When fact-checks go viral: a cross-national analysis of the dissemination of European fact-checkers on Twitter

Lorenzo Federico*, *LUISS University,*
lfederico@luiss.it

Mariavittoria Masotina, *University of Liverpool,*
mariavittoria.masotina@liverpool.ac.uk

*Corresponding author

Abstract:
To be effective in countering misinformation, it is paramount for fact-checkers to reach a wide audience. This study investigates the dynamics that lead to broader engagement with fact-checking content published on social networks. Specifically, it analyzes the dissemination activity on Twitter of a cross-national sample of European fact-checkers over a span of 4 months. We employ Network Analysis and Natural Language Processing techniques (sentiment analysis and keyword extraction), to address four questions: 1. Are there specific tweets that attract the majority of engagement?; 2. Do these tweets draw engagement from audiences beyond their usual reach?; 3. What is the prevailing sentiment expressed in these tweets—positive, neutral, or negative?; 4. What topics are covered in these highly engaging tweets? The results show that certain tweets receive significantly higher levels of engagement than the average and that this engagement extends beyond the typical audience. Furthermore, our findings suggest that while the topics of the most popular tweets are country-specific, audiences in most of the considered countries tend to interact more with tweets expressing negative sentiments.

Keywords: fact-checking, sentiment analysis, social networks, keywords extraction

1. Introduction

The last decades have seen an unprecedented rise in the availability to the public of enormous quantities of unfiltered information thanks to the development of the so-called networked society. The mass adoption of the World Wide Web and in particular of large-scale social media has made it easy for any individual or organization to reach out to massive audiences. Inevitably this has resulted also in the broad diffusion of false or unreliable information both by careless users (misinformation) and malicious actors (disinformation). Any major event in recent

years has been accompanied by a great wealth of dangerously unverified or outright fabricated information such as the *Infodemic* recognized by the World Health Organization besides the Covid-19 pandemic (Wilhelm et al., 2023). To help the public to navigate the vast mass of available news and tell apart correct information from misleading, false or unreliable ones, specialized fact-checking organizations have become commonplace. These outlets have the mission to parse the information circulating in the online infosphere and verify its correctness. It applies both to the pieces of information coming from official sources such as politicians and news outlets and to the one circulating virally from below, pushed by common users. While many fact-checking organizations have developed spontaneously around the world over the years, and the specific practices of each of them can vary significantly, there are recognized international organs, such as the International Fact-Checking Network (https://www.poynter.org/ifcn/), that coordinate their activities and set the professional standards they have to adhere to.

The scholarly discussion on the effectiveness of fact-checking as a tool to contain the diffusion of misinformation and disinformation has resulted in the identification of several critical points. The first is the difficulty for the fact-checkers to keep up with the enormous volume of new available information. The verification process of news is intrinsically more time-consuming than its production, in particular considering the increasing ease of access to content production tools as Generative AI, which is increasingly capable of producing both text that is hard to distinguish from human-written one and highly realistic images and voices (Kreps et al., 2022). Second, fact-checking content could be unable to reach out to the audience of people who are more vulnerable to misinformation and disinformation or could be ineffective, triggering cognitive biases that reinforce the beliefs of those already convinced of false information. Due to cognitive biases such as confirmation bias, most people are unlikely to search for material that contradicts their beliefs and are more likely to react negatively when they are exposed to it (Oswald, & Grosjean, 2004; Beauvais, 2022).

The literature on the quantitative study of online communication is extensive, but little attention has been given to making fact-checking content more visible for the general public. The necessity to employ complex Network Analysis tools was understood rapidly after the mass adoption of the World Wide Web as a major venue to exchange information (Otto, & Russeau, 2002). This has in parallel given great impulse to the theoretical development of new models of complex networks, and of new parameters to investigate their properties (see e.g., Newman 2001).

Twitter (now rebranded as X) has been a particularly important environment for the development of the theory and practice of social network analysis, thanks to the simplicity of its interactions and the comparatively open access to data offered for years to academics (see Karami et al., 2020 for meta-study of Twitter-related research). This analysis of the dynamics behind the spread of content has been applied to the diffusion of misinformation (Caldarelli et al., 2021), as well as other forms of online discourse, such as political communication (Gaumont et al., 2018) and marketing (Liu et al., 2021). In these contexts, factors such as the sentiment

expressed during the communication and the topics covered were taken into account as being drivers for enhancing engagement from the public.

The academic study about the impact of fact-checking instead has been much more interested in the effect that the fact-checking content has on the users that consume it, rather than on its ability to reach a broad and diverse audience. In most papers, such as those considered in the meta-study of the impact of fact-checking by Walter et al (2020), the fact that the participants of the study are reading fact-checking is taken as an assumption. This is true also for studies on the cognitive biases that can be triggered by fact-checking. An example is represented by the backfire effect (strengthening of wrong beliefs after being exposed to fact-checks), whose importance is debated in the work of Swire-Thompson et al (2020). Much less common are studies on the spreading dynamics of fact-checking content. the study on the Covid-related disinformation and fact-checking by Burel et al (2020) represents an exception. Here, the authors called in their conclusions for further studies in this area, possibly using both semantic and Network Analysis. Nonetheless, to the best of our knowledge, no study specifically delved into the dynamics that lead to broader engagement with fact-checking content published on social networks.

In this paper, we will focus on the study of if and how the content produced by fact-checkers is able to reach out beyond the core audience of users who are regularly and proactively looking for it. While it is desirable that users who are looking for fact-checking can easily find it, typically these users are not the most infodemically vulnerable. It is thus even more important that even those who are not regularly consuming fact-checking content and are not proactively researching it, are exposed to it. For this purpose, it is crucial to understand which messages from the fact-checkers are the most likely to reach out to an audience outside of those who regularly search for them.

2. The present study

The present study aims at understanding the trends of engagement with fact-checking content. We analyze the outreach activity of 19 European fact-checking outlets on Twitter (now rebranded as X), looking at the posts published on their official accounts and the engagement they received over a 4-month period from late 2022 to early 2023. The final aim is to understand which characteristics of the tweets published by the fact-checkers attract the most engagement (measured by the number of retweets) - in particular, from users who do not interact regularly with these accounts - and if there are significant differences in the patterns of engagement among different countries.

We tackle this problem using a multidisciplinary approach. We combine Network Analysis techniques to understand the engagement patterns between the tweets published by the fact-checkers and the broader Twitter public, and Natural Language Processing (NLP) to focus on the specific textual characteristics of the most popular

tweets for each fact-checker. Our study aims to answer four research questions:

- **RQ1:** Are there specific tweets that attract most of the retweets or are retweets spread out rather uniformly over all tweets?
- **RQ2:** Do the most popular tweets attract a higher proportion of their retweets from outside the usual audience than less popular ones?
- **RQ3:** Is it more likely for tweets expressing positive, neutral or negative sentiment to be among those with the highest number of retweets?
- **RQ4:** What are the topics touched on in the most popular tweets?

We will use Network Analysis techniques to answer RQ1 and RQ2, and NLP to answer RQ3 and RQ4. In this study, we divide the fact-checkers by nation and explore which patterns of engagement emerge. We also assess whether the patterns of engagement are consistent across Europe or change according to the information environment of each specific country.

3. Data and Methods

3.1 Data collection and filtering

For the purpose of this study, we used the now-discontinued Twitter Academic API to download the activity of 69 European fact-checkers affiliated with the International Fact-Checking Network (https://ifcncodeofprinciples.poynter.org/signatories) and the European Digital Media Observatory (https://edmo.eu/fact-checking-community/) in the period from 16/12/2022 to 16/03/2023. For each account, we gathered their own activity (original tweets, retweets, quote tweets and replies) and all the retweets, replies and quote tweets their tweets received. This resulted in a dataset of 1,210,094 tweets.

From the 69 accounts we then selected 19 to focus on based on the following filters:

- Accounts that were not specifically dedicated to fact-checking were discarded (e.g., @franceinfo, a news outlet that primarily publishes news, but produces also fact-checking contents).
- Accounts that were publishing tweets in multiple languages were also not considered, as that would impact the keyword extraction (e.g., @StopFakingNews, which posted both in English and Ukrainian).
- From the remaining accounts, we selected those with sufficient activity in the time span to make a statistical analysis possible, as measured by the total number of retweets received during the period considered (see Table 1).

The final list counts 19 accounts: ZDDK_ (Austria), DemagogCZ (Czech Republic), AfpFactuel (France), CheckNewsfr (France), decodeurs (France), Observateurs (France), correctiv_org (Germany), thejournal_ie (Ireland), PagellaPolitica (Italy), DemagogPL (Poland), fakenews_pl (Poland), JornalPoligrafo (Portugal),

Raskrikavanje (Serbia) , EFEVerifica (Spain), maldita (Spain), Newtral (Spain), veri_fi_cat (Spain), FerretScot (UK) and FullFact (UK).

These accounts covered a total of 11 countries and published tweets in 9 different languages.

3.2 RQ1 and RQ2 - Network Analysis

We start our investigation by employing Network theory to answer RQ1 and RQ2. Social Network Analysis is a well-established technique in social and information science, in which social interactions and information flow are abstracted as a network. Here individuals and/or pieces of content are represented as nodes of the network and various forms of social interaction (friendship, comments, reposts, co-authorships…) are represented as edges that link two nodes. We represent all the retweets received by each of the 19 fact-checkers as a *bipartite network,* that is, a network which contains two types of nodes, where connections are allowed only between nodes of different types (see Figure 1). This network, which is an example of a bipartite user-content network (cfr. Zhu et al. 2015), is identified by the following 3 sets:

- The right side of the network **T**, formed by all the tweets published by the fact-checker (this includes replies and quotes to tweets from other accounts) during the period examined.
- The left side of the network **U**, formed by all the users who retweeted at least one of the tweets during the same period.
- The edge set **E** where each edge joins a tweet to one of the users who retweeted it.

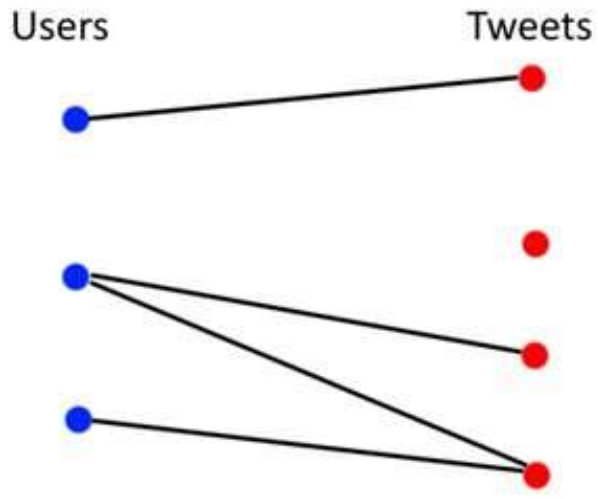

Figure 1. Representation of a small example of a bipartite user-content network.

This network thus encodes all the information about the number of retweets each tweet received and the identity of the individual retweeters.

	Total Tweets	Tweets with Retweets	Total Retweets	Individual Retweeters
FullFact	194	194	19418	10634
PagellaPolitica	1201	743	16490	6051
CheckNewsfr	611	426	14832	8042
AfpFactuel	670	633	13689	6024
thejournal_ie	4341	2420	12569	4975
Newtral	6521	3532	7857	3019
ZDDK_	1698	1627	7578	2026
FerretScot	737	542	5544	2707
JornalPoligrafo	4806	1110	5293	2472
correctiv_org	230	216	3041	1703
DemagogPL	886	497	2273	989
veri_fi_cat	647	257	1508	921
EFEVerifica	695	401	1498	718
maldita	513	338	1353	430
decodeurs	153	130	982	647
Observateurs	203	151	908	606
fakenews_pl	92	70	817	675
Raskrikavanje	108	82	813	291
DemagogCZ	206	125	795	495

Table 1. Number of tweets, retweets and individual retweeters for each account during the period considered for this study.

We have chosen retweets as the engagement measure to analyze in this study because the Twitter API allowed us to extract for each retweet the user id of the individual retweeter. This was not possible, for likes and impressions, as the Twitter APIs did not allow the extraction of the individual usernames but only of the total number of interactions.

We started by investigating RQ1, as the outcome of this first part of the study is necessary to ground RQ2-4. In fact, they build on the assumption that for each fact-checker there are some particularly popular tweets. We need to verify that, as is known in the literature to be the norm, see e.g. Lu et al. 2014, these distributions are heavy-tailed. This means that there is in each of these networks a small number of tweets that receive an amount of retweets that is much higher than the median and, equivalently, a small number of users that give a much higher number of retweets than the median. To do so, we measure the *degree distribution* on each side, that is, the distribution of the number of edges connected to a given tweet or users.

After computing the degrees of all tweets and users, we look for the tweets with the highest degree (i.e., those tweets that attract most of the engagement from the audience), to understand which properties are significantly more common among them than in the total corpora of tweets. We define for each fact-checking account the set of *hub tweets* as the set of all tweets whose degree is greater than $U^{1/3}$, where U is the total number of users that retweeted at least one tweet from that fact-checker. This threshold was set because networks which have nodes of such high degree show different macroscopic properties than those that do not, so we can treat them as playing a special role in shaping the engagement patterns (see e.g. Bhamidi et al 2020 for a technical explanation).

Further, we want to find if high-degree tweets are particularly able to attract the non-recurring audience to the fact-checking accounts. To do so we measure if said hub tweets are mostly receiving retweets from users that are not part of the core followers that are regularly retweeting the fact-checker's tweet. To do so in a rigorous way, we investigate whether these networks are *degree-disassortative*, that is, whether the tweets with very high degrees are more likely to be connected with the users of low degrees and vice versa. There are multiple measures of degree-assortativity for networks, and for our purpose, we will use the method introduced by van der Hoorn and Litvak (2015), which is designed to be robust even in the presence of very high-degree nodes. What we do is consider the set of all edges in the network (that is, of all retweets) as a bivariate dataset, where we record for each edge the degrees of the tweet at its right end and the user at its left end. We then compute the Spearman's ρ correlation coefficient between the tweet and user degrees as a measure of degree-assortativity of the network. A negative Spearman's ρ correlation coefficient indicates a disassortative network.

3.3 RQ3 - Sentiment Analysis

After identifying the hub tweets, we explored whether there were any significant differences in the expressed sentiment between them and the other tweets. We created a corpus composed by the collection of tweets described in Section 3.1 and employed the NLP technique of Sentiment Analysis (Medhat et al., 2014). Starting from the textual elements (lexicon in use and - depending on the model employed - emoji), this technique allows for the automatic classification of the sentiment expressed in a text at different levels of granularity. It has found application across various domains, including business, marketing, politics, health, and public initiatives (Drus & Khalid, 2019). When applied to the study of the phenomenon of tweet virality, previous studies used the polarity of the tweets as an independent variable and considered whether the lexicon employed conveyed a positive, negative, or neutral sentiment. The results showed that if a tweet has a positive or negative sentiment, it is more likely to be retweeted. Nonetheless, the effect was found to be stronger and more consistent for negative tweets (e.g., Stieglitz & Dang-Xuan, 2013; Bhattacharya et al., 2014; Jiménez-Zafra et al., 2021). Other studies employed more granular level of analysis, by categorizing not only the polarity of the tweets but also

the emotions expressed. For instance, Nanath and Joy, (2023) explored the factors that affect Covid-19 content sharing by Twitter users. Among the other metrics considered, they used NLP to categorize tweets in English according to the emotion expressed as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Also in this context, tweets containing negative emotions such as anger, disgust, fear, and sadness were more likely to be retweeted.

In our study, we concentrated on polarity and categorized tweet sentiment as positive, neutral, or negative. We took into consideration that, first, multi-label emotion classification is in general less accurate than polarity classification and, second, that pre-trained models accurate in classifying fine-grained emotions are specialized in a limited number of languages (e.g., BETO, a BERT-based model pre-trained for NLP tasks in the Spanish language; Cañete et al., 2023). The multilingual nature of our data would have required us to change the model in use during the analysis for the different country-based corpora, possibly introducing biases in the interpretation of potential differences across countries.

Our analysis followed four steps:

1. We pre-processed the texts of the tweets by removing the URL links, since they represented parts of the text not conveying information about the sentiment.
2. We grouped the accounts (along with their associated tweets) by country since there could be context-specific differences associated with the country that would not have been recognized by considering them all together. Consequently, we obtained 11 country-based subcorpora of tweets. Each subcorpus was monolingual and the language used was the one characterizing the country (e.g., Italian for Italy, English for the UK, etc.).
3. We categorized each tweet as expressing a negative, neutral, or positive sentiment using the OpenAI DaVinci-003 model (https://platform.openai.com/docs/models/gpt-3-5). This model is capable of handling different languages and has been proven to be accurate in this task across them (Ye et al., 2023; Rathje et al., 2023).
4. For each country-based subset, we conducted a series of statistical tests to examine whether there were significant differences in the proportions of tweets expressing negative, neutral, and positive sentiment between hub and non-hub tweets. When the expected frequencies of at least one group of tweets were below 5 (e.g., hub tweets expressing negative sentiment for the Czech Republic subset), we employed Fisher's exact tests, otherwise we run chi-squared tests. If the tests resulted significant, we run pairwise proportion tests with Bonferroni correction to identify which group was significantly different. All analyses were performed using RStudio Pro 2023.09.1 with a significance threshold set at $p < .05$.

3.4 RQ4 - Keywords extraction

After considering the expressed sentiment, we delved into the content of the hub tweets to explore whether certain topics were more likely to attract users' engagement. To address RQ4, we employed the NLP technique of keyword extraction and conducted a thematic analysis on the identified keywords, i.e., on those terms or phrases that succinctly represent the main topics within a piece of text. In the context of NLP and text analysis, keywords are crucial for identifying the key components of a text and understanding its central themes. As for RQ3, we considered the tweets grouped by country, and we began by pre-processing the texts. During this stage, we also removed emojis since they could be helpful for assessing sentiment (RQ3) but not informative about the linguistics contents.

Next, we utilized SketchEngine software (https://www.sketchengine.eu/) to extract a list of the nouns contained in the hub and non-hub tweets in their vocabulary form (i.e., lemma) with their associated frequencies. In this way, we obtained two frequency lists for each country-based subset: one for the hub tweets and one for the non-hub tweets. From these lists, we used the AntCONC software (https://www.laurenceanthony.net/software/antconc/) to extract the keywords. The software compared the two frequency lists by running a series of Log-Likelihood tests with Bonferroni correction – one for each country-based subset. This process allowed us to extract those words that characterized the hub tweets compared to non-hub tweets – i.e., those lemmas that were statistically more frequent in the hub tweets (target subcorpus) than in the non-hub tweets (reference subcorpus). As for RQ3, we set $p < .05$ as the significance threshold.

The final step involved a manual thematic analysis of the keywords found. We translated the keywords taking into account their context (i.e., the meaning they conveyed in the tweets) and categorized these words using a bottom-up approach (i.e., starting from the keywords and creating categories accordingly).

4. Results

4.1 RQ1 - Heavy-tailed degree distributions

We first present the answer to RQ1, as most of the other results hinge on knowing the degree distribution of the bipartite networks for both users and tweets. We found out that in the bipartite networks we created, both degree distributions are heavy-tailed, as is shown in the log-log plots in Figure 2. That said, we saw that the specificities of the distributions of degrees of users and tweets were quite different.

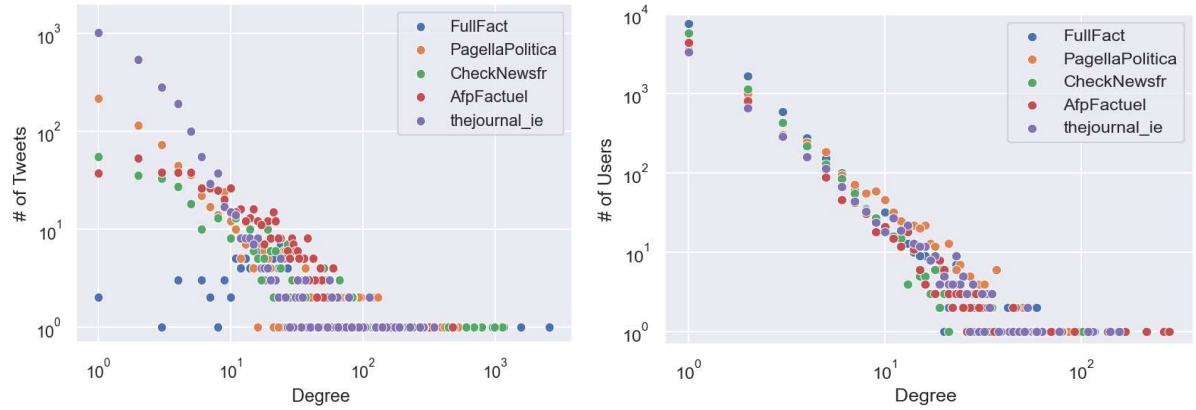

Figure 2. Log-log plots of the frequencies of degrees of tweets (left) and users (right) for the 5 fact-checking accounts with the highest number of retweets.

The degrees of the users followed very closely a power law, that is, the proportions of users of degree d in a network decrease proportionally to $d^{-\tau}$ for some $\tau<1$ as d grows. This can be observed as the right log-log plot in Figure 2 closely resembles a line with a negative slope for each fact-checker. This is often a sign of what is called a *preferential attachment* dynamic (Albert and Barabási 2002), that is, users tend to retweet new tweets from a fact-checker with a rate that is proportional to the number of tweets they have already retweeted.

The degree distribution of tweets instead depended on the average frequency of tweeting for each account. As we see from the left plot in Figure 2, for fact-checkers that tweet at a very high rate, such as *thejounal_ie* and *PagellaPolitica*, the distribution still resembles a power law, while for accounts with less frequent tweeting, like *FullFact*, there is still a power law tail of high-degree tweets, but the distribution of low-degree tweets is flatter, without a high frequency of degree 1 tweets. This indicates that for each fact-checker there is a limited amount of tweets that can become highly popular. In case of an extremely frequent production of content, the other tweets will end up receiving little to no engagement.

After computing the degree distributions, we could separate the tweets for each account into hubs and non-hubs, following the rule presented in Section 3.2. As we expected, for most of the fact-checking accounts, hub tweets were quite rare. We see from Table 2 that they were the majority only in the case of FullFact and in multiple cases were less than 5% of the total tweets.

	Hub Tweets	**Non-hub Tweets**	**Percentage of hubs**
FullFact	127	67	65.46 %
PagellaPolitica	125	1076	10.41%
CheckNewsfr	107	504	17.51%
AfpFactuel	204	466	30.45%
thejournal_ie	91	4250	2.10%
Newtral	15	6506	0.23%

	Hub Tweets	Non-hub Tweets	Percentage of hubs
ZDDK_	65	1633	3.83%
FerretScot	80	657	10.85%
JornalPoligrafo	66	4740	1.37%
correctiv_org	65	165	28.26%
DemagogPL	39	847	4.40%
veri_fi_cat	21	626	3.25%
EFEVerifica	40	655	5.76%
maldita	71	442	13.84%
decodeurs	34	119	22.22%
Observateurs	25	178	12.32%
fakenews_pl	11	81	11.96%
Raskrikavanje	42	66	38.89%
DemagogCZ	24	182	11.65%

Table 2. Number of Hub and Non-hub tweets for each of the fact-checkers.

4.2 RQ2 - Measure of network disassortativity

We next show the results of the analysis of assortativity over the 19 bipartite networks we built. A positive assortativity coefficient means that the most popular tweets draw engagement from the regular audience of the fact-checker, while a negative one indicates that they draw engagement mostly from users who share content from the fact-checker sporadically. As explained in Section 3.2, we computed the Spearman's ρ coefficient between the degrees of the tweet and the user at the end of each edge of the network. The Spearman's ρ coefficient by definition takes values in the interval $[-1,1]$, with a coefficient of 1 (respectively, -1) indicating a perfectly increasing (respectively, decreasing) relation between the two quantities.

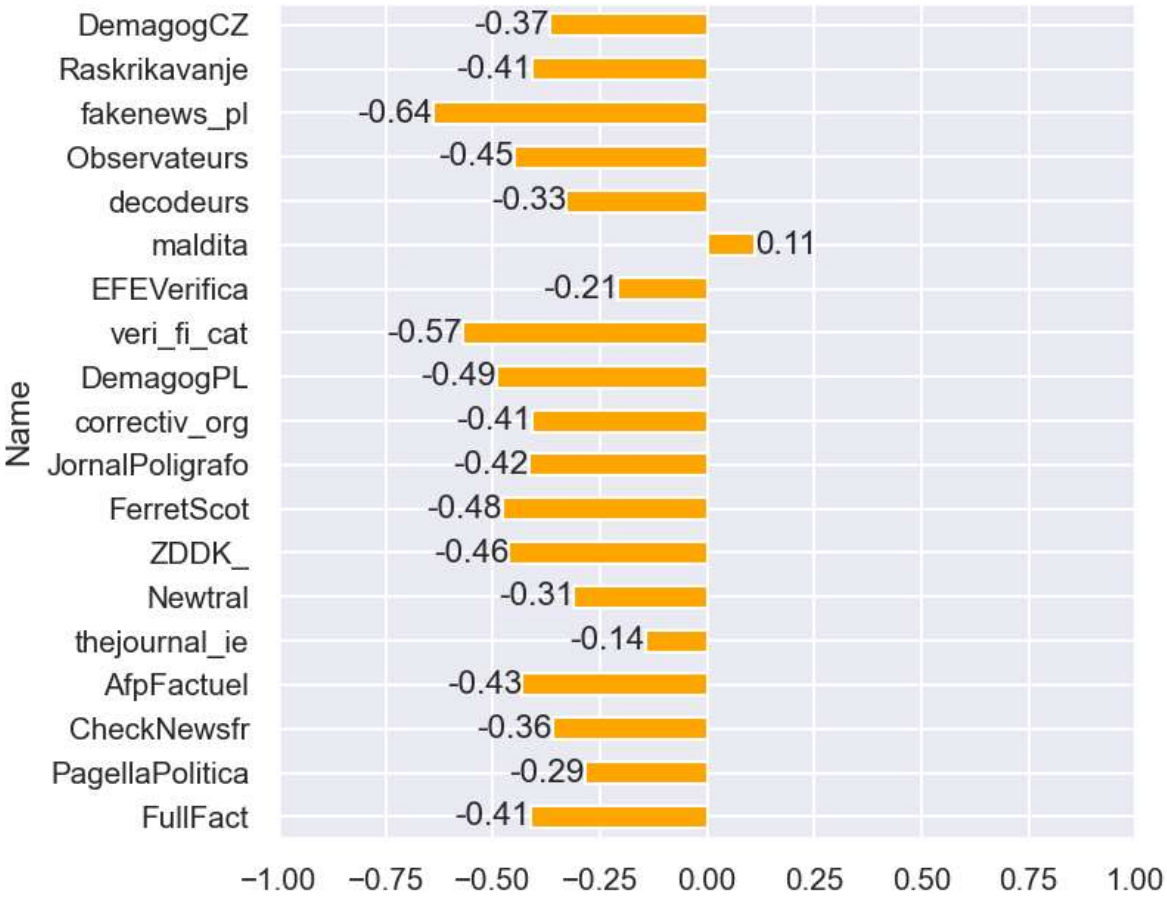

Figure 3. Barplot of the assortativity coefficients of all the networks. The bars represent the Spearman's ρ coefficients.

van der Hoorn and Litvak (2015) proved that in large random networks with a prescribed degree distribution, the coefficient is close to 0, so we interpret a negative Spearman's ρ as a sign of disassortativity and a positive one as a sign of assortativity. We found that, with the exception of *maldita*, all the networks were degree-disassortative, and actually all had an assortativity coefficient smaller than -0.1 (Figure 3). The existence of a single outlier to the disassortativity rule might be a sign that there could be something particular about *maldita*'s outreach tactics. In fact, this was the only case in our sample for which the hub tweets were more likely to be shared by its recurrent audience and not by other users who more rarely interacted with maldita's account.

4.3 RQ3 - Sentiment analysis

Regarding the sentiment of the tweets posted by fact-checker accounts, a neutral sentiment characterized the majority of the tweets in all the countries. Nonetheless, in most countries (Austria, Czech, France, Germany, Italy, Poland, Portugal, Serbia, UK) the audience tended to interact more with negative tweets, which were proportionally more likely to become hubs (Figure 4).

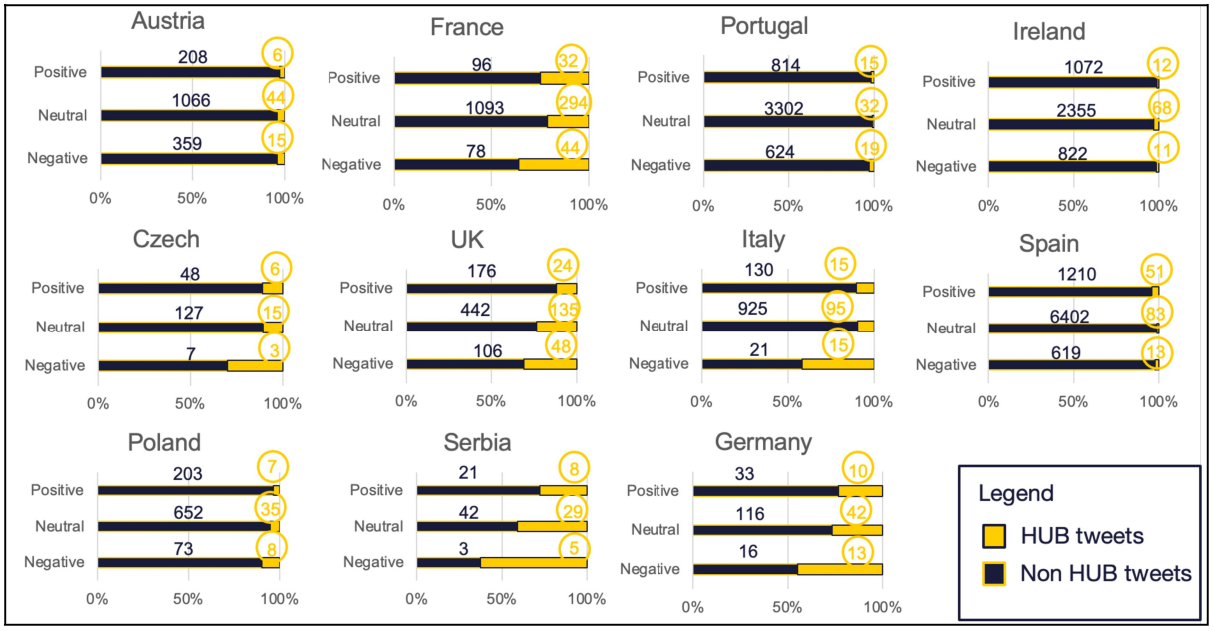

Figure 4. Percentage of positive, neutral, and negative tweets in hub tweets and non-hub tweets. The numbers represent the raw frequencies.

The results of the statistical tests (chi-squared and Fisher's exact texts, applied following the rationale specified in Section 3.3) showed, overall, a significant difference in the proportion of hub tweets depending on their sentiment in six countries (Table 3). The post-hoc tests confirmed that the trend observed at a descriptive level was significant in four countries, where negative hub tweets were in proportion more frequent than neutral (France, Portugal, Italy) and positive ones (Italy, UK) – Table 4. In the UK, positive tweets were less likely to become hubs also compared to neutral ones. For Ireland and Spain, no significant differences were found between tweets expressing negative sentiment and either positive or neutral ones. Instead, these countries exhibited a significant difference between positive and neutral hub tweets, but in opposite directions. In Ireland, neutral hub tweets were proportionally more frequent than positive, while in Spain, the situation was reversed.

Country	Test	X^2	df	p
Austria	X^2	0.70	2	.71
France	X^2	14.63	2	<.001***
Germany	X^2	4.68	2	.10
Ireland	X^2	13.56	2	.001**
Portugal	X^2	17.25	2	<.001***
Spain	X^2	47.17	2	<.001***
UK	X^2	19.68	2	<.001***
Czech	Fisher's exact test	-	-	.22
Italy	Fisher's exact test	-	-	<.001***
Poland	Fisher's exact test	-	-	.09
Serbia	Fisher's exact test	-	-	.17

Table 3. Tests of independence results. ** p = <.05, ** = p = <.01, *** = p = <.001.*

Country	Negative vs Neutral (*p* adj)	Negative vs Positive (*p* adj)	Positive vs Neutral (*p* adj)
France	<.001***	.23	1
Ireland	.07	1	.008**
Italy	<.001***	<.001***	1
Portugal	<.001***	.61	.18
Spain	.50	.10	<.001***
UK	.18	<.001***	.003**

Table 4. Post-hoc test results. p adj = p adjusted with Bonferroni correction. ** p = <.05, ** = p = <.01, *** = p = <.001.*

4.4 RQ4 - Keywords

For answering RQ4, we extracted the keywords by comparing the terms used in the hub tweets with the ones used in the other tweets. No terms were found to be significantly more frequent in the hub tweets than in the non-hub tweets in five countries (Austria, Czech Republic, Germany, Poland, and Serbia). This might suggest that the terms used in both the hub and non-hub tweets in these countries were the same. Nonetheless, these corpora were composed by a small number of tweets. Consequently, the absence of significant differences might be also due to the limited sample size.

Regarding the other countries, Table 5 reports all the keywords identified during the analysis.

Country	Keyword	% in hub	% in non-hub	LL
Ireland	assistance	47.25	0.26	246.13
	Gardai (*local police)*	67.03	3.51	182.03
	area	29.67	1.41	84.22
	height	15.38	0.64	46.60
	hair	14.29	0.73	39.11
	home	25.27	3.27	36.61
	Dublin	26.37	3.91	33.49
	appeal	15.38	1.39	29.87
	january	10.99	0.66	27.62
	eye	9.89	0.66	23.40
	Tallaght *(area of Dublin)*	7.69	0.35	22.18
Spain	telegram	33.33	0.38	269.29
	whatsapp	33.33	1.23	189.65

Country	Keyword	% in hub	% in non-hub	LL
	dia *('day')*	38.78	2.42	173.26
	malditatwitcheria (*maldita's twitch channel*)	17.01	0.18	138.83
	tema *('theme')*	17.69	0.51	110.13
	resumen *('summary')*	16.33	0.43	105.26
	canal *('channel')*	17.69	0.78	93.36
	redaccion *('editorial team')*	7.48	0.10	58.55
	curiosidad *('curiosities')*	7.48	0.11	56.93
	mito *('myth')*	8.84	0.30	51.63
	timo *('scam')*	7.48	0.16	51.53
	viernes *('Friday')*	8.16	0.24	50.24
	semana *('week')*	7.48	0.28	42.27
	bulo *('hoax')*	11.56	1.19	37.90
	spotifyivooxapple (*spotify channel*)	15.65	3.28	26.53
Italy	leader	12	1.95	22.03
	Salvini *(Matteo Salvini, politician)*	13.60	2.88	19.69
France	gouvernement ('*government*')	7.03	2.05	24.61
	femme ('*female*')	6.49	2.05	20.88
	contribuable (*'taxpayers'*)	1.62	0.00	19.73
	insulte (*'information obstraction'*)	1.62	0.00	19.73
	impots (*'taxes'*)	1.62	0.00	19.73
	reforme (*'reform'*)	17.30	10.50	19.23
Portugal	centimos ('*cents')*	13.64	0.63	43.20
	Costa (*Antonio Costa, politician)*	21.21	3.48	36.69
	Antonio (*Antonio Costa, politician)*	15.15	2.59	25.46
	margem *('margin')*	6.06	0.11	25.40
	gasolineira (*'gas station')*	6.06	0.11	25.40
UK	vaccine	10.63	0.69	38.68
	Sunak (*Rishi Sunak, politician)*	6.28	0.00	36.47
	Rishi *(Rishi Sunak, politician)*	5.80	0.00	33.66
	Ferret *(FerretScot, account)*	5.31	17.54	25.96
	BBC	3.86	0.00	22.44

Country	Keyword	% in hub	% in non-hub	LL
	MP (*member of parliament)*	10.14	1.80	21.02
	nurse	3.38	0.00	19.63
	asylum	8.21	1.24	19.23
	covid	4.83	0.28	18.37
	post	8.70	1.66	16.89
	heart	2.90	0.00	16.83
	democracy	2.90	0.00	16.83

Table 5. Results of the log-likelihood tests for each country. All the words included resulted to have p <.05 after Bonferroni correction.

Ireland and Spain exhibited a countertrend in the sentiment analysis. Considering the former, all the keywords in the hub tweets could be related to announcements about missing people. For instance, in Example 1, the account reports a call for help made by the Gardai—the local police force—to find a missing boy. The result suggests that the account provided this public service, and that people were likely to retweet these contents.

Example 1 (keywords in bold):

"**Gardai** are seeking the public's **assistance** in tracing the whereabouts of [name of missing person] (15) who has been missing from the **Tallaght area** of **Dublin** 24 since Tuesday 24 **January**. He is described as being approximately 5'3" in **height**, of slim build with short black **hair.**"

In Spain, instead, all the keywords could be related to self-promotion. For instance, in this tweet (Example 2), the account promotes a recurrent appointment on Twitch by describing the fact-checking activities they perform. Only the UK exhibited the same trend, as suggested by the keyword "ferret", which refers to the account's name (FerretScott).

Example 2 (keywords in bold):

" **Viernes** de **redaccion** en [link]! Vente a la **malditatwitcheria**: a las 11h00 desmentimos los **bulos** de la **semana**, te ensenamos a protegerte de los **timos**, , desmontamos **mitos** y charlamos sobre las **curiosidades** científicas"

'**Friday** with the **editorial team** at [link]! Come to **malditatwitcheria**: at 11:00, we debunk the **hoaxes** of the **week**, teach you how to protect yourself from **scams**, dismantle **myths**, and chat about scientific **curiosities**'.

Politics emerged as a recurring theme in the hubs across various countries. Notably, the terms used were linked to the names of local politicians (such as Matteo Salvini

in Italy, Antonio Costa in Portugal, and Rishi Sunak in the UK), institutions like “gouvernement” ('government') in France, political roles such as “leader” in Italy, and “MP” (Member of Parliament) in the UK. Additionally, the hubs reflected concerns related to local politics, as evidenced by the discussion of topics like the pension reform in France (see Example 3).

Example 3 (keywords in bold):

"**Pension** minimale a 1200 euros, impact de la **reforme** sur les **femmes**, mere de Darmanin... le **gouvernement** accroche a ses bobards".

‘Minimum **pension** at 1,200 euros, impact of the **reform** on **women**, Darmanin’s mother... the **government** sticks to its lies’

The other topics were specific depending on the Country and were not shared across them. In fact, those tweets addressed themes locally relevant during that time frame, as tweets disconfirming fake news about the price of diesel in Portugal (e.g., Example 4) or related to Covid in the UK (“BBC”, “covid”, “vaccine”, “nurse”, Example 5).

Example 4 (keywords in bold):

“Estado cobra 66 **centimos** por cada litro de gasoleo e **margem** das **gasolineiras** e de 12 **centimos**?”

‘The government charges 66 **cents** for each liter of **diesel** and the **gas station margin** is 12 **cents**?’

Example 5 (keywords in bold):

"Dr Aseem Malhotra claimed on **BBC** news that MRNA **Covid vaccines** were ‘a likely contributing factor’ in the number of excess deaths in the UK. But the **vaccines** will have reduced the rise in excess deaths, overall.”

While “post” referred to social media posts being fact-checked, and “heart” to contents related to health conditions (e.g., heart disease, heart attack), also the remaining keywords emerged in the UK could be connected with popular debates at the time, as the asylum seekers’ rights (“asylum”), or e.g., the suspension of journalists’ accounts by Twitter (Example 6).

Example 6 (keywords in bold):

"Media scrutiny is essential in any **democracy**. The suspension of several journalists from twitter is a reminder that our freedom of expression is too important to be left in the hands of any internet company.”

5. Discussion and Conclusions

In this study, we employed a multidisciplinary approach to explore engagement patterns in tweets published by fact-checkers across Europe. We started by utilizing Network Analysis techniques to determine whether certain tweets attracted the most engagement (RQ1). We found that the distribution of the number of retweets received by each individual tweet is heavy-tailed for all fact-checkers, but the distribution of low-degree nodes changes significantly. This suggests that these networks are generated by a preferential attachment process: on the one hand, tweets receive retweets at a rate proportional to the number of already received tweets; on the other hand, users retweet at a rate proportional to how many retweets they have already given. Such a phenomenon seems also compatible with the inner workings of the content recommendation algorithms, which give higher priority to popular content and tweets from accounts the user has already interacted with. Consequently, a relatively small proportion of tweets could attract the most engagement. This pattern was confirmed for all the accounts considered during the study. The same trend was observed in Zhang et al (2013) on different online user-content bipartite networks.

But can these tweets exit the boundaries of the already consolidated audience and reach a public who would usually not engage with fact-checking contents (RQ2)? Since we also found that almost all the networks are disassortative, we can conclude that the few most popular tweets are responsible for most of the engagement of a new audience, while the least popular ones are only receiving attention from the recurring audience. This is also something observed by Mironov et al. (2021) in preferential attachment networks. The only exception to this rule among the networks observed is *maldita*, likely due to their advertisement strategy. In this regard, Spain exhibited a countertrend during the sentiment analysis, as positive tweets were more likely to become hubs. The keyword analysis revealed that terms related to self-promotion were more common among the hub tweets of Spanish fact-checkers than among non-hub tweets. This included terms specific to Maldita's account, such as 'malditatwitcheria'. Both findings converge, indicating an outreach strategy associated with self-promotion. This alignment is consistent with marketing studies that suggest advertisements conveying a positive tone are more likely to be shared (e.g., Kulkarni et al., 2020). It also corresponds with the peculiar result regarding the positive assortativity coefficient of maldita's network, which indicates that its popular tweets draw retweets from the users already used to retweet the account. Self-promotion tweets are more likely to be shared by the recurrent audience of the account, which is already familiar with the promoted content, than by users who do not regularly consume the account's content. Other insights derived from our delving into the features of the most popular tweets. By investigating whether patterns emerged based on the expressed sentiment (RQ3) and the touched-upon topic (RQ4), we found that topic in the hub tweets was not generalizable, being connected to specific and country-based political figures, or to what was happening in the different countries in that time frame. On the other hand,

it emerged that in most countries people tended to interact more with tweets expressing negative sentiment, a result in line with previous studies (e.g., Stieglitz & Dang-Xuan, 2013; Bhattacharya et al., 2014; Jiménez-Zafra et al., 2021). Our findings align with a broader audience motivation of sharing content perceived as urgent and impactful to their peers. Events and developments specific to the local context are more likely to directly impact users' lives and surroundings, fostering a personal connection that prompts them to share such content as it resonates with their daily experiences. In this regard, users may perceive tweets related to local happenings as more informative and useful.

Moreover, events and stories tied to the local context can elicit stronger emotional responses from individuals with a direct connection to the subject matter. As demonstrated in previous studies (see e.g., Berger, & Milkman, 2012), content capable of inducing a sense of activation (arousal) is more likely to attract engagement, specifically through emotional responses such as e.g., anger, fear, anxiety, amusement, and anticipation (e.g., for an event, as seen in the case of the Spanish account advertisements). Nonetheless, in line with our overall interpretation, some neutral contents - as in the case of the missing people theme found for Ireland - can become particularly engaging due to their perceived relevance for the local community.

6. Limits and future work

Our contribution has focused on the dissemination of fact-checkers around Europe. We gave a comprehensive overview of the properties of the content produced by 19 European fact-checkers and the corresponding patterns of engagement from the audience using a multidisciplinary approach. We acknowledge that there is an opportunity for a more in-depth analysis of each particular question.

To begin, our analysis focused on a specific social media platform, and we recognize the inherent limitations associated with extrapolating analyses and predictions from interactions on an online platform characterized by a biased population distribution (see Miranda Filho et al., 2015) and pseudonymous identity (see Peddinti et al., 2014). Future studies may explore various venues, both online and offline, where fact-checking content can be disseminated.

Second, in this contribution, we classified the tweets according to their polarity (positive, neutral, negative). A more granular classification based on emotions (i.e., anger, joy, fear, etc.) might deepen the understanding of some patterns we found. Nonetheless, some challenges might emerge, such as testing emotion classification algorithms consistently across multiple languages and collecting larger corpora of fact-checking content, which would enable a more detailed statistical analysis.

Third, in this paper, we restrict ourselves to the study of the retweet networks. Retweets consist only in sharing the content of the original author without adding to it any personal comment. Metaxas et al (2015) have shown that they can be generally considered an expression of support to the original author and agreement with the message, but it is impossible to gauge the actual stance of the retweeter on a case-by-case basis. Extending the study to replies and quote tweets we could use NLP

algorithms to analyze the explicitly expressed stance of the user interacting with the fact-checking content. This would allow a more detailed understanding of the patterns of different kinds of engagement, coming either from the users supporting the work of fact-checkers or those criticizing it.

On a similar note, our study delved into the contents published by fact-checking accounts. To further explore the hypothesis of the generalization of the "emotion-driven" pattern for audience engagement also in this context, it would be interesting to consider the sentiment expressed in the replies to these contents. We are planning to explore this in an upcoming study, where we will use users' replies as the basis for corpus creation.

References

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

Beauvais, C. (2022). Fake news: Why do we believe it?. Joint bone spine, 89(4), 105371.

Berger, J., & Milkman, K. L. (2013). Emotion and virality: what makes online content go viral?. NIM Marketing Intelligence Review, 5(1), 18-23.

Bhamidi, S., Dhara, S., van der Hofstad, R., & Sen, S. (2020). Universality for critical heavy-tailed network models: Metric structure of maximal components.

Burel, G., Farrell, T., Mensio, M., Khare, P., & Alani, H. (2020). Co-spread of misinformation and fact-checking content during the Covid-19 pandemic. In Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12 (pp. 28-42). Springer International Publishing.

Caldarelli, G., De Nicola, R., Petrocchi, M., Pratelli, M., & Saracco, F. (2021). Flow of online misinformation during the peak of the COVID-19 pandemic in Italy. EPJ data science, 10(1), 34.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2023). Spanish pre-trained bert model and evaluation data. arXiv preprint arXiv:2308.02976.

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. Procedia Computer Science, 161, 707-714.

Gaumont, N., Panahi, M., & Chavalarias, D. (2018). Reconstruction of the socio-semantic dynamics of political activist Twitter networks—Method and application to the 2017 French presidential election. PloS one, 13(9), e0201879.

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and research: A systematic literature review through text mining. IEEE access, 8, 67698-67717.

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that’s fit to fabricate: AI-generated text as a tool of media misinformation. Journal of experimental political science, 9(1), 104-117.

Kulkarni, K. K., Kalro, A. D., Sharma, D., & Sharma, P. (2020). A typology of viral ad sharers using sentiment analysis. Journal of Retailing and Consumer Services, 53, 101739.

Liu, X., Shin, H., & Burns, A. C. (2021). Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing. Journal of Business research, 125, 815-826.

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

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.

Metaxas, P., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., & Finn, S. (2015). What do retweets indicate? Results from user survey and meta-review of research. In Proceedings of the international AAAI conference on web and social media (Vol. 9, No. 1, pp. 658-661).

Miranda Filho, R., Almeida, J. M., & Pappa, G. L. (2015). Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In 2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM) (pp. 1254-1261). IEEE.

Mironov, S., Sidorov, S., & Malinskii, I. (2021). Degree-degree correlation in networks with preferential attachment based growth. In Complex Networks XII: Proceedings of the 12th Conference on Complex Networks CompleNet 2021 (pp. 51-58). Springer International Publishing.

Nanath, K., & Joy, G. (2023). Leveraging Twitter data to analyze the virality of Covid-19 tweets: a text mining approach. Behaviour & Information Technology, 42(2), 196-214.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. Physical review E, 64(2), 025102.

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory, 79.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. Journal of Information Science, 28(6), 441-453. https://doi.org/10.1177/016555150202800601

Peddinti, S. T., Ross, K. W., & Cappos, J. (2014). " On the internet, nobody knows you're a dog" a twitter case study of anonymity in social networks. In Proceedings of the second ACM conference on Online social networks (pp. 83-94).

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis.

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. Journal of applied research in memory and cognition, 9(3), 286-299.

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.

Wilhelm, E., Ballalai, I., Belanger, M. E., Benjamin, P., Bertrand-Ferrandis, C., Bezbaruah, S., Briand, S., Brooks, I., Bruns, R., Bucci, L., Calleja, N., Chiou, H., Devaria, A., Dini, L., D'Souza, H., Dunn, A., Eichstaedt, J., Evers, S., Gobat, N., ... & Purnat, T. D. (2023). Measuring the burden of infodemics: Summary of the methods and results of the Fifth WHO Infodemic Management Conference. JMIR infodemiology, 3(1), e44207.

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

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., ... & Huang, X. (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. arXiv preprint arXiv:2303.10420.

Zhang, C. X., Zhang, Z. K., & Liu, C. (2013). An evolving model of online bipartite networks. Physica A: Statistical Mechanics and its Applications, 392(23), 6100-6106.

Zhu, Z., Su, J., & Kong, L. (2015). Measuring influence in online social network based on the user-content bipartite graph. Computers in Human Behavior, 52, 184-189.