers tend to share most popular tweets only. The most effective strategy for fact-checkers to gain more active users is, thus, that of increasing the likeability of the largest possible subset of tweets: tweets that attract the most likes and, thus visualizations, promise to engage both active and passive users in re-sharing activities.
- **Craft messages inducing high-arousal negative emotions to foster passive engagement**: the results of the content analysis in terms of both sentiment/emotion and topic offer hints about best practices to foster audience engagement. As to sentiment, regardless of the dataset, messages with negative polarity trigger both more active and passive engagement. Zooming into negative emotion types, those entailing arousal (anger and fear) catalyze more passive engagement, with anger being more prominent in the FT than in the DS dataset where fear constitutes the major drive. This situation is in line with the the general audience motivation of sharing content to be useful to their peers: on one side, fake news debunked as blatantly false (and thus through 'angry terms'), are felt as the most worth resharing since perceived as harmful, while disinformation which induces-fear is likely to be shared to warn about bad outcomes perceived as worth disclosing by followers of conspiratorial thought. Overall, a stylistic strategy to promote likes and retweets of fact-checks seem to lie in the stylistic choice of terms which express a critique through intense rather than neutral tones. In this way fact-checks of news which did not originally contain angry tones despite being detrimental will more likely receive attention.

- **Design messages which prompt replies expressing agreement rather than disagreement**: when it comes to active engagement the DS dataset contains more replies, confirming that followers of disinformation accounts are more active. The relation emotion-type/replies shows slightly different patterns between FT and DS: in both corpora content evoking anger is the one inducing more replies, but in the FT dataset messages expressing sadness outperform those expressing fear, while in the DS dataset the trend is reversed. The emotion induced by replies to messages and the messages themselves tends to be the same in the DS dataset, while this is not the case in the FS dataset where users, for instance, criticize fact-checks that evoke sadness. It is clear that disagreement patterns slow down fact-checks' popularity: instead of amplifying their content they might cause a potential backfire effect flagging fact-checks as untrustworthy.
- **Avoid polarizing expressions and name calling to counter blame culture**: the topic modeling analysis has, not surprisingly, shown that covid and vaccine represent the hottest topics in both datasets. However, the comparison with the other topics shows that, in the DS dataset, single political figures or categories are in focus, in line with blame culture, while the FS dataset privileges highly polarizing terms and sarcastic expressions which frame controversial issues as dichotomic choices. Such a rhetorical choice is risky since it, on one side, excludes the audience which does not recognize in clear cut parties and, on the other, might trigger angry reactions radicalizing the debat, especially when name calling expressions are used.

## References

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. *Electronics*, *10*(11), 1348.

Amazeen, M. A. (2016). Checking the fact-checkers in 2008: Predicting political ad scrutiny and assessing consistency. Journal of Political Marketing, 15, 433–464.

Anoop, K., Deepak, P., & Lajish, V. L. (2020, August). Emotion cognizance improves health fake news identification. In *IDEAS* (Vol. 2020, p. 24th).

Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, *164*, 114006.

Berger, J. 2011. "Arousal Increases Social Transmission of Information." Psychological Science 22 (7): 891–893.

Berger, J., and K. L. Milkman. 2012. "What Makes Online Content Viral?" Journal of Marketing Research 49 (2): 192–205

Bianchi, F., Nozza, D., & Hovy, D. (2021). Feel-it: Emotion and sentiment classification for the italian language. In *The 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Borgatti, S. P., & Everett, M. G. (2000). Models of core/periphery structures. *Social networks*, *21*(4), 375-395.

Brandtzaeg, P. B., Følstad, A., & Chaparro Domínguez, M. Á. (2018). How journalists and social media users perceive online fact-checking and verification services. Journalism practice, 12(9), 1109-1129.

Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Burel, G., Farrell, T., Mensio, M., Khare, P., & Alani, H. (2020, October). Co-spread of misinformation and fact-checking content during the Covid-19 pandemic. In *International Conference on Social Informatics* (pp. 28-42). Springer, Cham.

Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675-684).

Cohen, Geoffrey L.; Sherman, David K.; Bastardi, Anthony; Hsu, Lilian; McGoey, Michelle; Ross, Lee (2007). "Bridging the partisan divide: self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation". Journal of personality and social psychology, v. 93, n. 3, pp. 415-430.
https://doi.org/10.1037/0022-3514.93.3.415

Graves, L., & Cherubini, F. (2016). The rise of fact-checking sites in Europe. Available at:

https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Rise%2520of%2520Fact-Checking%2520Sites%2520in%2520Europe.pdf

Graves, L., Nyhan, B., & Reifler, J. (2016). Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. Journal of Communication, 66, 102–138. doi:10.1111/jcom.12198

Guo, C., Cao, J., Zhang, X., Shu, K., & Yu, M. (2019). Exploiting emotions for fake news detection on social media. *arXiv preprint arXiv:1903.01728*.

Himelboim, I. (2017). Social Network Analysis (Social Media). In The International Encyclopedia of Communication Research Methods (eds J. Matthes, C.S. Davis and R.F. Potter). https://doi.org/10.1002/9781118901731.iecrm0236

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600).

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., ... & Verlegh, P. (2005). The firm's management of social interactions. *Marketing letters*, *16*(3), 415-428.

Grootendorst, R., & Van Eemeren, F. H. (2004). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Jiang, S., & Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1-23.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital journalism*, *4*(1), 89-106.

Jiménez-Zafra, S. M., Sáez-Castillo, A. J., Conde-Sánchez, A., & Martín-Valdivia, M. T. (2021). How do sentiments affect virality on Twitter?. *Royal Society Open Science*, *8*(4), 201756.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Locke, S. (2009). Conspiracy culture, blame Guo et al., 2019, Zhang et al., 2019
culture, and rationalisation. *The Sociological Review*, *57*(4), 567-585.

Lu, Y., Zhang, P., Cao, Y., Hu, Y., & Guo, L. (2014). On the frequency distribution of retweets. *Procedia Computer Science*, *31*, 747-753.

Marietta, M., Barker, D. C., & Bowser, T. (2015, December). Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities?. In *The Forum* (Vol. 13, No. 4, pp. 577-596). De Gruyter.

Nanath, K., & Joy, G. (2021). Leveraging Twitter data to analyze the virality of Covid-19 tweets: a text mining approach. *Behaviour & Information Technology*, 1-19.

Negara, E. S., Triadi, D., & Andryani, R. (2019, October). Topic modelling twitter data with latent dirichlet allocation method. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)* (pp. 386-390). IEEE.

Nieminen, S., & Sankari, V. (2021). Checking politifact's fact-checks. *Journalism Studies*, *22*(3), 358-378.

Sen, S., & Lerman, D. (2007). Why are you telling me this? An examination into negative consumer reviews on the web. *Journal of interactive marketing*, *21*(4), 76-94.

Shu, K., Bernard, H. R., & Liu, H. (2019). Studying fake news via network analysis: detection and mitigation. In *Emerging research challenges and opportunities in computational social network analysis and mining* (pp. 43-65). Springer, Cham.

Temmerman, M., Moernaut, R., Coesemans, R., & Mast, J. (2019). Post-truth and the political: Constructions and distortions in representing political facts. *Discourse, Context & Media*, *27*, 1-6

Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, *25*(2), 162-180.

van der Hoorn, P., & Litvak, N. (2015). Degree-degree dependencies in directed networks with heavy-tailed degrees. *Internet mathematics*, *11*(2), 155-179.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. Political Communication, 37(3), 350-375

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135-163.

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021, April). Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021* (pp. 3465-3476).

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1-40.

LUISS

Data Lab