5th Multidisciplinary International Symposium on Disinformation in Online Open Media - MISDOOM 2023

November 21-22, 2023
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

EXTENDED ABSTRACTS

Editors

Davide Ceolin, Centrum Wiskunde & Informatica
Tommaso Caselli, University of Groningen
Marina Tulin, University of Amsterdam

Debunking and Exposing Misinformation among Fringe Communities: Testing Source Exposure and Debunking Anti-Ukrainian Misinformation among German Fringe Communities (Extended Abstract)

Christiern Santos Rasmussen$^{1[0000-0002-6764-0323]}$, Amir Ebrahimi Fard$^{2[0000-0003-4903-8918]}$, and Marijn ten Thij$^{2[0000-0002-7186-7344]}$

[1] Department of Political and Social Sciences, European University Institute, San Domenico di Fiesole (FI), Italy
johannes.santos@eui.eu
[2] Department of Advanced Computing Sciences, Maastricht University, Maastricht, the Netherlands

Abstract. In this study we conduct the first ever online field experiment testing traditional and novel counter-misinformation strategies among fringe communities. While traditional strategies have been found to effectively counter misinformation, these have yet to be tested among fringe communities that regularly consume misinformation online. Furthermore they do not address the infrastructure of misinformation sources supporting this consumption. We therefore propose to test if both traditional debunking and the novel counter-misinformation strategy, source exposure, can lower consumption of misinformation media among fringe communities. Based on a snowball sampling of German fringe communities on Facebook, we identified public Facebook groups who regularly consumed the two most popular misinformation sources in Germany. We then conducted an online field experiment to test the effect of debunking and source exposure on consumption levels. Results support a more active counter-misinformation approach to reduce consumption of misinformation sources.

Keywords: Misinformation · Fringe communities · Intervention · De bunking · Inoculation

Extended Abstract

In the current digital age, misinformation is a societal threat, as it erodes trust in democracy, increases polarization, and taps into extremist movements' ideology. As a result, ways to effectively stop misinformation are actively being investigated [3, 6, 5]. Short of censorship, debunking has become the gold-standard counter-strategy to misinformation [6], which is based on experiments that focus on a general audience, rather than members of fringe communities. Fringe communities are harder to reach with debunking approaches for as they 1) have higher consumption levels [1] of misinformation , 2) have been shown to be more susceptible to misinformation [2, 8], and 3) tend to avoid fact-checkers and traditional media [7]. In addition, misinformation spreaders have built an infrastructure of media, channels and blogs to facilitate their efforts [9] in which they can opportunistically jump between narratives and topics to flood a platform with relatable content [4]. Combining the fact that this misinformation infrastructure exists and these communities tend to fall outside of the scope of the existing misinformation treatment strategies, there is a need to investigate new ways to reach and inform this group. Exposing an outlet as a creator of misinformation within one of these fringe communities directly addresses both the aspects of this problem (i.e., combating the infrastructure and reaching the target audience). Consequently, such a strategy promises to be more effective than the existing techniques used by practitioners and researchers at this moment. Therefore, our study compares the use of source exposure directly within fringe communities as a technique to thwart consumption of misinformation, measured as number of URL mentions, as an alternative to existing approaches. Using a snowball sampling method, we identified public Facebook groups that regularly consume German misinformation sources (N = 35) and posted either debunks of the anti-Ukrainian misinformation claims or exposed the two most prominent misinformation sources' bad track record for spreading misinformation. In collaboration with the fact-checking organization VoxCheck, we conduct two types of interventions; 1) debunking anti-Ukrainian misinformation, or 2) exposing the source of the message as a spreader of misinformation. Serving as gatekeepers of posts in the groups, ten group administrators blocked our interventions. This allowed us to include another intervention, 3) targeting gatekeepers of fringe communities. Based on our experiment, we find that treated groups (i.e., debunking, source exposure, or gatekeeper targeting) do have a statistically significant lower consumption two weeks after treatment, compared to the control group (i.e., no intervention). Looking at the long-term effect of the interventions, we find that source exposure has a statistically significant long-term effect of lowering misinformation consumption, whereas we do not observe this outcome for debunking, in light with previous work [6]. More surprisingly, however, is the fact that we saw a statistically significantly lower consumption of misinformation sources for the fringe communities for which the group admins rejected our treatment. These results suggest that proactive counter-misinformation have an effect.

References

1. Bor, A., Petersen, M.B.: The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. American Political Science Review 116(1), 1–18 (2022). https://doi.org/10.1017/S0003055421000885
2. Bruder, M., Haffke, P., Neave, N., Nouripanah, N., Imhoff, R.: Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire. Frontiers in Psychology 4 (2013). https://doi.org/10.3389/fpsyg.2013.00225
3. Chan, M.P.S., Jones, C.R., Jamieson, K.H., Albarrac´ın, D.: Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. Psychological Science 28(11), 1531–1546 (2017). https://doi.org/10.1177/0956797617714579
4. Kriel, C., Pavliuc, A.: Reverse engineering Russian internet research agency tactics through network analysis. Defence Strategic Communications 6, 199–227 (2019)
5. van der Linden, S., Leiserowitz, A., Rosenthal, S., Maibach, E.: Inoculating the pub lic against misinformation about climate change. Global Challenges 1(2), 1600008 (2017). https://doi.org/10.1002/gch2.201600008
6. McCabe, D.P., Smith, A.D.: The effect of warnings on false memories in young and older adults. Memory & Cognition 30(7), 1065–1077 (2002). https://doi.org/10.3758/BF03194324
7. Nouri, L., Lorenzo-Dus, N., Watkin, A.L.: Impacts of radical right groups' movements across social media platforms – a case study of changes to Britain first's visual strategy in its removal from facebook to gab. Studies in Conflict & Terrorism 0(0), 1–27 (2021). https://doi.org/10.1080/1057610X.2020.1866737
8. Petersen, M.B., Osmundsen, M., Arceneaux, K.: The "Need for Chaos" and Motiva tions to Share Hostile Political Rumors. In: 114th Annual Meeting of the American Political Science Association (2018). https://doi.org/10.31234/osf.io/6m4ts
9. Starbird, K.: Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. Proceedings of the International AAAI Conference on Web and Social Media 11(1), 230–239 (May 2017). https://doi.org/10.1609/icwsm.v11i1.14878

Debunking Effectiveness of Corporate Social Responsibility (CSR) Washing (Extended Abstract)

Britta C. Brugman[1], Dian van Huijstee[2], Ellen Droog[2]

[1] University of Amsterdam, Amsterdam School of Communication Research. Corresponding author: Britta Brugman, Nieuwe Achtergracht 166, 1018WV Amsterdam, the Netherlands, b.c.brugman@uva.nl
[2] Vrije Universiteit Amsterdam, Department of Communication Science, the Netherlands

Much of disinformation research has focused on the domains of political communication, journalism, and health and science communication (Walter & Murphy, 2018). While these studies are extremely valuable in advancing scholarly understanding of, for instance, the effectiveness of corrective strategies such as debunks, it is still unclear to what extent findings generalize to other relevant communication domains, such as corporate communication. The current study therefore aims to integrate the *debunking* literature with a current topic that has still received limited attention in disinformation research: *CSR washing*.

In order to maintain and build public support for their activities, organizations are increasingly required by the general public to show that they conduct their activities in a corporate socially responsible (CSR) manner (e.g., Velte, 2022), for instance through advertisements (Perks et al., 2013). However, sometimes organizations fabricate or exaggerate the positive impact of their CSR initiatives. This has been referred to as CSR washing (Bernardino, 2022). Misleading claims involve those about organizations' support for environmental actions (e.g., stopping climate change; *greenwashing*), support for social and economic change (e.g., promoting legal justice; reducing poverty; *bluewashing*), and support for the emancipation and empowerment of women (e.g., achieving gender equality; *purplewashing*).

When undetected, CSR washing can positively enhance the organization's reputation and subsequently improve brand attitudes and purchase intentions (Chu et al., 2019). To test debunking effectiveness for this type of disinformation, we designed an experiment with a 2 (debunking: present vs. not) x 3 (washing type: green vs. blue vs. purple) between subjects design. As the dependent variables, we measured participants' brand attitudes and product purchase intentions before and after exposure to the stimulus materials. As the stimuli, we used Instagram ads of a fictional clothing company that contained exaggerated claims about the organization's ethical behavior in relation to sustainable production (greenwashing), child free labor (bluewashing) and gender equality in the workplace (purplewashing). The debunk was designed to look like a third-party fact-check you would see on Instagram. A total of 657 British participants participated in the study via Prolific in January of 2023. Our results revealed that debunking represents a promising strategy against CSR washing across different types. Participants who were only exposed to the Instagram ads that contained CSR washing indicated slightly positive brand attitudes and somewhat present purchase intentions. Participants who were also exposed to subsequent debunks of the ads, however, indicated moderately negative brand attitudes and purchase intentions. The debunks were thus found to significantly reverse persuasive effects of all three CSR washing types.

Our findings contradict previous debunking research (e.g., Walter & Tukachinsky, 2020) since earlier studies often showed continued influence of debunked disinformation. Possible explanations for the discrepancy in findings include (a) that identity politics and political ideology may play a less important role in the processing debunks of corporate washing than of other forms of disinformation, and (b) that disinformation generally focuses on negativity while corporate washing by default consists of positive statements. We therefore hope our study inspires more attention in future disinformation research to CSR washing.

References

Bernardino, P. (2021). Responsible CSR communications: Avoid "washing" your corporate social responsibility (CSR) reports and messages. Journal of Leadership, Accountability & Ethics, 18(1), 102-113.

Chu, S. C., & Chen, H. T. (2019). Impact of consumers' corporate social responsibility‐related activities in social media on brand attitude, electronic word‐of‐mouth intention, and purchase intention: A study of Chinese consumer behavior. Journal of Consumer Behaviour, 18(6), 453-462. https://doi.org/10.1002/cb.1784

Perks, K. J., Farache, F., Shukla, P., & Berry, A. (2013). Communicating responsibility practicing irresponsibility in CSR advertisements. Journal of Business Research, 66(10), 1881-1888. https://doi.org/10.1016/j.jbusres.2013.02.009

Velte, P. (2022). Meta-analyses on corporate social responsibility (CSR): a literature review. Management Review Quarterly, 72(3), 627-675. https://doi.org/10.1007/s11301-021- 00211-2

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. Communication Monographs, 85(3), 423-441. https://doi.org/10.1080/03637751.2018.1467564

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it?. Communication Research, 47(2), 155-177. https://doi.org/10.1177/0093650219854600

Data Stream Clustering on Systematically Collected Social Media Benchmarks Incorporating Semantic Similarities (Extended Abstract)

Janina Pohl*, Dennis Assenmacher**, Christian Grimme*, and Heike Trautmann*
[janina.pohl, christian.grimme, heike.trautmann]@uni-muenster.de, dennis.assenamcher@gesis.org

* Data Science: Statistics and Optimization, University of Münster ** GESIS: Leibniz Institute for the Social Science

1 Abstract

Stream clustering algorithms are vital for processing high-velocity data streams, particularly from social media platforms, facing challenges in handling dynamic data and detecting concept drifts. On social media, stream clustering is helpful for tasks like event identification and campaign detection. However, existing social media benchmarks lack essential features like emojis, URLs, and hashtags, and exhibit low entropy, raising doubts about their representativeness in real-world scenarios. To address these limitations, the project introduces three new benchmark datasets from Twitter, Reddit, and Telegram, along with an artificial dataset generated using GPT-Neo. The datasets are labeled with broader topics and incorporate pseudonymization techniques to protect user privacy. The study also proposes an enhancement to the textClust algorithm by incorporating soft cosine similarity based on word embeddings to improve its ability to handle diverse clusters. The new version of textClust will be tested in a grand benchmarking study alongside other stream clustering algorithms.

Keywords: computer science · stream clustering · campaign detection

2 Extended Abstract

Stream clustering algorithms are crucial in processing high-velocity, potentially unbounded, continuously generated data streams, particularly prevalent on social media. These algorithms encounter distinctive challenges inherent to the dynamic nature of multi-modal data, including the timely detection of concept drifts and often only being able to inspect a record once (Silva et al., 2013). Use cases of stream clustering on social media include not only (crisis) event or topic identification but also campaign detection. Here, identifying outliers or anomalous patterns in identified clusters can offer insights into coordinated or inauthentic behavior (Anwar et al., 2023; Assenmacher et al., 2020). Consequently, assessing and comparing different stream clustering algorithms on benchmark datasets resembling real-world scenarios become imperative for informed decision-making regarding selecting a suitable algorithm for a specific use case.

However, most of the existing social media benchmarks such as News-T or Tweets T (Yin et al., 2018) underwent extensive preprocessing that excludes important features like emojis, URLs, and hashtags, which are commonly present on various social media platforms. Moreover, our research demonstrates that these benchmarks exhibit low entropy, indicating a prevalence of repetitive and similar words. This, in turn, leads to disruptions in the semantic structure of the texts. Thus, the dataset's ability to accurately represent real-world use cases is questionable.

To address this limitation, our project introduces three novel benchmark datasets from Twitter, Reddit, and Telegram. Each textual entry within these datasets is labeled with its broader topic: the Sub-Reddit for Reddit data, the keywords used to filter the stream for Twitter, and the designated group or channel topic for Telegram. To ensure compliance with platform terms of service and privacy regulations, we employ pseudonymization techniques by substituting user-related information with synthetic data generated using the Python package Faker (Faraglia et al., 2023). Additionally, we decouple the textual content from its associated ID and author information. These genuine datasets, collected from existing social media platforms, represent a contemporary and challenging novel benchmark for clustering algorithms.

In addition to our new real-world datasets, we propose an artificial dataset generated with GPT-Neo (Black et al., 2021). The model was trained on social media data to generate genuine tweets (Pohl et al., 2022), and thus, represents a valid new type of benchmark dataset for sensitive social media data. Further, especially since the advent of large language models, generating social media posts automatically is a viable option for (malign) campaign creators. While identifying these agents in the past was possible by applying consistent rules to analyze account features due to their predictable behavior (Howard & Kollany, 2016; Varol et al., 2017), the utilization of automated text generation, resulting in a heterogeneous stream of posts conveying identical messages, significantly hinders their detection (Pohl et al., 2022). Therefore, we propose an enhancement to a state-of-the-art stream clustering algorithm, namely textClust. This one-pass algorithm employs the cosine similarity based on TF-IDF vectors and exponential fading to update evolving clusters incrementally (Assenmacher & Trautmann, 2022). However, prior studies have demonstrated that relying solely on term frequencies for text and cluster similarity calculations proves inadequate in effectively detecting more diverse clusters, such as those generated by GPT (Pohl et al., 2022).

Our proposed improvement involves incorporating the concept of soft cosine similarity into the cluster formation process of textClust (Sidorov et al., 2014). This integration

leverages semantic similarities of words based on word embeddings, enabling the algorithm to handle diverse clusters more proficiently. Naturally, we will incorporate the new version of textClust into a new grand benchmarking study to test the algorithms, its predecessor, and other competitive stream clustering algorithms on the novel benchmarks.

References

Anwar, T., Nepal, S., Paris, C., Yang, J., Wu, J., & Sheng, Q. Z. (2023). Tracking the evolution of clusters in social media streams. *IEEE Transactions on Big Data*, *9* (2), 701–715. https://doi.org/10.1109/TBDATA.2022.3204207

Assenmacher, D., Clever, L., Pohl, J., Trautmann, H., & Grimme, C. (2020). A two-phase framework for detecting manipulation campaigns in social media. In G. Meiselwitz (Ed.), *Proceedings of the international conference on human-computer interaction (hcii 2020): Social computing and social media. design, ethics, user behavior, and social network analysis* (pp. 201–214). Springer International Publishing. https: //doi.org/10.1007/978-3-030-49570-1 14

Assenmacher, D., & Trautmann, H. (2022). Textual one-pass stream clustering with automated distance threshold adaption [Publication status: Published]. In T. et al. Tran (Ed.), *Intelligent information and database systems* (pp. 3–16). Springer International Publishing. https://doi.org/10.1007/978-3-031-21743-2 1

Black, S., Leo, G., Wang, P., Leahy, C., & Biderman, S. (2021). *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow* (tech. rep.). Zenodo. https://doi.org/10.5281/zenodo.5297715

Faraglia, D., many, other, & contributors. (2023, June 24). *Faker* (Version 18.11.1). https://github.com/joke2k/faker

Howard, P. N., & Kollany, B. (2016). Bots, #strongerin and #brexit: Computational Propaganda during the UK-EU Referendum. *Social Science Research Network*. https://doi.org/10.2139/ssrn.2798311

Pohl, J. S., Assenmacher, D., Seiler, M. V., Trautmann, H., & Grimme, C. (2022). Artificial social media campaign creation for benchmarking and challenging detection approaches. *Workshop Proceedings of the 16th International Conference on Web and Social Media (ICWSM)*, 1–10. https://doi.org/10.36190/2022.91

Sidorov, G., Gelbukh, A., Gomez Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, *18*. https://doi.org/10.13053/cys-18-3-2043

Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., Carvalho, A. C. P. L. F. d., & Gama, J. (2013). Data stream clustering: A survey. *ACM Comput. Surv.*, *46* (1). https://doi.org/10.1145/2522968.2522981

Varol, O., Ferrara, E., Davis, C., Menczer, F., & Flammini, A. (2017). Online human bot interactions: Detection, estimation, and characterization. *Proceedings of the International AAAI Conference on Web and Social Media*, *11* (1), 280–289.

Yin, J., Chao, D., Liu, Z., Zhang, W., Yu, X., & Wang, J. (2018). Model-based clustering of short text streams. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2634–2642.

Correcting misinformation before or after exposure: What works better to reduce continued influence? (Extended Abstract)

Dian van Huijstee, Ivar Vermeulen, Ellen Droog, & Peter Kerkhof

Vrije Universiteit Amsterdam

The prevalence of false information in online news has become a significant concern, largely due to our growing reliance on social networking sites (SNS) as primary news sources. Because misinformation can create general distrust in media, democracy, and institutions, it threatens deliberative democracy (Van Aelst et al., 2017), increasing the importance to counter it. Correcting misinformation, however, does not seem to be unequivocally effective: corrections help to make news consumers aware that information is incorrect, but do not always neutralize its persuasive influence, resulting in *continued influence* of misinformation (Lewandowsky et al., 2012).

While continued influence research typically focuses on post-consumption corrections (debunks), current practice on SNS involve issuing misinformation warnings beforehand. Forewarnings accompanying other types of persuasive messages (e.g., ads), have been shown to evoke persuasive resistance, increasing counterarguing, and resulting in reduced opinion change (e.g., Boerman et al., 2012). Thus, theory suggests that forewarnings could effectively counter misinformation effects, as was empirically supported by Ecker and colleagues (2010) in their study demonstrating the reduction of continued influence due to specific forewarnings.

Interestingly, recent studies alternatively measuring recall and veracity judgements show that post-exposure corrections are actually more effective than forewarnings (Dai et al., 2021; Brashier et al., 2021). In the current experiment (preregistered at AsPredicted #84663), we directly compare the continued influence effects associated with both strategies empirically.

We used a 3(forewarning vs. debunk. vs. no-correction) by 2(positive vs. negative[1] (mis)information) between-subject design on a Prolific sample of 657 British participants (75.8% female; *M*age=40.40, *SD*age=12.97; political orientation (left to right 1-11): *M*=5.28, *SD*=2.03). Pre- and post-exposure attitudes and voting intentions were measured (-10 to +10), and persuasive influence served as DV, representing attitude/intention change from t1 to t2 congruent to the news article's valence.

For both the forewarning and debunking messages, persuasive influence significantly exceeded 0 for attitude and voting intention (one-sample t-tests; forewarning: *Matt*=5.10,*SDatt*=4.84, *t*(220)=15.65, *p*<.001, *d*=1.05 and *Mvot*=5.29,*SDvot*=5.67, *t*(220)=13.85, *p*<.001, *d*=0.93; debunking: *Matt*=3.24,*SDatt*=5.35, *t*(216)=8.91, *p*<.001, *d*=0.61 and *Mvot*=3.55,*SDvot*=6.12, *t*(216)=8.55, *p*<.001, *d*=0.58), thus signaling continued influence for both correction strategies. ANOVAs showed a significant main effect for correction type (*F*(2,654)=18.88, *p*<.001, η^2=0.06) for the persuasive impact on attitude, where all strategies significantly differed (*p's*<.001) except for forewarning and no-correction (*p*=.074). The same was found for the persuasive impact on voting intention: *F*(2,654)=14.17, *p*<.001, η^2=0.04; all strategies again significantly differed (*p's*<.001) except for forewarning and no-correction (*p*=.076). Debunks were most effective in lowering persuasive impact, while forewarnings had a limited impact compared to no correction.

Our study suggests that correcting political misinformation after exposure is preferable, as it was more effective in reducing its persuasive impact compared to forewarnings. Forewarnings, although lowering

[1] Due to the word-limit, valence effects will be discussed during the conference but cannot be included in this abstract.

credibility, did not significantly reduce persuasive influence, similar to no correction. This research informs social media developers and managers using (fore)warnings to combat the influence of misinformation. Collaboration between researchers and practitioners is essential to prevent misinformation from becoming a more significant societal problem.

References

Boerman, S. C., van Reijmersdal, E. A., & Neijens, P. C. (2012). Sponsorship Disclosure: Effects of Duration on Persuasion Knowledge and Brand Responses. Journal of Communication, 62, 1047-1064.

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. Proceedings of the National Academy of Sciences, 118(5), e2020043118.

Dai, Y., Yu, W., & Shen, F. (2021). The effects of message order and debiasing information in misinformation correction. International Journal of Communication, 15, 1039–1059. Ecker, U.K.H.

Lewandowsky, S., & Tang ,D.T.W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory and Cognition, 38(8), 1087–1100.

Lewandowsky, S., Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychological Science in the Public Interest,13(3), 106–131.

Van Aelst, P., & Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? Annals of the International Communication Association, 41(1), 3–27.

Combatting the Disinformation Crisis: A Systematic Literature Review of the Characteristics and Effectiveness of Media Literacy Interventions (Extended Abstract)

Ellen Droog, Ivar Vermeulen & Dian van Huijstee

Vrije Universiteit Amsterdam

Due to the sheer amount of disinformation in our media environment, there is an urgent need for the development of effective media literacy interventions that can broadly protect people from the negative effects of disinformation by enhancing their ability to critically evaluate the information they encounter. However, research suggests that such media literacy interventions are not always effective in increasing these abilities (e.g., Ecker et al., 2022). This calls for a better understanding of the characteristics and effectiveness of media literacy interventions developed to counter the effects of disinformation. We aimed to improve this understanding by conducting a systematic literature review of experimental research that tested the effectiveness of different media literacy interventions.

Following the PRISMA checklist, the systematic literature review consisted of five steps: (1) database search, (2) duplicate removal, (3) abstract screening through ASReview (van de Schoot et al., 2021), (4) full content screening, and (5) article coding. A total of 67 articles comprising 80 studies were included in the final review.

The media literacy interventions varied in nature and effectiveness. Using existing typologies (Ecker et al., 2022; van der Linden, 2022), we categorized interventions and found that inoculation interventions, which involved warning users about disinformation and providing examples and refutations, were the most studied type. Of those interventions, passive inoculation, where users passively consumed the intervention, was the most prevalent (46%). Active inoculation, where users actively engaged with the intervention, was used less frequently (22%). General media literacy interventions, such as using infographics to identify disinformation, were relatively more common (30%). Finally, some interventions were logic- (11%) or source-based (1%), prompting users to identify disinformation based on faulty arguments or on a non-credible source.

The effectiveness of media literacy interventions depended on specific outcome variables. Most studies focused on assessing the veracity of the disinformation (64%). Psychological outcomes such as beliefs (33%), attitudes (30%), intentions (16%), and behaviors (9%) were also examined, though to a much lesser extent. Surprisingly, only 15% of studies explored an actual improvement in (perceived) media literacy skills. The majority of the interventions were effective in improving users' ability to identify disinformation. However, the effects on psychological outcomes were more mixed, with less than half of the interventions effectively countering disinformation's impact on beliefs, attitudes, intentions, and behaviors. Long-term effects varied, some studies showed persistent effects for several weeks post-intervention, while others did not. Of the few studies that assessed actual improvement in media literacy skills, most reported positive effects. Overall, no specific type of media literacy intervention stood out as particularly successful.

Based on the findings, we propose four recommendations for practitioners. Firstly, interventions should activate pre-existing media literacy knowledge rather than solely educate users. Secondly, they should prioritize improving users' information selection skills over information processing skills. Thirdly, technique-based interventions should be emphasized, because they were found to be more effective than issue-based ones. Finally, interventions should prevent a skeptical attitude towards news in general by equipping users with skills to differentiate between true and false information.

References

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., … & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology, 1*(1), 13-29. https://doi.org/10.1038/s44159-021-00006-y

Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., ... & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence, 3*(2), 125-133. https://doi.org/10.1038/s42256-020-00287-7

Van Der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine, 28*(3), 460-467. https://doi.org/10.1038/s41591- 022-01713-6

The Gamification of Disinformation and Cross-Platform Negotiation

Cassian Osborne-Carey, Robert Topinka

Social media platforms have become more active in banning groups spreading abuse and disinformation. Partly in response to criticism from researchers mapping the spread of such abuse - platforms have restricted access to APIs, making it difficult to track how users respond to and negotiate moderation (Bruns, 2019). Researchers must adapt methodologically to evaluate the consequences of deplatforming. They must be attuned to cross-platform practice and the challenges arising when following actors across the disinformation landscape.

We study a community of supporters of Donald Trump as they migrate from Reddit to multiple sites on the '.win' platform. Members of r/The_Donald were accused of manipulating platform mechanics to increase their content's visibility, harassing platform opponents and promoting violence. After being banned they found a new home on the .win platform, a Reddit clone, first at 'TheDonald.win' and finally 'Patriots.win'. Drawing on a dataset incorporating a thousand posts and 13,000 comments collected from r/The_Donald before Reddit reduced access to its API, and working with two open-access datasets featuring millions of posts and comments from TheDonald.win and Patriots.win, we analyse how the community respond to and remediate political news and events as they negotiate platform moderation.

As platforms limit API access, researchers need flexible, 'quali-quantitative' (Moats and Borra, 2018) methods that combine computational analyses of large datasets with manual collection and close interpretive analysis. Taking this approach, we examine the 'ensemble of practices' (Pink et al, 2016) across platforms, incorporating 'distant' quantitative analysis of post frequency, popularity and commenting patterns with 'close' qualitative analysis of discursive themes present in user content. We code the ways community members negotiate affordances as they remediate news and participate in the 'gamification of disinformation', building on a definition of gamification as 'the use of game design elements in non-game contexts'(Deterding et al 2011)

The link between gaming and politics has attracted recent attention with research on extremism and radicalisation pointing to gaming communities as targets for infiltration by political activists (Condis 2021; Lamphere-Englund & White 2023; Wells et al 2023). Within gaming communities these actors amplify hate speech, spread extremist material, and collectively re-design games, including the production and dissemination of discriminatory mods. However, we argue one must also attend to the gaming *of* politics, as social media users discursively remediate politics as a game of salvation featuring heroic characters, evil enemies, non-player characters, and apocalyptic landscapes, wherein community members weaponise 'truth' through collective reinterpretation and dissemination of news.

We show how pro-Trump users respond to breaking news by weaving a dramatic narrative of victimhood and conspiracy, enacting roles as protagonists fighting 'Info War' for personal, national and global salvation. We note shifts in practice following deplatforming that impact community identity, conflict and the symbolic construction of enemies, and evaluate the wider consequences for the spread of disinformation when a community loses access to a mainstream platform audience but gains firmer control over the means of ideological production.

References

Bruns, A. (2019) After the ‘APIcalypse’: social media platforms and their fight against critical scholarly research, Information, Communication & Society, 22:11, 1544-1566, DOI: 10.1080/1369118X.2019.1637447

Condis, M. (2021). Playing at racism: White supremacist recruitment in online video game culture. In White Supremacy and the American Media (pp. 246-274). Routledge.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining "gamification". In Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments (pp. 9-15).

Lamphere-Englund, G., White, J., & GNET (Global Network on Extremism & Technology) (2023). The Online Gaming Ecosystem: Assessing Digital Socialisation, Extremism Risks and Harms Mitigation Efforts. Global Network on Extremism and Technology. https://doi.org/10.18742/pub01-133

Moats D & Borra E (2018) Quali-quantitative methods beyond networks: Studying information diffusion on Twitter with the Modulation Sequencer. Big Data & Society 5(1): 1–17. https://doi.org/10.1177/2053951718772137.

Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T., & Tacchi, J. (2015). Digital ethnography: Principles and practice. Sage. Chicago

Wells, G., Romhanyi, A., Reitman, J. G., Gardner, R., Squire, K., & Steinkuehler, C. (2023). Right-Wing Extremism in Mainstream Games: A Review of the Literature. Games and Culture, 15554120231167214.

Analysing Political Bias of News Outlets by Clustering Social Media Posts[2]

(Extended Abstract)

Wilcke WX[1[0000−0003−2415−8438]] and Kuhn T[1[0000−0002−1267−0234]]

Dept. of Computer Science
Vrije Universiteit Amsterdam
The Netherlands
{w.x.wilcke, t.kuhn} @ vu.nl

Abstract. Social media platforms are rapidly becoming the primary source of information for many people around the world. Much of this information is created by online news outlets, who have successfully made the transition to social media and who now have the ability to potentially influence the topical stance and political views of millions of followers, simply by how they frame the news. Raising awareness about these effects can help these people to put things into perspective, reducing their impact. In this ongoing work, we introduce an approach for investigating bias in social media networks, by modelling and comparing online news outlets based on the content and context of their posts. This is accomplished by treating the post history as a graph with metadata and multimodal features, and by learning meaningful embeddings for each news outlet, which we can compare in a two-dimensional space. We evaluate our approach quantitatively, against different models, and qualitatively, with the help of domain experts.

Social media platforms are rapidly becoming the primary source of information for many people around the world [6]. Much of this information is created by regular users and prominent figures, yet many of the once traditional news outlets have now also successfully made the transition to online platforms such as Facebook, Twitter, and YouTube. Catering to an audience worldwide, these news outlets have the ability to influence the topical stance and political views of millions of people, potentially affecting elections and increasing polarisation. These effects might be the intended aim in a handful of cases, but it is far more often a side effect of how online news outlets select and frame the news [5]. For example, in the news coverage of the 2014 annexation of Crimea by Russia, both Russian and Belarusian news outlets spoke of the liberation of the Crimean people, whereas most other news outlets in Europe and America spoke of an aggressive invasion and subsequent occupation. Neither of these views is an universal truth, yet both are valid interpretations of the same event.

Raising awareness about bias is an important first step towards reducing its effects: understanding that every news item that you consume is coloured in some way, on purpose or otherwise, can help a person put things into perspective. One way to achieve this is by comparing the news items about the same event from different outlets [1]. Likewise, we can compare the topical and political stance of different news outlets by comparing all of their news items. In this case, we can even forgo matching items by events since the aggregation of news items is likely to converge to the overall mean stance of the news outlet that published them.

In this ongoing work, we introduce an approach for investigating bias in social media networks, by modelling and comparing online news outlets based on the content and context of their posts. By

[2] This research is party funded by the CaptureBias project.

treating the post history not as a set of documents but as an interconnected graph with accompanying metadata and multimodal features, our approach is capable of learning representations that embed the complex dynamics and information necessary for a faithful comparison. Under the hood, our approach makes use of a multimodal graph autoencoder [7] to learn embedding vectors for each published post, which are aggregated per news outlet and subsequently projected in a two-dimension space for visual comparison. We evaluate our approach quantitatively, by comparing the learned embedding landscape to that learned with A) a regular graph autoencoder [2], B) word2vec [4], and C) doc2vec [3], and qualitatively, by discussing the learned landscapes with experts on news and bias during a full day workshop.

References

[1] Abeer ALDayel and Walid Magdy. “Stance detection on social media: State of the art and trends”. In: Information Processing & Management 58.4 (2021), p. 102597.

[2] Thomas N Kipf and Max Welling. “Variational graph auto-encoders”. In: arXiv preprint arXiv:1611.07308 (2016).

[3] Quoc Le and Tomas Mikolov. “Distributed representations of sentences and documents”. In: International conference on machine learning. PMLR. 2014, pp. 1188–1196.

[4] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: arXiv preprint arXiv:1301.3781 (2013).

[5] Gordon Pennycook and David G Rand. “Fighting misinformation on social media using crowdsourced judgments of news source quality”. In: Proceedings of the National Academy of Sciences 116.7 (2019), pp. 2521–2526.

[6] Elisa Shearer and Jeffrey Gottfried. “News use across social media platforms 2017”. In: (2017).

[7] WX Wilcke et al. “End-to-End Learning on Multimodal Knowledge Graphs”. In: Under Submission (2023).

When Does Conspiracy Exposure Affect Behavioural Intentions? The Moderating Role of the Need to Evaluate (Extended Abstract)

Valentin Mang, Kai Epstude, Bob M. Fennis

University of Groningen

Exposure to conspiracy theories has been shown to detrimentally affect behavioural intentions, for instance, reducing intentions to get vaccinated [1], to engage in pro-environmental behaviour, and to engage in politics [2]. While research on conspiracy beliefs has recently gained much attention, experimental research on the cognitive processes underlying belief in conspiracy theories and their outcomes is scarce [3], [4]. To address this gap, we experimentally investigated the effects of exposure to conspiracy narratives on behavioural intentions as well as the mediating role of specific conspiracy beliefs across domains (climate change, COVID-19). Additionally, we examined the potential moderating role of individual differences in the need to evaluate (NE). The NE is the tendency to engage in evaluative responding when exposed to objects or issues [5], and has to date not been examined in the context of conspiracy beliefs. We expected that for individuals with a high NE, exposure to conspiracy theories should affect evaluative responses in the form of behavioural intentions more strongly compared to those with a low NE. More specifically, in the hypothesised process, namely that conspiracy exposure affects conspiracy beliefs, which then affect behavioural intentions, we predicted that the effect of conspiracy beliefs on intentions would be stronger for individuals with a high (vs. low) NE.

We tested our predictions in two experimental studies. In study 1 (N = 220), exposure to a conspiracy narrative about climate change (vs. a non-conspiracy narrative) increased belief in climate change conspiracy theories. Belief in climate change conspiracy theories then reduced intentions to engage in pro-environmental behaviour, but more strongly for those with a high (vs. low) NE. In study 2 (N = 358), exposure to a vaccination conspiracy narrative increased belief in general vaccination conspiracy theories, in turn reducing vaccination-related intentions regarding COVID-19. The indirect effect of conspiracy exposure on behavioural intentions via conspiracy beliefs was again stronger for those with a high (vs. low) NE. In other words, individuals with a high NE seem to be more at risk of the negative effects of conspiracy beliefs than those with a low NE.

Overall, these studies shed light on the process by which conspiracy exposure affects conspiracy beliefs and behavioural intentions causally across contexts and suggest that the detrimental effects of conspiracy beliefs might be more pronounced for individuals with a high (vs. low) NE. In addition to broadening understanding of how conspiracy theories affect individuals' behavioural intentions, the present research could help inform interventions aimed at reducing the negative behavioural consequences of conspiracy beliefs. Previous research has shown that high-NE individuals tend to exhibit more online engagement (e.g., creating and seeking content), which can help identifying these individuals and targeting them with online content [6]. Therefore, targeting high-engagement individuals online with interventions aimed at reducing the negative effects of conspiracy theories [7] could help in reaching a larger share of individuals who are especially at risk of the negative effects of conspiracy beliefs (i.e., high-NE individuals), thereby potentially improving such intervention efforts.

Bibliography

[1] D. Jolley and K. M. Douglas, ‘The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions’, PLOS ONE, vol. 9, no. 2, p. e89177, Feb. 2014, doi: 10.1371/journal.pone.0089177.

[2] D. Jolley and K. M. Douglas, ‘The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one’s carbon footprint’, Br. J. Psychol., vol. 105, no. 1, pp. 35–56, 2014, doi: 10.1111/bjop.12018.

[3] A. M. Enders, J. Uscinski, C. Klofstad, and J. Stoler, ‘On the relationship between conspiracy theory beliefs, misinformation, and vaccine hesitancy’, PLOS ONE, vol. 17, no. 10, p. e0276082, Oct. 2022, doi: 10.1371/journal.pone.0276082.

[4] K. Sassenberg, P. Bertin, K. M. Douglas, and M. J. Hornsey, ‘Engaging with conspiracy theories: Causes and consequences’, J. Exp. Soc. Psychol., vol. 105, p. 104425, Mar. 2023, doi: 10.1016/j.jesp.2022.104425. [5] W. B. G. Jarvis and R. E. Petty, ‘The need to evaluate’, J. Pers. Soc. Psychol., vol. 70, pp. 172–194, 1996, doi: 10.1037/0022-3514.70.1.172.

[6] M. Xu, R. W. Reczek, and R. E. Petty, ‘Need to evaluate as a predictor of creating and seeking online word of mouth’, Mark. Lett., Apr. 2023, doi: 10.1007/s11002-023-09676-5.

[7] J. Roozenbeek, S. van der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky, ‘Psychological inoculation improves resilience against misinformation on social media’, Sci. Adv., vol. 8, no. 34, p. eabo6254, Aug. 2022, doi: 10.1126/sciadv.abo6254.

Controversy detection and automated characterization of polarized communities by compression distances - Extended Abstract

Luis Pérez-Miguel[1] and David Arroyo[1[0000−0001−8894−9779]]

Institute of Physics and Information Technologies "Leonardo Torres Quevedo" (ITEFI), Spanish
National Research Council (CSIC), Spain
luis.perezdavid.arroyo@csic.es

Abstract. A critical characteristic of current discussions in social networks is determined by the high level of polarization over certain controversial topics. This same polarization is a key component of an increasing number of advanced manipulation campaigns in social networks.

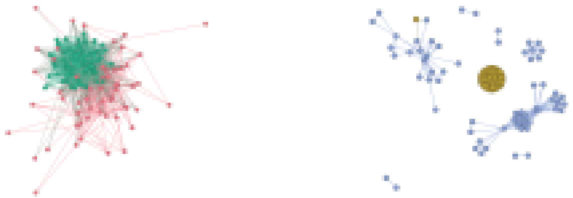
We present a methodology for the characterization and visualization of controversial discussions in online fora. Aiming to provide users with automatic means to identify the different sides of a debate. This can be later applied to improve our previous work on authorship attribution using natural language processing tools. Due to the disappearance of academic Twitter API, this work is presented with a single use case.

In this work we propose a variance of methodologies for the identification and visualization of controversial discussions in online social networks. We aimed to provide a framework for easily understanding the discourse around a given topic in Twitter. However, due to the elimination of academic access to X (Previously Twitter) API this was reduced to a single topic case study. We believe the same basis could be used to discourse of other Online Social Networks or discourse on media coverage. The methodology consists of the following steps.

Topic modelling Considering a collection of tweets within a given time frame. We search for further tweets including a hashtag from the set of those appearing in the corpus, and compute the similarity between the obtained results and the original set of tweets, utilizing the Normalized Compression Distance (NCD) [4]. A topic is then defined as the top-20 most similar hashtags. The accuracy of the method has been tested in a simple setup by analyzing Twitter threads about the new Spanish housing law and comparing the results with those obtained using the cosine similarity metric [1], where our method enlist more terms related to said law, instead of general politics.

NCD		Cosine - similarity	
LeyDeVivienda	LikeValencia	LeyDeVivienda	28M
NosotrasPodemos	NoaEstaLeydeVivienda	DebateMadridRTVE	SiSePuede
SouadSeQueda	leyantiocupas	AnaRosa	noticias
CuestionDePrioridades	aspe	colombia	ReformaLaboral
Botànic	Barbano	vivienda	empleo
VotoÚtil	eutanasia	SoloQuedaVox	LoQueVotasImporta
YoConPodemosSiempre	ElkarrekinPodemos	Corrupcion	turismo
cloacasdelperiodismo	UnidasSíPodemos	pp	VotaSeguro
VotaMásMadrid	TrabajoDigno	Almería	Alicante
claves	inversor	blog	bogota
LaLlaveParaMadrid	28Maig	derecho	leonesp

Connection graph We construct a connection graph representing the users participating in the conversation employing two different methods. First, based on the use of at least two hashtags in common in order to identify users focusing on a specific aspect of a given topic. In the second method, we establish con nections between users by applying the NCD to measure the similarity of their discourse during the given period of time. In our given case, the first approach divides the discourse completely; this might be due to a small size of the corpus .



(a) Graph by NCD similarity (b) Graph by common hashtag use

Group characterization After a separation of the graph in two clusters, we can describe each community according to the methodology proposed in [3, 2]. Filtering the central users: those with both a big amount of connections and a large number of tweets; these users are treated as generators, while the rest are considered as the training set for a SVM model. The model is trained over the NCD measure between the outer users and the generators. Following this approach, we could perform an automatic classification of new participants in the discourse.

With this work we extended our previous contributions on applying the NCD to authorship attribution and to the identification of polarized discussions in Twitter in a accurate way. However, we need to pivot to proving the applicability of this approach to different social media environments.

References

1. Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quan tifying controversy in social

media. CoRR abs/1507.05224 (2015), http://arxiv.org/abs/1507.05224

2. Mułoz, S.P., Oliva, C., Lago-Ferńandez, L.F., Arroyo, D.: Advancing the use of information compression distances in authorship attribution. In: Disinformation in Open Online Media: 4th Multidisciplinary International Symposium, MISDOOM 2022, Boise, ID, USA, October 11–12, 2022, Proceedings. pp. 114–122. Springer (2022)

3. Mułoz, S.P., Oliva, C., Lago-Fern´andez, L.F., Arroyo, D.: Advancing the use of in formation compression distances in authorship attribution. In: Spezzano, F., Ama ral, A., Ceolin, D., Fazio, L., Serra, E. (eds.) Disinformation in Open Online Media. pp. 114–122. Springer International Publishing, Cham (2022)

4. de la Torre-Abaitua, G., Lago-Fern´andez, L.F., Arroyo, D.: A compression-based method for detecting anomalies in textual data. Entropy 23(5), 618 (2021). https://doi.org/10.3390/e23050618, https://doi.org/10.3390/e23050618

Mood, Threat, and Gamified Psychological Inoculation Against Misinformation (Extended Abstract)

Dan Loughnan[1, 2] and Kai Epstude[2]
Radboud University[1], University of Groningen[2]
Email: dan.loughnan@ru.nl

Psychological Inoculation Theory [1] describes a two-step process of resistance to persuasion: (i) a warning to instil 'motivational threat', conceptualised as a perceived threat that motivates effortful attention; and (ii) the pre-emptive refutation of a persuasive attack. In applying the theory to address misinformation, researchers developed choice-based online games to inoculate against manipulative techniques of persuasion [2]. *Go Viral!* [3] is one such 'inoculation game', wherein players score 'likes' by successfully spreading misinformation on COVID-19. Go Viral! resembles 'casual video games' that clinical research has shown promote happiness [4], that being a mood state that drives information processing effects that could disrupt inoculation [5, 6, 7]. However, such effects may be extinguished by promoting sufficient motivation [8, 9]. It is therefore pertinent to consider mood and threat in the context of inoculation games and their effects. This is the first study to do so.

We ran a between-subjects, double-blinded, randomised controlled experiment on *Prolific* (UK residents; 18-68 years; N=368). Participants submitted to happy or sad audio-visual mood inductions validated for online use [10], then played Go Viral! or watched a control video. Susceptibility to misinformation was assessed by dichotomous ratings of (un)reliable news headlines pre/post-intervention (T1/T2), and the following day (T3). We also measured post-intervention motivational threat. Analyses employed two current metrics of susceptibility to misinformation: mean group ratings of unreliable news headlines only (metric 1), and signal detection analyses to determine discernment between reliable and unreliable headlines, and overall scepticism (metric 2).

- **Effects of Mood:** Manipulation checks showed mood inductions were successful. As predicted, accuracy in discernment between reliable and unreliable headlines was increased by sadness, and reduced by happiness. Both groups became more sceptical, especially the sad group (see Figure 1).

- **Motivational Threat:** There were no differences in motivational threat between treatment and control conditions. As level of perceived threat provides a manipulation check for inoculation, inoculation apparently did not occur.

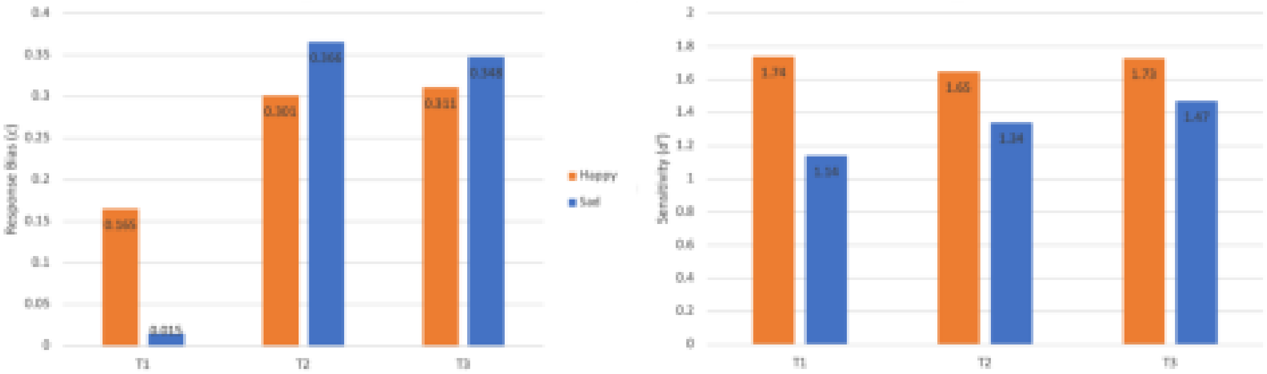
- **Effects of Go Viral!:** Post-intervention mean group ratings of unreliable headlines (metric 1) suggested susceptibility to misinformation was reduced from playing the game, a finding consistent with previous research employing this metric. However, signal detection analyses (metric 2) indicated better post-intervention discernment between reliable and unreliable headlines in the control group.

- **Mood and Inoculation:** No effects of mood while playing Go Viral! were detected once inductions had decayed. However, as inoculation did not occur the role of mood remains an open question.

Findings for the effects of mood on the detection of (un)reliable information have implications for micro-targeting of online misinformation (e.g., via textual emotion detection). Regarding Go Viral! and inoculation, the essential element of motivational threat was absent, and there were no benefits to reliability discernment resultant from gameplay. Despite this, the game did appear to reduce susceptibility to misinformation by the more-common metric employed in inoculation studies: mean ratings of unreliable news headlines only. Implications include that Go Viral!, and other interventions assessed by the usual metric, may not produce the claimed effects. Standardisation of assessment procedures in future research is encouraged.

Figure 1

Discernment (d') and Scepticism (c) on the MIST-20, Across Timepoints, by Mood Group



Note. In both graphs, differences between groups at each timepoint were significant for α = .001, as were T1-T2 differences for each group, and the difference-in-differences T1-T2, and T1-T3.

References

[1] W. J. McGuire. Advances in Experimental Social Psychology, volume 1, Some contemporary approaches, pages 191–229. Elsevier, 1964. https://doi.org/10.1016/S0065-2601(08)60052-0

[2] S. Lewandowsky and S. van der Linden. Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2):348–384, 2021. https://doi.org/10.1080/10463283.2021.1876983

[3] M. Basol, J. Roozenbeek, M. Berriche, F. Uenal, W. P. McClanahan, and S. van der Linden. Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. Big Data & Society, 8(1), 2021. https://doi.org/10.1177/20539517211013868

[4] R. Pine, T. Fleming, S. McCallum, and K. Sutcliffe. The effects of casual videogames on anxiety, depression, stress, and low mood: A systematic review. Games for Health Journal, 9(4):255–264, 2020. https://doi.org/10.1089/g4h.2019.0132

[5] J. Compton, B. Ivanov, and E. Hester. Inoculation theory and affect. International Journal of Communication, 16:3470– 3483, 2022. https://ijoc.org/index.php/ijoc/article/view/19094

[6] J. P. Forgas. Mood and judgement: The affect infusion model (AIM). Psychological Bulletin, 117(1):39–66, 1995. https://doi.org/10.1037/0033-2909.117.1.39

[7] J. P. Forgas. Feeling and doing: Affective influences on interpersonal behavior. Psychological Inquiry, 13(1):1–28, 2002. https://doi.org/10.1207/S15327965PLI1301_01

[8] J. A. Banas, and S. A. Richards. Apprehension or motivation to defend attitudes? Exploring the underlying threat mechanism in inoculation-induced resistance to persuasion. Communication Monographs, 84(2):164–178, 2017. https://doi.org/10.1080/03637751.2017.1307999

[9] J. P. Forgas. Happy believers and sad skeptics? Affective influences on gullibility. Current Directions in Psychological Science, 28(3):306–313, 2019. https://doi.org/10.1177/0963721419834543

[10] D. Marcusson-Clavertz, O. N. E. Kjell, S. D. Persson, and E. Cardeña. Online validation of combined mood induction procedures. PLOS ONE, 14(6):e0217848, 2019. https://doi.org/10.1371/journal.pone.0217848

Conference Program Abstract: The game *Go Viral!* aims to reduce susceptibility to misinformation via psychological inoculation. Similar games promote happiness, which causes information processing effects that could disrupt inoculation. However, effects may be extinguished by a motivating perception of threat, the provision of which is also a mandated element of inoculation. It is therefore important to consider mood and threat in the context of 'inoculation games'. We ran a randomised controlled experiment over two phases to assess effects of mood (happy/sad) and Go Viral! on susceptibility to misinformation. Happy/sad mood resulted in worse/better discernment between reliable and unreliable news headlines, while playing Go Viral! led to worse discernment relative to controls. Further, motivational threat was not higher in the Go Viral! group, suggesting inoculation did not occur. Findings have implications for micro-targeting misinformation by mood, the metrics by which misinformation susceptibility is assessed, and the status of Go Viral! as an inoculation intervention.

Emotions in misinformation studies: Distinguishing affective state from emotional response and misinformation recognition from acceptance
(Extended Abstract)

Jula Lühring[1,2]*, Apeksha Shetty[1,2]*, Corinna Koschmieder[3,4], David Garcia[2,5,6], Annie Waldherr[1] & Hannah Metzler[2,7,8 #]

*equal contributions

[#] corresponding author, metzler@csh.ac.at

[1] Department of Communication, University of Vienna, Austria

[2]Complexity Science Hub Vienna, Austria

[3]Institute of Psychology, University of Graz, Austria

[4] Center for Research Support, University College for Teacher Education, Graz, Austria [5] Department of Politics and Public Administration, University of Konstanz, Germany

[6]Institute of Interactive Systems and Data Science, Faculty of Computer Science and Biomedical Engineering, Graz University of Technology, Austria

[7]Center for Medical Data Science, Medical University of Vienna, Austria [8]Institute for Globally Distributed Open Research and Education

Misinformation is said to elicit emotion and trigger reactions such as commenting or sharing online. In particular high-arousal emotions like anxiety and anger may hinder critical reflection and elicit rapid, intuitive thinking — leaving people vulnerable to misinformation (Berger & Milkman, 2013; Boyer, 2021; Weeks, 2015). Observational (Chuai & Zhao, 2020; Pröllochs et al., 2021; Zollo et al., 2015) and experimental studies (Greenstein & Franklin, 2020; Martel et al., 2020; Weeks, 2015) on emotions and misinformation have measured aggregate emotions across individuals in different contexts. However, identical content can trigger different emotional reactions, often depending on people's prior beliefs (Mercier, 2020). Yet, earlier misinformation studies have not considered people's reasons for emotional reactions, nor the timing of emotions (Oatley & Johnson-Laird, 2014), which is related to their origin. Therefore, we investigated the role of emotions in processing misinformation assessing prior beliefs as well as emotions before and after exposure to real and false news.

In a pre-registered survey (*N* = 422; https://osf.io/tgzxr) with an Austrian sample (mean age: 33.97 +/- 15 years) collected by university students, participants rated the accuracy of false and real news headlines about COVID-19 and vaccine safety. Affective state measured before misinformation exposure, self-reported for "the last few days", did not correlate with discernment abilities between false and real news in contrast to a previous study by Martel et al. (2020). These null findings could be related to specifics of our sample (e.g., well-educated, left-leaning), or highlight the importance of the timing of emotions concerning misinformation processing. Second, we explored participants' emotional response to false and real news. Linear mixed effect models showed that participants reported anger when news contrasted with their COVID-19 beliefs: most participants with accurate COVID-19 beliefs felt significantly more angry after the exposure to false news, while participants with more false beliefs about COVID-19 reported more anger when exposed to real news. Text analysis of open-ended responses further underlined that only a minority of participants seemed angry because they believed the false news, whereas most expressed anger about the existence of such news. Consistent with this, the relationship between people's news discernment ability and elicited emotions was curvilinear: Both lower and higher discernment abilities co-occurred with higher anger. Our findings align with previous research suggesting that when people encounter

information that conflicts with their preexisting beliefs, they perceive it as a threat to their worldview, which elicits a negative emotional response (Nyhan & Reifler, 2010; Trevors, 2022). In summary, our results show that emotions do not just broadly increase susceptibility to misinformation and decrease people's ability to discern false from real news. Instead, their potential influence varies based on factors such as prior beliefs and whether emotions are experienced before or in response to exposure to misinformation. Misinformation studies on emotions therefore need to consider people's reasons for feeling the emotion, as well as the timing of the emotional experience.

References

Berger, J., & Milkman, K. L. (2013). Emotion and Virality: What Makes Online Content Go Viral? Marketing Intelligence Review, 5(1), 18–23. https://doi.org/10.2478/gfkmir-2014-0022

Boyer, M. M. (2021). Aroused Argumentation: How the News Exacerbates Motivated Reasoning. The International Journal of Press/Politics, 19401612211010576. https://doi.org/10.1177/19401612211010577

Chuai, Y., & Zhao, J. (2020). Anger makes fake news viral online (arXiv:2004.10399). arXiv. https://doi.org/10.48550/arXiv.2004.10399

Greenstein, M., & Franklin, N. (2020). Anger Increases Susceptibility to Misinformation. Experimental Psychology, 67(3), 202–209. https://doi.org/10.1027/1618-3169/a000489

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. Cognitive Research: Principles and Implications, 5. https://doi.org/10.1186/s41235-020-00252-3

Mercier, H. (2020). Not Born Yesterday. https://press.princeton.edu/books/hardcover/9780691178707/not-born-yesterday Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. Political Behavior, 32(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Oatley, K., & Johnson-Laird, P. N. (2014). Cognitive approaches to emotions. Trends in Cognitive Sciences, 18(3), 134–140. https://doi.org/10.1016/j.tics.2013.12.004

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021). Emotions in online rumor diffusion. EPJ Data Science, 10(1), 51. https://doi.org/10.1140/epjds/s13688-021-00307-5

Trevors, G. J. (2022). The Roles of Identity Conflict, Emotion, and Threat in Learning from Refutation Texts on Vaccination and Immigration. Discourse Processes, 59(1–2), 36–51. https://doi.org/10.1080/0163853X.2021.1917950

Weeks, B. E. (2015). Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation: Emotions and Misperceptions. Journal of Communication, 65(4), 699–719. https://doi.org/10.1111/jcom.12164

Zollo, F., Novak, P. K., Vicario, M. D., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Emotional Dynamics in the Age of Misinformation. PLOS ONE, 10(9), e0138740. https://doi.org/10.1371/journal.pone.0138740

Preventing Profiling for Ethical Fake News Detection by Correlating Articles and Their Online Spreaders (Extended Abstract)

Liesbeth Allein[1], Marie-Francine Moens[1], and Domenico Perrotta[2]

[1] Department of Computer Science, KU Leuven, Belgium
[2] European Commission, Joint Research Centre (JRC), Italy

1 Introduction

Social media users are important actors in a fake news article's dissemination online and provide valuable insights into its appeal. It is imperative for automated detection methods to heed such social context. However, existing approaches that take both textual content and spreader information as input risk relying on *profiling*. We address the unethical nature of profiling-dependent decision-making [2, 3] in the fake news detection task and introduce a novel method for models to avoid profiling while still leveraging the rich insights on social context held by social media users[3].

2 Methodology

2.1 Task Definition

A fake news classifier *f* with a text encoder *h* and a classification layer *c* predicts whether a given news article *a* is *true* (y = 1) or *fake* (y = 0).

$$h : a \ 7\longrightarrow a', a' \in R^{m} \ (1)$$

$$c : a' 7\longrightarrow y, y \in \{0, 1\} \ (2)$$

During model training, user encoder *g* independently models *N* Twitter users *u* who spread *a* online and projects them onto the same latent space as a'. Here, *u* is represented by their profile description and 200 latest tweets.

$$g : u \ 7\longrightarrow u', u' \in R^{m} \ (3)$$

2.2 The Proposed Learning Algorithm

The parameters of *f* and *g* are optimised using a weighted combination of three loss functions.

Classification loss Cross-entropy loss on *y* to optimise prediction performance.

User-article correlation loss Loss based on the cosine distance between the latent representation of the article, a', and that of each user, u'. It correlates an article with its spreaders.

User-user correlation loss Loss based on the cosine distance between the latent repre sentations of all users, u'. It correlates users spreading the same article on Twitter.

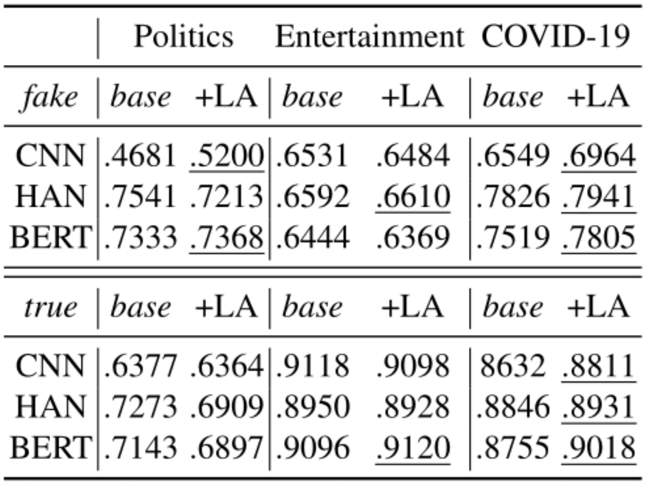
[3] This extended abstract presents published work: Liesbeth Allein, Marie-Francine Moens and Domenico Perrotta (2023). Preventing profiling for ethical fake news detection. *Information Processing & Management, 60*(2), 103206.

The latter two losses follow the proposed correlated identity assumption. By enforcing correlation between articles and their spreaders, and between those spreaders, the learning algorithm integrates the social context of *a* in the parameters of *f*. This way, *f* is inspired by social context, yet it does not perform profiling since it does not take *u* as input.

Correlated identity assumption: The identity contained in a news article correlates with the identity of its spreaders, and vice versa (user-article correlation). If an article correlates with each of its spreaders, then all spreaders should portray commonalities among each other (user-user correlation). In the case of Twitter, a spreader's identity is reflected in their tweet collection and profile description, among other things.

3 Experiments

The experiments are conducted using data from different news areas: politics, entertainment, and COVID-19 [6, 8]. We experiment with three different text encoders for *h* and *g*: CNN [4], HAN [7], DistilBERT [5]. Table 1 presents the classification results, showing consistent improvement for COVID-19 news.



	Politics		Entertainment		COVID-19	
fake	*base*	+LA	*base*	+LA	*base*	+LA
CNN	.4681	.5200	.6531	.6484	.6549	.6964
HAN	.7541	.7213	.6592	.6610	.7826	.7941
BERT	.7333	.7368	.6444	.6369	.7519	.7805
true	*base*	+LA	*base*	+LA	*base*	+LA
CNN	.6377	.6364	.9118	.9098	8632	.8811
HAN	.7273	.6909	.8950	.8928	.8846	.8931
BERT	.7143	.6897	.9096	.9120	.8755	.9018

Table 1: Overview of the performance results (F1-score) for the *fake* and *true* labels. The underlined results indicate that the classifier optimised using the learning algorithm (+LA) outperforms the classifier optimised only on the classification loss (*base*).

4 Main Findings

- Classifiers are equally competitive when looking at the characteristics of an article's early and late dissemination audience.
- The success of the learning algorithm remains steady when changing the selection of tweets to represent the users.
- Statistical visualisation and dimension reduction techniques show that the user inspired classifiers better discriminate between unseen fake and true news in their latent spaces.

5 Conclusion

In contrast to popular profiling-avoiding methods, we show that a fake news classifier does not need to ignore available user information or artificially alter user representations using debiasing techniques to prevent profiling. Instead, the classifier can still benefit from the user modality by indirectly integrating social context in its parameters via the loss function. Our study serves as a stepping stone to resolve the underexplored issue of profiling-dependent decision-making in user-informed fake news detection.

References

1. Allein, L., Moens, M.F., Perrotta, D.: Preventing profiling for ethical fake news detection. Information Processing & Management 60(2), 103206 (2023). https://doi.org/10.1016/j.ipm.2022.103206

2. European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union L110 59, 1–88 (4 May 2016)

3. High-Level Expert Group on AI: Ethics Guidelines for Trustworthy AI. Tech. rep., European Commission, Brussels, Belgium (2019)

4. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1746–1751 (Oct 2014). https://doi.org/10.3115/v1/D14-1181

5. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)

6. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 8(3), 171–188 (2020). https://doi.org/10.1089/big.2020.0062

7. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages=1480–1489 (2016). https://doi.org/10.18653/v1/N16-1174

8. Zhou, X., Mulay, A., Ferrara, E., Zafarani, R.: ReCOVery: A multimodal repository for covid-19 news credibility research. In: Proceedings of the 29th ACM Interna tional Conference on Information & Knowledge Management. pp. 3205–3212 (2020). https://doi.org/10.1145/3340531.3412880

Psychological inoculation strategies to fight climate disinformation across 12 countries

Tobia Spampatti[1,2]*, Ulf J. J. Hahnel[1,3], Evelina Trutnevyte[4], and Tobias Brosch*[1,2]

[1] Swiss Centre for Affective Sciences, University of Geneva, Geneva, Switzerland

[2] Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

[3] Faculty of Psychology, University of Basel, Basel, Switzerland

[4] Renewable Energy Systems, University of Geneva, Switzerland

* Corresponding authors: Tobia Spampatti (tobia.spampatti@unige.ch), Tobias Brosch (tobias.brosch@unige.ch).

Abstract

Decades after the scientific debate about the anthropogenic causes of climate change has been settled, climate disinformation still challenge the scientific evidence in public discourse. Here, we present a comprehensive theoretical framework of (anti)science belief formation and updating to account for the psychological factors (cognitive and socioaffective, Ecker et al., 2022) and core communicational bases (receiver, sender of a message, message itself; Phillip

Muller et al., 2022) that influence acceptance or rejection of scientific messages. We experimentally investigated, across twelve countries (USA, Canada, UK, Ireland, Australia, New Zealand, Singapore, Philippines, India, Pakistan, Nigeria, and South Africa; N=6816), the effectiveness of six theory-derived inoculation strategies each targeting one of the six factors identified in the framework – scientific consensus, trust in scientists, transparent communication, moralization of climate action, accuracy, and positive emotions – to fight real

world disinformation about climate science and climate mitigation actions. The effects of the psychological inoculations and the climate disinformation statements were measured against participants' climate beliefs, affective reactions towards climate mitigation action, climate related truth discernment, and pro-environmental behavior. While exposure to disinformation had strong detrimental effects on participants' climate change beliefs (Cohen's δ=-0.16), affect towards climate mitigation action (Cohen's δ=-0.33), ability to detect disinformation (Cohen's δ=-0.14), and pro-environmental behavior (Cohen's δ=-0.24), we found almost no evidence for protective effects of the theory-driven psychological inoculations (all Cohen's δs<0.20). We discuss the implications of these findings and propose ways forward to systemically fight climate disinformation.

COM-PRESS: An Image Manipulation Analysis Dashboard for Fact-checkers (Extended Abstract)

Hannes Mareen[1], Stephanie D'haeseleer[2], Kristin Van Damme[2],
Tom Evens[2], Peter Lambert[1], Glenn Van Wallendael[1]
[1]IDLab, Ghent University – imec, Ghent, Belgium
[2]imec-mict-UGent, Communication Sciences, Ghent University – imec, Ghent, Belgium
e-mail: firstname.lastname@ugent.be

The advancement of artificial intelligence has made it easier to manipulate images. For example, FireFly or Photoshop Generative Fill enables users to realistically add or delete objects in an image without the need for advanced technical skills. As a result, there is a growing concern that manipulated images will be used more frequently for disinformation purposes. To address this challenge, we present COM-PRESS, a tool designed to equip journalists with efficient image manipulation detection methods, aiding them in fact-checking processes. COM-PRESS results from an interdisciplinary project, which fosters collaboration between computer and communication scientists. The tool is in its alpha version and publicly available on https://com-press.ilabt.imec.be/.

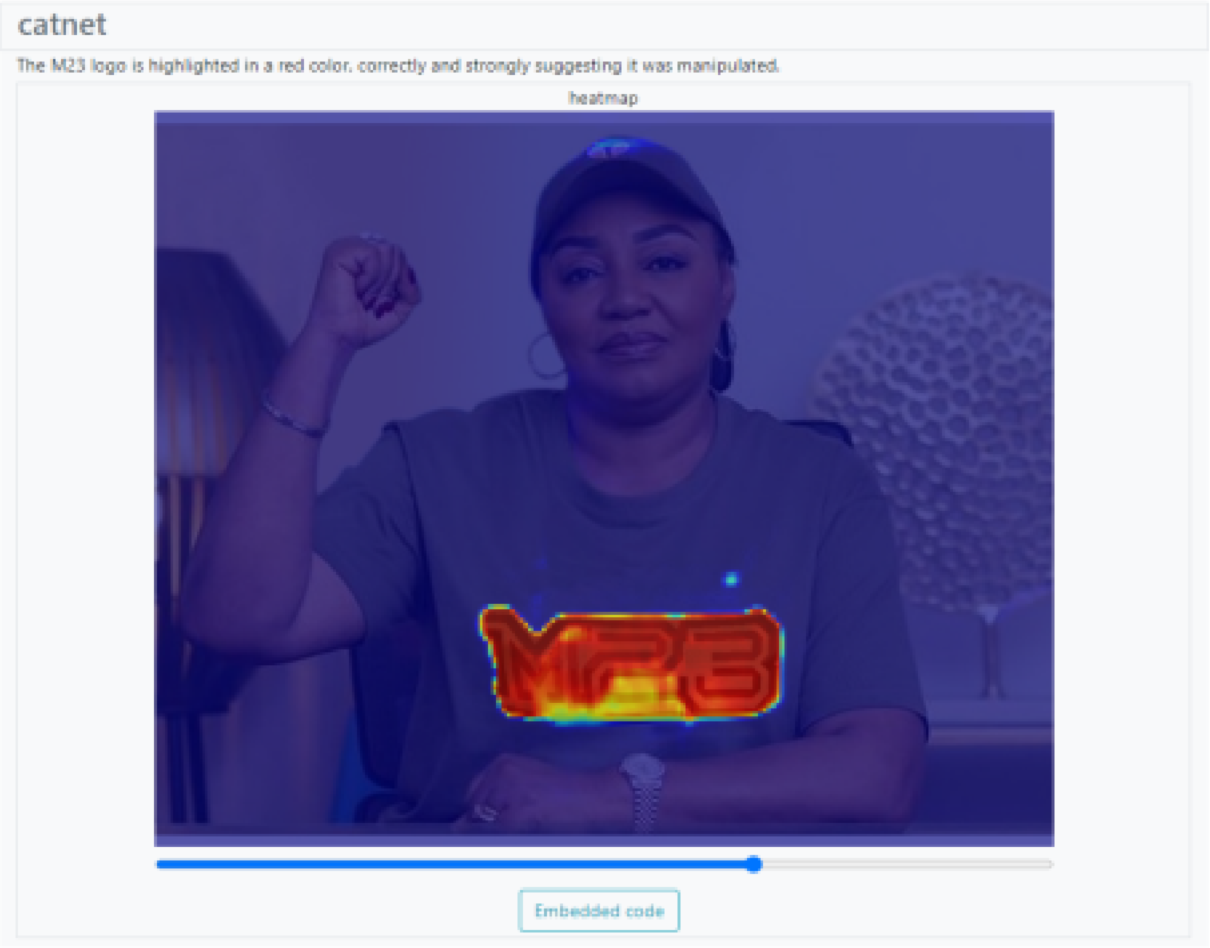
The workflow to analyze images in COM-PRESS is as follows. First, users upload an image of interest (i.e., fact check worthy). The image is then analyzed using multiple state-of-the-art forgery detection methods. The results are then visually presented on the website as heatmaps that highlight potential manipulations or inconsistencies.

To aid the online publication of fact-checks, the dashboard provides two additional features which were derived from interviews with fact-checkers at the beginning of the research project. First, a button to show embedded code is provided, which journalists can copy and paste in their online article. The embedded code enables the visualization of a heatmap image along with a slider to change the transparency to make the underlying image of interest visible. Within fact checking, transparency is key, and such an embedded code thus strengthens the online fact-check publication. Second, fact-checkers or consulted experts can add comments to the images, so that the COM-PRESS result page can be linked to in articles, and readers are provided with contextual information and interpretations. Fact-checkers oftentimes rely on expert interpretations, hence this was an essential feature to fit the tool in the current fact-checking practices. An example of such a result page is shown in Fig. 1 and on https://com-press.ilabt.imec.be/result/kp0RcotFa.

The dashboard currently incorporates the following forgery detection methods: BLK [4], CAGI [2], DCT [6], Noiseprint [1], Comprint [5], and CAT-Net [3]. The first three are conventional methods, whereas the latter three are more recent deep learning methods. In future versions of COM-PRESS, we will incorporate additional (fused) detection methods. More over, we will provide more transparency on the performance of the methods, as well as a tutorial on how to interpret the results, to meet previously detected fact-checker's needs. Furthermore, we will perform a user study to pinpoint opportunities to improve the dashboard.

Compared to existing dashboards (e.g., MeVer Image Verification Assistant on https://mever.iti.gr/forensics/), COM-PRESS has the following novelties: integration of Comprint [5] and CAT-Net [3], embedded-code buttons, and the ability to add comments.

In conclusion, COM-PRESS offers journalists a valuable resource to enhance image fact-checking and combat dis information. By leveraging multidisciplinary approaches and open-source detection methods, it equips journalists with additional tools necessary to verify the authenticity of images and promote the dissemination of accurate information in the face of growing online disinformation.



The M23 logo is highlighted in a red color, correctly and strongly suggesting it was manipulated.

Figure 1: Example result page: https://com-press.ilabt.imec.be/result/kp0RcotFa, with comments to aid interpretation, a slider to change the heatmap transparency, and an embedded-code button.

References

[1] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNN-based camera model fingerprint. IEEE Trans. Inf. Forensics Security, 15:144–159, 2020.

[2] Chryssanthi Iakovidou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. Content-aware de tection of JPEG grid inconsistencies for intuitive image forensics. Journal of Visual Communication and Image Representation, 54:155–170, 2018.

[3] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. International Journal of Computer Vision, 130(8):1875– 1895, 2022.

[4] Weihai Li, Yuan Yuan, and Nenghai Yu. Passive detection of doctored JPEG image via block artifact grid extraction. Signal Processing, 89(9):1821–1829, 2009.

[5] Mareen, Hannes and Vanden Bussche, Dante and Guillaro, Fabrizio and Cozzolino, Davide and Van Wallendael, Glenn and Lambert, Peter and Verdoliva, Luisa. Comprint: Image Forgery Detection and Localization Using Com pression Fingerprints. In Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges, Lecture Notes in Computer Science, pages 281–299. Springer Nature Switzerland, 2023.

[6] Shuiming Ye, Qibin Sun, and Ee-Chien Chang. Detecting digital image forgeries by measuring inconsistencies of blocking artifact. In IEEE International Conference on Multimedia and Expo, pages 12–15, 2007.

How did I end up here? Advocating for situated media literacy

Tamara Witschge, Jeroen de Vos & Sabine Niederer

Amsterdam University of Applied Sciences

Media literacy is an important element in countering the harmful effects of dis- and misinformation. To date, the main perspective on media literacy can be described as a 'liberalist' view of media literacy, focused on the nature and source of individual messages (Phillips and Milner, 2021). In this approach, media literacy initiatives mainly focus on equipping 'citizens with the necessary skills to make sense of the message they read, see and hear' (Phillips and Milner, 2021: 151). Such initiatives, though important and well-intended, are limited in effect and in some cases even strengthen rather than diminish the harmful impact of specific content and messages. In this paper, we will report and reflect on the insights from the project "Putting Disinformation on the Map" in which we use creative, participative, visual and digital methods to understand how we can expand our understanding of media literacy. A central question to our inquiry is how to orientate and situate yourself in a media landscape which is inherently interconnected, networked and interdependent, also understanding yourself as a contributing actor: How can we experience our place and role in the media landscape and the workings of more structural elements of platforms, algorithms and media logic explicit and tangible?

In our research, we develop and test creative interventions that invite people to become aware of the workings of online networks and algorithms: How one might easily travel from one point of a particular debate or discourse to another, rapidly jumping to different visual cultures, narratives and vernaculars, ending up somewhere deep within a coherent frame which is hard to match and thus counter from within another frame. Particularly as this often happens without being explicitly aware of the inherent connectedness of this media landscape, it is beneficial to show how one "ends up" in such a place, and also to consider one's own role in shaping the network. How can you understand where you are in relation to these different parts of the landscape, different types of more or less problematic online discourses? What places are in the vicinity, and what might show up when you make a certain turn opposed to taking another exit - and what signposts and signposting practices might look like to help the traveller understand their position?

Operationalisting the extensive conceptualisation of ecological literacy developed by Philips and Milner (2021) in their book "You are here," we analyse the online content on the theme of "weight loss" in the online platforms TikTok, Telegram and Instagram (see figure 1 for one of the outputs). We employ three larger strands of design research; 1) collaborative mapping and mapmaking , 2) co-designing navigation and signposting materials and 3) artistic creation of storytelling devices drawing on generative AI. Under the header of "situated media literacy", in this paper, we endeavour to help build a more expansive view of the agency of individuals in a complex media landscape which is inherently connected, ever-changing with different moving elements being interdependent of one another.

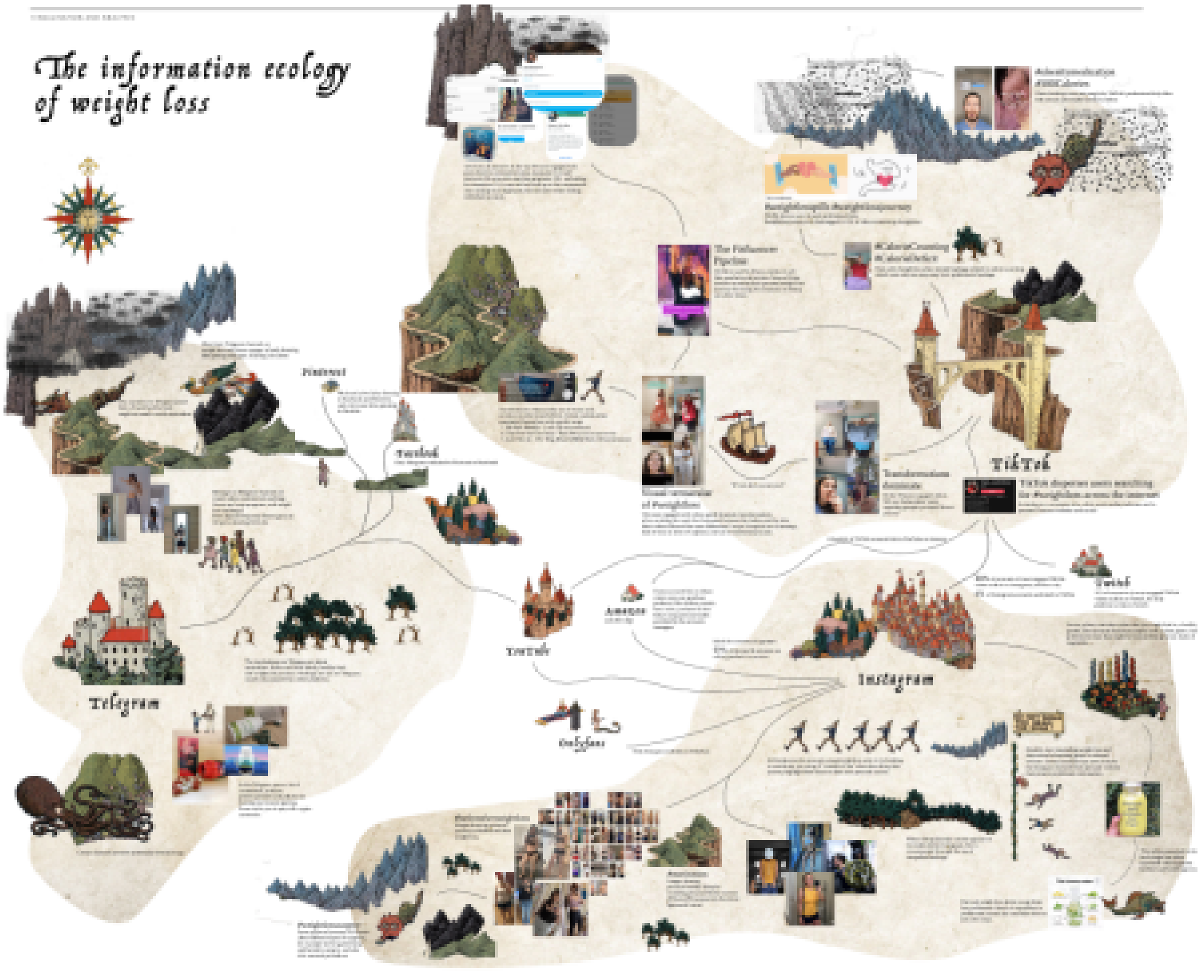


Fig 1. Media landscape of ‘weight loss’, this map was produced using the content, narratives and visual culture one might encounter if you would query different platforms for ‘weighloss’, reappropriated drawing of the visual elements of the Land of Make Believe (Jaro Hess from the 1930s) in a collaborative map-making program.

This ongoing research project is funded by the SIDN fund, project partners include the Dutch Sound and Vision (Beeld & Geluid), KB (Nationale Bibliotheek) and journalism partner Verspers.

References

Phillips, W., & Milner, R. M. (2021). You are here: A field guide for navigating polarized speech, conspiracy theories, and our polluted media landscape. MIT Press.

Additional short abstract

Media literacy is an important element in countering the harmful effects of dis- and misinformation. To date, the main perspective on media literacy can be described as a ‘liberalist’ view of media literacy, focused on the nature and source of individual messages (Phillips and Milner, 2021). In this approach, media literacy initiatives mainly focus on equipping ‘citizens with the necessary skills to make sense of the message they read, see and hear’ (Phillips and Milner, 2021: 151). In this paper, we report and reflect on the insights from the project “Putting Disinformation on the Map” in which we use creative, participative, visual and digital methods to

understand how we can expand our understanding of media literacy. A central question to our inquiry is how to situate yourself in a media landscape which is inherently interconnected, networked and interdependent, also understanding yourself as a contributing actor.

False or not true: should fact-check headlines avoid negations in favour of fact affirmations? (Extended Abstract)

Ferre Wouters (KU Leuven)
Michaël Opgenhaffen (KU Leuven)
Marina Tulin (University of Amsterdam)
Michael Hameleers (University of Amsterdam)

This study aims to investigate which type of headlines work best in debunking false claims. In the first part of our research, we coded the types of headlines of fact check articles in the Dutch language over a period of 12 months (N=976). We found that the majority used either one of two opposite strategies to correct misinformation: negative and affirmative phrasings. The former, called the 'negation-tag-model', repeats the false claims and adds a negation tag (e.g. 'NO, head of Ukrainian army does NOT wear bracelet with swastika'). In affirmative headlines, also called 'fusion model', the negation is fused into an affirmative by using antonyms or stating the correct fact (e.g. 'Head of Ukrainian army wears bracelet with Viking symbols').

The second part of our research builds on psycholinguistic studies (Fiedler et al., 1996; Giora et al., 2005; Mayo et al., 2014) that suggest negation tags in headlines gradually disappear from people's memory, leading to false memory. E.g. in an experiment by Maciuszek & Polczyk (2017), participants were exposed to descriptions of a room, and misremembered the negated features of the room as if they were present. An explanation for this is that negation tags are more exhausting to process. Sentences containing a negation evoke longer reaction times compared to sentences that do not; the so-called 'polarity effect' (Agmon, 2022).

A headline is a very prominent and visible part of a fact check article, especially in online and social media environments (Piotrkowicz et al., 2017), where fact-checking organisations mostly operate. Although previous research in the context of fact-checking does not point to a backfire effect of negations (Swire et al., 2016 & 2017; Ecker et al., 2020), negative headlines might still be less effective in reducing belief in false claims vis-à-vis affirmative phrasings. We tested this hypothesis. Furthermore, the study goes beyond claim belief, also considering the impact on the reader's perception of the fact check.

An online survey experiment was conducted with 1,500 participants from Flanders. The participants were divided into different groups and exposed to headlines using either the negation-plus-tag model or the fusion model. Two separate formats were used: one where participants read a complete fact-check article and another that simulated a social media environment with only the fact-check headline visible. There was also a control condition, where respondents were exposed to the false claim instead of a fact check. The experiment used different stories about food production in Brazil and a parking law in Wales. We used real but foreign articles to minimise familiarity. Participants were asked questions about their perception of the fact-check, with a 12-hour delay for questions about belief in the claim.

The results suggest that both kinds of fact check headlines are as effective in discerning true from wrong claims, and thus do not confirm the hypothesis. However, full articles demonstrate better accuracy estimation and foster more positive perceptions compared to fact checks on social media. This research provides insights into optimal headline approaches for debunking false claims and combating misinformation.

References

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy, 3, e32.

Levine, T. R. (2014). Truth-default theory (TDT) a theory of human deception and deception detection. Journal of Language and Social Psychology, 33(4), 378-392.

Who checks the fact-checkers? Studying the work of External Assessors behind fact-checking organizations (*Extended abstract)*

Marilia Gehrke & Ansgard Heinrich
Centre for Media and Journalism Studies, University of Groningen

Fact-checking has emerged as a set of practices that aim to invent a new style of political news based on truth-seeking and holding public figures accountable (Graves, 2016). When verifying claims *a posteriori*, independent fact-checking organizations are commonly seen not to compete but add to the journalism repertoire. Conventional news outlets have been criticized for reproducing politicians' claims in a clickbait strategy and non-critical manner, potentially increasing the noise around journalism (Gehrke et al., 2022). Yet, fact-checking organizations have also been accused of bias. Such criticisms have triggered a dispute about truth claims. With more voices adding to information exchange, with an increase in (digital) platforms promoting themselves as ultimate arbiters of truth, drawing boundaries between information, disinformation, and misinformation appears to become a struggle over authority and over the question of who determines what is 'true' and 'false.' In an anecdotal but genuine inquiry, people often ask: **who checks the fact-checkers?** This paper attempts to contribute to this debate by investigating the accountability operations behind fact-checking organizations. Specifically, we focus on the 'behind-the-scenes' of fact-checking organizations and ask: **How do external assessors and advisory boards assist in establishing credibility around fact-checking operations?**

In research on fact-checking organizations, scholars have so far addressed topics such as fact checkers' work, perceptions, and their role on social media in distributing corrections (e.g., Dafonte-Gómez, Míguez-González & Ramahí-García, 2022; Rodríguez-Pérez & Seibt, 2022;
Tsang, Feng & Lee, 2022). In addition, researchers have examined the effects of fact-checking on the population, particularly voting behavior (Nyhan et al., 2020). Others have investigated the efficacy of this practice by establishing labeling systems (Oeldorf-Hirsch et al., 2023). However, codes of ethics and conduct have so far rarely been scrutinized. In addition, little attention has been given to those who operate *behind the scenes*, namely external assessors and advisory board members. Yet, who are these people who help to build and professionalize accountability structures around fact-checking organizations, and how do they operate? With this overarching question in mind, we further explore 1) what expertise they bring to the job, 2) which codes and principles guide them, and 3) how they perceive their role vis-à-vis the respective fact-checkers.

To shed light on these questions, we will conduct 20 in-depth interviews with external assessors and advisory board members working for the International Fact-Checking Network (IFCN). As an internationally leading organization, IFCN has established a detailed code of principles to guide fact-checking practices (IFCN, 2023). In order to become part of it, organizations need to run through a rigorous verification process that is overseen by IFCN's Advisory Board and external assessors. This study uses the list of signatories of ICFN as a starting point to identify potential interviewees on those assessors overseeing fact-checking operations in multiple countries, ensuring continent diversity. Initiatives such as the recently created European Fact Checking Standards Network (EFCSN) will also be a departure point to find interviewees. We will make sure that our selection includes diverse countries and political systems. All in all, this study thus sets out to enhance our knowledge about how fact-checking organizations attempt to build credibility. Ultimately, it draws the curtain to examine who checks the fact-checkers to better understand the accountability mechanisms at play in the independent fact-checking scene.

References

IFCN Code of Principles (2023, July 12). https://ifcncodeofprinciples.poynter.org/know more/the-advisory-board-and-our-pool-of-assessors

Gehrke, M., Träsel, M., Ramos, Á. K., & Ozorio, J. (2023). All the President's Lies: How Brazilian News Media Addressed False and Inaccurate Claims in Their Titles. Journalism Practice, 1–18. https://doi.org/10.1080/17512786.2023.2174579

Graves, L. (2016). Deciding what's true: the rise of political fact-checking in American Journalism. New York: Columbia University Press.

Gómez, A. D., Míguez-González, M. I., & Ramahí-García, D. (2022). Fact-checkers on social networks: analysis of their presence and content distribution channels. Comunicacion Y Sociedad, 35(3), 73–89. https://doi.org/10.15581/003.35.3.73-89

Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2023). The Influence of Fact-Checking Is Disputed! The Role of Party Identification in Processing and Sharing Fact Checked Social Media Posts. American Behavioral Scientist, 000276422311743. https://doi.org/10.1177/00027642231174335

Nyhan, B., Porter, E., Reifler, J., & Wood, T. K. (2019). Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability. Political Behavior, 42(3), 939–960. https://doi.org/10.1007/s11109-019-09528-x

Rodríguez-Pérez, C. V., & Seibt, T. (2022). critérios dos fact-checkers brasileiros. Brazilian Journalism Research, 18(2), 350–373. https://doi.org/10.25200/bjr.v18n2.2022.1510

Tsang, N. L. T., Feng, M., & Lee, F. L. F. (2022). How fact-checkers delimit their scope of practices and use sources: Comparing professional and partisan practitioners. Journalism: Theory, Practice & Criticism, 146488492211008. https://doi.org/10.1177/14648849221100862

A decolonial feminist approach to gendered disinformation (Extended abstract)

Marilia Gehrke
Centre for Media and Journalism Studies, University of Groningen, The Netherlands

False and misleading content has been produced and targeted at women to delegitimize their trajectory and work in prominent societal positions. Still barely explored in mis- and disinformation studies, *gendered disinformation* mainly encompasses politicians, activists, and journalists. In line with Jankowicz *et al.* (2021), I argue it is a long-term misogynist strategy to undermine women's participation and decision-making in public life. Because only what gets counted counts (D'Ignazio & Klein, 2020), fabricated content targeting women urges to be categorized.

This study analyzes how Brazilian female politicians have been portrayed in misogynist and fabricated narratives. Besides being a prominent country in South America, Brazil has a violent past of colonization and slavery. Thus, domination and oppression have been part of society and resonate today. Through multiple case study (Stake, 2006) and documentary research as methodological approaches, the sampling includes six women with different ethnical backgrounds and political experiences. They are the former president Dilma Rousseff; former senator and Minister of Environment and Climate Change, Marina Silva; former councilwoman Marielle Franco; the Minister of Indigenous Peoples, Sonia Guajajara; former congresswoman Manuela D'Avila, and First Lady Rosangela da Silva (Janja).

Rousseff, Marina Silva, Guajajara, and D'Avila were chosen due to their political prominence in national politics since they ran for president or vice president in recent years. Franco was a victim of murder due to her political activities, and the crime remains unresolved five years later.
Janja became first lady at the beginning of 2023, and her performance in public life is active and engaging – different than the stereotype that these women should behave in a certain way and stay behind the curtains. With different ethnic backgrounds and ages, these female politicians symbolize the intersection of power and race.

Historically, black women in Brazil have been particularly subject to violence: 45% of them reported they were victims of aggression at some point. Among white Brazilians, this percentage drops by almost ten points (Bueno et al., 2023). My assumption is that (online) political violence repeats this pattern. Through decolonial feminism lenses (Vergès, 2021) and adopting a multimodal approach, I explore the commonalities and differences related to word choice and images used to delegitimize women with different backgrounds in a country historically marked by oppressing minorities. The data selected for the analysis includes fact-checking texts, self biographical information published in books and social media, and the database used by *Aos Fatos* fact-checking organization in a news article about misogyny against politicians (Rudnitzki & Barbosa, 2023).

According to Sobieraj (2020), black women receive gender and race-based attacks that deploy the specific stereotypes and myths that have been used to pathologize them. Additionally, when explaining this intersectionality, Vergès (2021) provokes us with the question, "Who cleans the world?". In colonized countries, it is not expected for (especially black) women to ascend to powerful positions. Consequently, the anger toward them is potentialized.

Despite presenting several cases, my contribution here is not only empirical. Instead of discussing gendered disinformation as something that affects women equally, I argue that there are layers of complexity behind those false and misleading narratives. Decolonial feminism was not yet explored within the gendered disinformation framework. Thus, the results encountered in this pilot study will guide the direction of new investigations in a broader research project.

References

Bueno, S. et al. (2023). Fórum Brasileiro de Segurança Pública & Instituto Datafolha (2023, July 2023).https://forumseguranca.org.br/publicacoes_posts/visivel-e-invisivel-a-vitimizacao-de mulheres-no-brasil-4a-edicao/

D'Ignazio, C. & Klein, L. (2020). Data Feminism. Cambridge: MIT Press.

Jankowicz, N. Hunchak, J. Pavliuc, A. Davies, C. Pierson, S. Kaufmann, Z. (2021). Malign creativity: how gender, sex, and lies are weaponized against women online. Wilson Center: Science and Technology Innovation Program.

Rudnitzki, E. & Barbosa, J. (2023). WhatsApp concentra ondas de ataques misóginos a mulheres na política. Aos Fatos. https://www.aosfatos.org/bipe/whatsapp-ataques-misoginos-mulheres politica/

Stake, Robert. (2006). Multiple case study analysis. New York: The Guilford Press.

Sobieraj, Sarah. (2020). Credible threat: attacks against women online and the future of democracy. Oxford: Oxford University Press.

Vergès, Françoise. (2021). A decolonial feminism. London: Pluto Press.

Generative AI tools and disinformation perceptions (Extended Abstract)

Authors: Michael Sivolap, Marina Tulin, Christopher Starke, Tom Dobber

Affiliation: University of Amsterdam, ASCoR

The recent introduction of generative artificial intelligence (AI) tools both in business and private life, have raised concerns over how these tools may be used for malicious purposes such as the fabrication of deceptive content(Bontridder & Poullet, 2021).

This study focuses specifically on how generative AI tools like ChatGPT and Midjourney relate to disinformation perceptions. Because these tools can easily be abused to fabricate false information, this study asks to what extent learning about these tools contributes to more distrust and cynicism in information. While people have a general tendency to trust the information they see, a tendency known as truth default (Levine, 2014), we are interested in whether generative AI tools have the ability to undermine this. This study employs an experimental between-subjects design with three conditions. Participants in the experimental conditions watch one of two explainer videos on generative AI tools, namely ChatGPT (for text) combined with Midjourney (for images). The videos explain how these tools can be used to generate text and matching images. The explainer videos vary in the purpose the AI tools are used for. In one experimental condition, the explainer video shows how the AI tools can be used for entertainment purposes, namely to generate art and a poem. This is to mimic the intended use of these tools as advertised by the creators. In the other experimental condition the explainer video shows how the AI tools can be used to generate false information, namely a fake news article and matching photograph. This is to mimic the ways that the news media have been reporting on the possible dangers of generative AI tools. A control condition was also included where participants are not exposed to any AI tools. The dependent variable consists of ratings of 10 news media items showing a headline and matching image. Five of these stimuli contain true information taken from news media, while the other five will contain false information generated by AI tools. Participants are asked to judge the veracity of each stimulus as well as the extent to which they are certain of their answers. The stimuli are presented in random order. We hypothesize that individuals are overall more likely to rate online information as false after being exposed to generative AI tools, irrespective of whether the information was actually true or false. In other words, we expect that exposure to AI tools undermines individuals' truth default bias. We also expect that

this effect is stronger for individuals who are exposed to AI tools being used to generate false information compared to individuals who learn that AI tools are used for entertainment purposes. Finally, we hypothesize that uncertainty is a mediator in these relationships, such that individuals who learn about AI tools experience more uncertainty, and that this in turn increases individuals' tendency to judge information as false.

At the time of writing, this study has received ethical approval from the Ethics Review Board of the university of the research team. The pre-test has been launched and the main study is set up in collaboration with the survey company Dynata. The data collection is set for August 2023 with a target sample size of 900 adults (18 years or older) from the population of Canada. Soft quotas are used on sex, age, and education level to achieve a sample that approximates the distribution of the latest Canadian census data.

References

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy, 3, e32.

Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. Journal of Language and Social Psychology, 33(4), 378–392. https://doi.org/10.1177/0261927X14535916

Public perceptions about media as sources of (dis)information about the Ukraine war: Evidence from Romania

Georgiana Udrea, Alina Bărgăoanu, Nicoleta Corbu

Extended abstract

In recent years, disinformation has become not only a structural problem of the European Union and beyond, but alsoamatter of public perception. In the context of the Ukrainian war, the subject of disinformation vehiculated in both mainstream and social media has received much public attention and debate, especially in war bordering countries, suchasRomania. The current literature in the field is still in its infancy, with papers investigating people's wellbeing (Vintilă et al., 2023), dominant Russian narratives in the media (Grinko & Baeriswyl, 2023), incidence of pro-Ukraine vs. pro-Russianviews in social media (Dastgeer & Thapaliya, 2023), manipulation of Russian propaganda in online (Kling et al, 2022; Williams & Carly, 2023) and social media (Geissler et al., 2022). Additionally, in war-bordering countries, people report seeing misleading information about the war in the news more than twice as much as people in distant countries (Newmanet al., 2023). However, little is still known about citizens' perception of main sources of potential disinformation relatedtothe Ukraine war. In this context, news users in Romania are likely to perceive that much of the information coming fromvarious sources is false. Hence, the overarching research question of the study is: What are the main predictors of people'sperception about the sources of disinformation about the Ukraine war? Using a national survey (N=1000) conductedoneyear after the beginning of the war (March 2023) in Romania using an online panel (conducted by Dynata, using soft quotas for age, gender, and education), we found that people estimate that most of the information coming fromRussianmedia is false (M=5.56, SD=1.66 on a 7-point Likert scale), but equally, even though to a lesser extent, informationcomingfrom Ukrainian media (M=4.41, SD=1.71), Romanian media (M=4.36, SD=1.62), and Western countries in general (M=4.14, SD=1.62). Results of OLS regression models predicting people's perception of each type of media as potential source of disinformation show that holding a conspiracy mindset, along with a general perception of a high incidence of disinformation in the media have the potential to influence the general perception of misleading information comingfromUkrainian, Romanian, and Western media, but not Russian. On the other hand, news consumption significantly correlates only with citizens' perception about the Russian media as a source of false information, and only when in the form of news vehiculated in social media. Additionally, higher trust in mainstream media makes people more convinced that media inUkraine, Russia, and Romania provide accurate information most of the time, while trust in social media does not have a significant impact on public perceptions about disinformation sources. (for an overview of the OLS models see Appendix). We controlled for age, gender, education, and ideology. The study has its limitations, among which the non-probabilistic sample (inherent for online panels), as well as single item measurement of the dependent variables. Results could be usedt o address the negative consequences of disinformation on negative public perceptions about the media with regards totheUkraine war and beyond.

References

Dastgeer, S., & Thapaliya, R. (2023). Information and Disinformation about the Ukraine War on Social Media. ESSACHESS–Journal for Communication Studies, 16(1(31)).

Geissler, D., Bär, D., Pröllochs, N., & Feuerriegel, S. (2022). Russian propaganda on social media during the 2022 invasionof Ukraine. arXiv preprint arXiv:2211.04154.

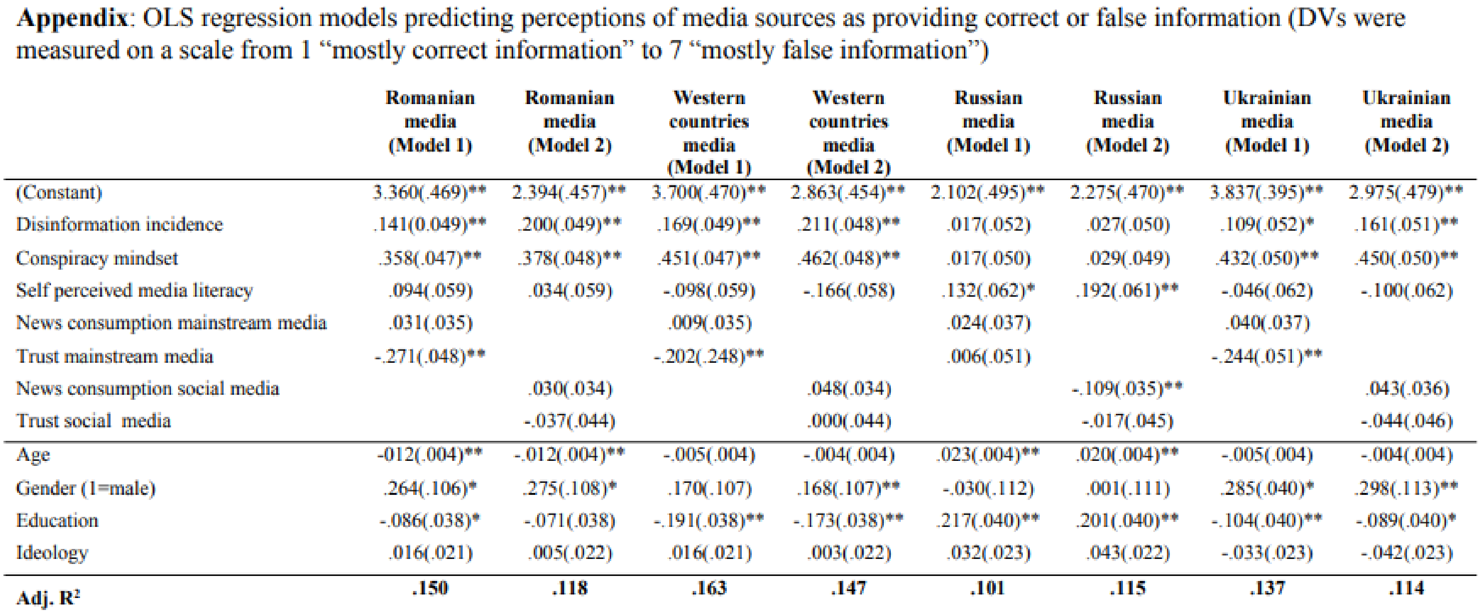
Grynko, A., & Baeriswyl, O. (2023). Digital Disinformation Campaign Around the War in Ukraine: Case of AlternativeMedia in Switzerland. ESSACHESS–Journal for Communication Studies, 16(1(31)), 183-202.

Kling, J., Toepfl, F., Thurman, N., & Fletcher, R. (2022). Mapping the website and mobile app audiences of Russia's foreign communication outlets, RT and Sputnik, across 21 countries. Harvard Kennedy School Misinformation Review.

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., & Nielsen, R. K. (2023). Reuters Institute digital news report 2023. Reuters Institute for the study of Journalism.

Vintilă, M., Lăzărescu, G. M., Kalaitzaki, A., Tudorel, O. I., & Goian, C. (2023). Fake news during the war in Ukraine: coping strategies and fear of war in the general population of Romania and in aid workers. Frontiers in psychology, 14, 1151794.

Williams, E. M., & Carley, K. M. (2023). Search engine manipulation to spread pro-Kremlin propaganda. Harvard Kennedy School Misinformation Review.



Appendix: OLS regression models predicting perceptions of media sources as providing correct or false information (DVs were measured on a scale from 1 "mostly correct information" to 7 "mostly false information")

	Romanian media (Model 1)	Romanian media (Model 2)	Western countries media (Model 1)	Western countries media (Model 2)	Russian media (Model 1)	Russian media (Model 2)	Ukrainian media (Model 1)	Ukrainian media (Model 2)
(Constant)	3.360(.469)**	2.394(.457)**	3.700(.470)**	2.863(.454)**	2.102(.495)**	2.275(.470)**	3.837(.395)**	2.975(.479)**
Disinformation incidence	.141(0.049)**	.200(.049)**	.169(.049)**	.211(.048)**	.017(.052)	.027(.050)	.109(.052)*	.161(.051)**
Conspiracy mindset	.358(.047)**	.378(.048)**	.451(.047)**	.462(.048)**	.017(.050)	.029(.049)	.432(.050)**	.450(.050)**
Self perceived media literacy	.094(.059)	.034(.059)	-.098(.059)	-.166(.058)	.132(.062)*	.192(.061)**	-.046(.062)	-.100(.062)
News consumption mainstream media	.031(.035)		.009(.035)		.024(.037)		.040(.037)	
Trust mainstream media	-.271(.048)**		-.202(.248)**		.006(.051)		-.244(.051)**	
News consumption social media		.030(.034)		.048(.034)		-.109(.035)**		.043(.036)
Trust social media		-.037(.044)		.000(.044)		-.017(.045)		-.044(.046)
Age	-012(.004)**	-.012(.004)**	-.005(.004)	-.004(.004)	.023(.004)**	.020(.004)**	-.005(.004)	-.004(.004)
Gender (1=male)	.264(.106)*	.275(.108)*	.170(.107)	.168(.107)**	-.030(.112)	.001(.111)	.285(.040)*	.298(.113)**
Education	-.086(.038)*	-.071(.038)	-.191(.038)**	-.173(.038)**	.217(.040)**	.201(.040)**	-.104(.040)**	-.089(.040)*
Ideology	.016(.021)	.005(.022)	.016(.021)	.003(.022)	.032(.023)	.043(.022)	-.033(.023)	-.042(.023)
Adj. R^2	**.150**	**.118**	**.163**	**.147**	**.101**	**.115**	**.137**	**.114**

Fighting the Health Misinformation Infodemic on Social Media: Can Digital Nudging Help?

Extended Abstract (Research in progress)

Muhammed Sadiq T, Saji K Mathew

Introduction: Misinformation spreads with ease and speed on social media platforms as it is easy to generate and disseminate information freely and instantly on social media (McAfee and Brynjolfsson 2017). Recently the surge of "Infodemic" has accompanied COVID-19 as a major threat to the world. World Health Organization (WHO) has defined "infodemic" as the information overload that occurs during a disease outbreak, which includes false and misleading content both offline and online (WHO 2021).

Although there is rich academic engagement on misinformation, current scholarly research on understanding the nature of the 'infodemic' and its cure is quite limited (Gu and Hong 2019; Li et al. 2019). our research seeks to contribute to the growing literature on health misinformation management by (i) *examining the psychological drivers of users engaging in disseminating fake health information* and (ii) a*ssessing the potential impacts of digital nudging in influencing the user's health information-sharing behavior*. We draw upon psychological ownership theory and digital nudging to develop an understanding of the motives behind spreading health-related misinformation and what kind of digital interventions could mitigate the spread of such misinformation.

Theoretical Foundations: We theorize that the need to share misinformation emerges from psychological ownership. "Psychological ownership is the state in which individual feels as though the target of the ownership or as a piece of target is 'theirs'" (Pierce et al. 2001). When a user shares a content on social media platform by investing energy, time and effort in it, that person develops a psychological ownership towards the platform and content (Karahanna et al. 2015; Pierce et al. 2001).

Nudging involves "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (Thaler and Sunstein 2008). Although digital nudge has been discussed in social media context, we observed that there is less research on digital nudge interventions to control or curtail the spread of misinformation on social media platforms, especially in an Asian context. Our study addresses this critical gap and analyzes digital nudging as an intervention strategy to control the spread of misinformation, while understanding Infodemic through the theoretical lens of psychological ownership motivation.

Proposed Study: This research adopts a mixed methods approach by fusing qualitative insights from exploratory focus-group discussions and quantitative insights from a randomized controlled lab experiment. As an exploratory phase, we initially conducted a questionnaire survey to understand the social media users' perception on health misinformation, especially to understand what kind of health misinformation users encountered during COVID-19. We have built an in house web platform to conduct our experiment. The Web platform will give the participants a virtual experience resembling the WhatsApp messaging platform. The participants are exposed to varied health information, including scientifically validated as well as fake news. The fake information was collected from the International Fact-Checking Network (IFCN) partner sites, and the fact information was extracted from mainstream media sources. We introduce specific digital nudging interventions and assess the influence of those interventions on participant's behavior to share information with their target group(s). Our experiment intends to use one control group and three treatment groups. Participants in three treatment groups are exposed to different digital nudges to understand the effects and impact of digital nudges on misinformation sharing. We seek to present our research ideas and preliminary findings at the MISDOOM'23 conference to obtain feedback and improve our research.

References:

Gu, R., and Hong, Y. K. 2019. "Addressing Health Misinformation Dissemination on Mobile Social Media," in 40th International Conference on Information Systems, ICIS 2019, November 12.

Karahanna, E., Xu, S. X., and Zhang, N. 2015. "Psychological Ownership Motivation and Use of Social Media," Journal of Marketing Theory and Practice (23:2), pp. 185–207. (https://doi.org/10.1080/10696679.2015.1002336).

Li, Y.-J., Cheung, C. M. K., Shen, X.-L., and Lee, M. K. O. 2019. Health Misinformation on Social Media: A Literature Review, Association for Information Systems.

McAfee, A., and Brynjolfsson, E. 2017. "Machine, Platform, Crowd : Harnessing Our Digital Future," W. W. Norton & Company.

Pierce, J. L., Kostova, T., and Dirks, K. T. 2001. "Toward a Theory of Psychological Ownership in Organizations," Academy of Management Review (26:2), pp. 298–310. (https://doi.org/10.5465/AMR.2001.4378028).

Thaler, R. H., and Sunstein, C. R. 2008. Nudge - Improving Decisions about Health, Wealth and Happiness, Yale University Press.

WHO. 2021. "Infodemic," Who.Int. (https://www.who.int/health-topics/infodemic, accessed February 21, 2022).

Understanding the Use of WhatsApp Groups as a Source of (Mis)Information: A user centric mixed method study in a polarized authoritarian context

Suncem Koçer, Ph.D.
Koç University
sukocer@ku.edu.tr

Ozen Bas
Kadir Has University, Ph.D.
ozen.bas@khas.edu.tr

In a highly polarized sociopolitical context with a conflict-ridden media landscape, as in Turkey, news users increasingly distrust political news. The recent rise of authoritarianism has enabled multiple legal and social surveillance mechanisms to the detriment of dissenting citizens, extending various forms of (self-)censorship (Çelik, 2020). Whatsapp offers a safe space for citizens to obtain trustable information (Koçer & Bozdağ, 2020:5303). While recent studies (Kuru et al., 2022) investigate the role of social dynamics and informational activity in understanding misinformation processing in instant messaging groups, these analyses do not consider contextual particularities. This study aims to comprehensively understand the role of social relationships (intimacy) in different types of WhatsApp groups for individuals of varying political affiliations in a polarized authoritarian country such as Turkey. We ask the following questions:

RQ1: Is a group's intimacy level, that is, if a group is based on weak ties vs. strong ties (Granovetter, 1973), associated with (a) the level of political information received from WhatsApp groups, (b) the level of trust users feel towards political information, and (c) the frequency of fact-checking political information received from WhatsApp groups?
RQ2: How do different WhatsApp groups serve as sources of political information for users with different political party affiliations in a highly polarized political context where freedom of expression and dissent online and offline is repressed?

We use mixed-methods research, including a nationally representative survey, media diaries, and semi-structured interviews with diary participants. User-centric, mixed-methodology studies help researchers overcome what Rossini (2023:4) calls data scarcity, highlighting the difficulties of pursuing false information on WhatsApp and other private communication platforms. Conducted in April-May 2023, the survey (N=951) determined the prevalence of WhatsApp usage and trust levels across various socio demographic groups that receive political (mis)information. We found that the type of WhatsApp group and political party affiliation are critical in understanding political (mis)information processing. First, the results suggest that the level of intimacy in the WhatsApp group tends to increase the frequency of receiving political information and the level of trust for the information.

Secondly, an unexpected finding emerged regarding the larger political context and misinformation processing practices of individuals. Those who voted for the two nationalistic parties (Iyi Party and MHP) used WhatsApp groups more as a source of information and trusted that information more, even though these two parties are on opposite sides of the Turkish political landscape. The main pillar parties of the opposing camps (AKP and CHP) tend to receive similar levels of political information from WhatsApp groups and feel more trust toward them than the nationalistic parties. We, therefore, employ qualitative techniques to examine the contingent role of immediate social relationships that circumvent the event of messaging (the type of WhatsApp groups) and the larger sociopolitical context (political party affiliations) on how users process political (mis)information received on WhatsApp. Currently, we are soliciting media diaries from 15 WhatsApp users. Following the one-week diary study, we will conduct semi-structured interviews with the participants to elucidate the examples they gave in their diaries and ask additional questions about how they used information throughout the course of the week via WhatsApp.

References

Çelik, B. (2020). Turkey's Communicative Authoritarianism. Global Media And Communication, 16(1), 102–120.

Granovetter, M. S. (2018). The Strength of Weak Ties. Social Stratification: Class, Race, and Gender in Sociological Perspective, 78(6), 653–657. https://doi.org/10.4324/9780429494642-79

Koçer, S., & Bozdağ, Ç. (2020). News Sharing Repertoires on Social Media in the Context of Networked Authoritarianism: The Case of Turkey". International Journal of Communication, 40.

Kuru, O., Campbell, S. W., Bayer, J. B., Baruh, L., & Ling, R. (2022). Reconsidering Misinformation in WhatsApp Groups: Informational and Social Predictors of Risk Perceptions and Corrections. International Journal of Communication.

Rossini, P. (2023). Farewell to Big Data? Studying Misinformation in Mobile Messaging Applications. In Political Communication (pp. 1–6). Routledge. https://doi.org/10.1080/10584609.2023.2193563

VaxTwita: an Annotated Corpus of Italian Tweets Related to Covid-19 Disinformation (Extended Abstract)

Marta Maggioni

The easy accessibility of information on the internet creates new challenges in terms of discerning truth from falsehood. False information spreads more rapidly and more broadly than reliable information (Vosughi et al, 2018), especially on social media platforms that often promote echo chambers (Li & Chang, 2021). The uncontrolled circulation of false information can lead to ill-informed decisions, as seen during the Covid-19 pandemic when fake news casted doubts on vaccines and health measures (Monselise et al. 2021).

To investigate the perspectives surrounding Covid-19 vaccines, a corpus of Italian Twitter messages was created as part of a broader research endeavor. On Twitter, users shared opinions and concerns about vaccines through news articles and personal viewpoints, but the tendency to prioritize personal beliefs over factual accuracy (Van Aelst et al. 2017) may lead to the spread of false information. The study analyzed how individuals represent their worldview when they share true or false information.

The initial dataset consisted of approximately 140,000 Italian tweets posted between February and July 2021, collected by the University of Turin (Basile & Caselli, 2020, Basile et al. 2018). Data collection used 19 keywords related to Covid-19 vaccines, which were searched either as a unique string or as a hashtag. The size of the dataset was subsequently reduced to a representative sample of 6,280 tweets to perform the manual annotation of the corpus.

VaxTwita represents the first Italian corpus of Twitter messages related to disinformation about Covid-19 vaccines. The corpus was manually annotated to differentiate messages containing false claims and messages containing true claims and it was created to be homogeneous and representative: homogeneous as it portrays
social media interactions posted by common Twitter users excluding news outlets and political accounts. Representative because it samples online discourse within a specific timeframe, reflecting the same proportion of hashtags used in that timeframe.

The researcher adapted and utilized the annotation technique developed by Alam et al. (2021) to perform the annotation of the 6,280 tweets in the VaxTwita corpus. At first, the annotation identified which messages were relevant to the topic of vaccines and which messages were not relevant to the topic. The latter group was excluded from further steps of annotation. Considering topic-relevant messages, the researcher determined if the text contained a verifiable factual claim. If so, the researcher evaluated the veracity of the claim by verifying the information it contained. It is important to clarify that this annotation technique should not be regarded as fact-checking, as it does not involve examining the accuracy of the information or the reliability of the source. However, the study did incorporate common fact-checking techniques as needed during the annotation process.

At the end of the process, 5,083 messages out of the total are topic-relevant. Of the topic-relevant tweets, 2,132 contain verifiable claims and 2,951 contain opinions (non-verifiable claims). Lastly, out of the 2,132 verifiable claims, 1,132 are annotated as True claims and 1,000 as False claims.
Conference Program Abstract

Social media users share news and opinions online, but the tendency to prioritize personal beliefs over factual accuracy often leads to the spread of disinformation. As part of a broader research project, we created the corpus VaxTwita to analyze perspectives on Covid-19 vaccines. VaxTwita is the first Italian corpus of disinformation, as it consists of 6,280 tweets differentiated between messages containing a verifiable factual claim or a personal opinion. Messages with verifiable factual claims were further annotated to identify whether the information in the claim was true or false. The researcher adapted and employed previous annotation techniques to: assess the relevance of messages to the topic of study, identify verifiable claims, and evaluate their veracity. It's important to note that this annotation process does not involve fact-checking but incorporates fact-checking techniques when necessary.

References

Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Da San Martino, G., Abdelali, A., Sajjad, H., Darwish, K., & Nakov, P. (2021). Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to Arms. Proceedings of the International AAAI Conference on Web and Social Media, 15(1), 913-922.

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Darwish, K., Al- Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., & Nakov, P. (2021). Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In M-F. Moens, X. Huang, L. Specia, & S. Wen-tau Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 611–649). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.findings-emnlp.56

Basile, V. Lai, M. Sanguinetti, M. (2018). Long-term Social Media Data Collection at the University of Turin. In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018. DBLP:conf/clic-it/2018, http://ceur-ws.org/Vol-2253/paper48.pdf, Mon, 17 Dec 2018 17:18:40 +0100, https://dblp.org/rec/bib/conf/clic-it/BasileLS18.

Basile, V. Caselli, T. (2020). 40twita 1.0: A collection of Italian Tweets during the COVID-19 Pandemic. http://twita.di.unito.it/dataset/40wita

Li, J. and Chang, X. (2022). Combating Misinformation by Sharing the Truth: a Study on the Spread of Fact-Checks on Social Media. Inf Syst Front. https://doi.org/10.1007/s10796-022-10296-z Monselise M, Chang C, Ferreira G, Yang R, Yang C (2021) Topics and Sentiments of Public Concerns Regarding COVID-19 Vaccines: Social Media Trend Analysis. J Med Internet Res 2021;23(10):e30765. DOI: 10.2196/30765

Van Aelst, P. Strömbäck, J. Aalberg, T. Esser, F. deVreese, C. Matthes, J. Hopmann, D. Salgado, S. Hubé, N. Stępińska, A. Papathanassopoulos, S. Berganza, R. Legnante, G. Reinemann, C. Sheafer, T. Stanyer, J. (2017) Political communication in a high-choice media environment: a challenge for democracy? In Annals of the International Communication Association, 41:1, 3-27, DOI:10.1080/23808985.2017.1288551

Vosoughi S, Roy D, Aral S. (2018) The spread of true and false news online. Science. Mar 9;359(6380):1146-1151. doi: 10.1126/science.aap9559. PMID: 29590045.

Analysis of visual disinformation during the France riots: Evolution patterns and emotional appeals

Kun He, Jiapan Guo

Abstract:

Dis/misinformation, including fake news (Shu, et al, 2017) and hate speech (Warner & Hirschberg, 2012), is prevalent on social media, particularly in the context of theCovid-19 pandemic and the Ukraine-Russia war. While extensive research has been conducted to detect and analyse textual dis/misinformation (Choraś, Michał, et al, 2021), the evolution and spread of visual dis/misinformation and hateful memes remain understudied. Visual dis/misinformation and hateful memes are more effective emotionally and psychologically impacting the public, exacerbating social gaps, and polarising communities (Hameleers, et al., 2020).

This study examines the phenomenon of visual disinformation during the France Riots, which were sparked by the death of a teenager in a police shooting in June 2023. The Proliferation of visual disinformation across various media platforms, including TikTok, Twitter, Douyin, and Weibo, has been observed during these riots. Specifically, thisresearch focuses on analyzing visual disinformation circulating on Douyin, a Chinesesocial media platform, with a specific emphasis on short videos. The objective of this study is to provide insights into the patterns of evolution, emotional appeals, and computational methods used to analyze and detect visual disinformation.

Employing a computational approach, our methodology encompasses several essential steps: video retrieval and preprocessing, object recognition for disinformation debunking, context-based disinformation detection, and sentiment analysis. Initially, we collected and examined a sample of 175 popular short videos related to the France Riots. Through A manual process, we debunked disinformation by cross-referencing the information presented in these videos with BBC news reports. Subsequently, a combination of natural language processing algorithms and machine learning models was employed for a comprehensive analysis of the videos. These analytical techniques enabled us to identify patterns and characteristics of visual disinformation within the short videos. Notably, computer vision techniques were utilized to detect manipulated content, deep fake videos, and misleading visual narratives. Moreover, sentiment analysis provided valuable insights into the emotional appeals and potential manipulation tactics employed in these videos. By considering the contextual information surrounding thevideos, we evaluated the intent and purpose of the disinformation campaigns."

This study lays the groundwork for future research on analyzing disinformation, particularly focusing on short videos shared on social media platforms. Firstly, it enhances our comprehension of visual disinformation by examining its dissemination, evolution patterns, and emotional appeals within the context of the France Riots. Secondly, by specifically studying Douyin short videos, a platform with a substantial user

base and influential presence in China, it contributes to our understanding of theworldwide spread of disinformation during conflicts. Additionally, our findings illuminate the complexities and obstacles involved in countering visual disinformation, underscoring the necessity for robust computational tools and strategies.

Keywords: visual disinformation, Douyin, computational methods, emotional appeal, France riots, sentiment analysis

“No idea…” – What do German economy and journalism elites know about digital disinformation characteristics? – Extended Abstract

Nils Vief, HAW Hamburg
Marcus Bösch, HAW Hamburg
Christian Stöcker, HAW Hamburg

Conference program abstract:

The spread of false information is a growing problem in societies governed by platforms that allow any person with internet access to create, share, and disseminate false information to a global audience. Besides more traditional strategies to distribute false information via images, text, and video, new ways have developed in recent times including multimodal memes, audio-visual deepfakes, and audio misinformation. We conducted guided interviews with 25 key actors from large corporations and media outlets in Germany in order to answer the research questions: How do people in charge identify online disinformation? What do they know about specific characteristics of disinformation including new ways of distributing fake information like audio misinformation or multimodal memes? Preliminary results indicate a lack of knowledge and awareness concerning characteristics of disinformation amongst these key players. We map out the need for optimized counter-measures, and suggest basic strategies to increase disinformation literacy.

Extended Abstract:

The spread of false information is a growing problem in societies governed by platforms that allow any person with internet access to create, share, and disseminate false information to a global audience (Jeangène Vilmer et al., 2018). Besides more traditional strategies to distribute false information via images, text, and video, new ways have developed in recent times including multimodal memes (Kiela et al., 2020) audio-visual deepfakes (Mittal et al., 2020) and audio misinformation (El-Masri, Riedl & Woolley, 2022). These specific characteristics of disinformation distribution are further polluting the information ecosystem (Wardle, 2020) and the content curation systems that the largest platforms employ (Stöcker, 2020) through a complex ecology of attention hijacking. Yet they are not well researched and often not well understood by practitioners that are exposed to these new forms of disinformation. Our research questions therefore are: How do German journalism and business elites identify online disinformation? What do they know about specific characteristics of disinformation including new ways of distributing fake information like audio misinformation or multimodal memes? We conducted guided interviews with 25 key actors with particular expertise on the
topic of disinformation from deliberately selected large corporations and media outlets from relevant sectors in Germany through structured interviews. The interviewees either work as heads of the respective communication and security departments and are responsible for dealing with desinformation that affect their organizations (business sector), or they do journalistic work on the subject of disinformation (media sector). The interviews are part of a joint research project “Hybrid” - "Real-time detection and tracking of hybrid disinformation campaigns in online media" funded by the German Federal Ministry of Education and Research. Preliminary results show that the interviewed professionals can be classified in different groups based on their knowledge concerning the usage of different characteristics of disinformation. Our observations indicate a lack of knowledge and awareness among German journalism and business elites concerning crucial characteristics of disinformation. We observed severe knowledge gaps and strong differences between the participants

from both fields. The majority of corporate professionals in our sample are not aware of recent developments in disinformation technologies and characteristics while the interviewees from the media sector can describe analytical detection strategies. Therefore we map out the need for optimized counter-measures, and suggest basic strategies to increase disinformation literacy (Sharon & Baram-Tsabari, 2020) to minimize damage being done by bad actors.

References:

Jeangène Vilmer, J.-B. (2018). "Information Manipulation: A Challenge for Our Democracies, report by the Policy Planning Staff (CAPS) of the Ministry for Europe and Foreign Affairs and the Institute for Strategic Research (IRSEM) of the Ministry for the Armed Forces" https://www.diplomatie.gouv.fr/IMG/pdf/information_manipulation_rvb_cle838736.pdf

Khan, M. L., & Idris, I. K. (2019). Recognise misinformation and verify before sharing: a reasoned action and information literacy perspective. Behaviour & Information Technology, 38(12), 1194-1212.

Kiela, D. et al. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in neural information processing systems, 33, 2611-2624.

El-Masri, A., Riedl, M. J., & Woolley, S. (2022). Audio misinformation on WhatsApp: A case study from Lebanon. Harvard Kennedy School (HKS) Misinformation Review.

Mittal, T. et al. (2020, October). Emotions don't lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM international conference on multimedia (pp. 2823-2832).

Richards, A. (2022). A pro-Russia propaganda campaign is using over 180 TikTok influencers to promote the invasion of Ukraine. Media Matters. https://www.mediamatters.org/tiktok/pro-russia-propaganda-campaign-using-over-180-tiktok-influencers-promote-invasion-ukraine

Sharon, A. J., & Baram-Tsabari, A. (2020). Can science literacy help individuals identify misinformation in everyday life?. Science Education, 104(5), 873-894.

Stöcker, C. (2020). How Facebook and Google Accidentally Created a Perfect Ecosystem for Targeted Disinformation. In: Grimme et al. (Eds.) Disinformation in Open Online Media (pp. 129-149).

Wardle, C. (2020, September 20). Understanding information disorder. First Draft. https://firstdraftnews.org/long-form-article/understanding-information-disorder/

Breaking Boundaries: Cross-Platform Analysis of User and Information Dynamics in the Case of German Climate Change Discussion (Extended Abstract)

Janina Pohl[**], Saïd Unger[*], Lucas Stampe [**], Johanna Klapproth[*], Svenja Boberg[*], Christian Grimme[**], Thorsten Quandt[*] first name.last name@uni-muenster.de

[*] Online Communication, University of Münster

[**] Computational Social Science and Systems Analysis, University of Münster

1 Abstract

Social media has become an essential part of our lives, offering new possibilities for information sharing, social connections, and participation in public debates. However, existing research has predominantly focused on analyzing user behavior within single platforms, neglecting the diffusion of information across multiple platforms. To address this gap, our study introduces a framework for cross-platform analysis, aiming to investigate user and information dynamics within the German climate change discourse across Twitter, Telegram, and YouTube. Through an examination of user dispersion and information propagation patterns, this study seeks to provide insights into the polarized nature of the discourse. Moreover, we contribute to a broader understanding of cross-platform dynamics and information diffusion, offering insights for the development of disinformation detection techniques, analysis of self-radicalized subgroups, and network economics.

Keywords: cross-platform · information diffusion · social media · climate change · Germany

2 Extended Abstract

Social media platforms are now an essential part of modern lives. They serve as means of private socializing and as tools for civic, educational, and political engagement and information sharing. Consequently, various platforms have emerged, each with distinct formats and focuses, such as Twitter, a micro-blogging platform, and YouTube, a video sharing platform (Kim et al., 2022). However, the majority of researchers have predominantly concentrated their research on analyzing user behavior and reactions within the confines of a single platform rather than adopting a more comprehensive approach for their research topic (Assenmacher et al., 2020; Cinelli et al., 2022).

Cross-platform analysis has gained especially significant attention in radicalization and disinformation-sharing research. Studies examine the interplay between radical, less moderated social media platforms and mainstream platforms regarding user and information dynamics or the interaction solely within mainstream platforms. Research indicates that niche platforms construct narratives and disseminate radical content, while larger social media platforms are employed to attract new followers and increase visibility by reintroducing radical content during new events (Krafft & Donovan, 2020; L. Ng et al., 2022; Nizzoli et al., 2020). On mainstream platforms like Twitter and Facebook, users demonstrate coordinated behavior by sharing narratives via external links to, for example, YouTube videos (Lukito, 2020; K. Ng et al., 2021; Wilson & Starbird, 2021; Zhang et al., 2023).

Our ongoing project aims to investigate the dynamics of user engagement and information dissemination across multiple platforms, focusing on the German climate change discourse on social media. Given the highly polarized nature of this topic in Germany, with radical activists and climate change deniers, we collected data from Twitter, Telegram, and Youtube throughout 2022. In future studies, we intend to expand our analysis to encompass additional mainstream platforms such as

Instagram, Facebook, and Reddit, as well as niche platforms including Mastodon, Rumble, Odysee, and others. Based on this data, we want to conduct a twofold analysis:

First, we want to examine the users on the platforms. While identifying prominent figures such as Greta Thunberg across platforms presents minimal challenges, identifying ordinary users proves more difficult. However, we aim to explore this feasibility through a flexible search pattern that accommodates variations in punctuation, numbers, and abbreviations within usernames. Additionally, we intend to establish connections between platforms by investigating if users share common links, texts, images, or videos across multiple accounts. By successfully identifying users across platforms, our investigation seeks to ascertain whether German climate change discussion participants are dispersed among different platforms or utilize multiple accounts on various platforms.

Second, our study entails analyzing the shared content across diverse platforms. We aim to identify prominent topics within a specific platform at any given moment by em ploying techniques such as topic modeling and stream clustering. This analysis will enable us to investigate the flow of information between platforms. By tracing the trajectory of information from its origin on one platform to its dissemination on others, we can gain insights into the dynamics of various social media platforms in our case study. Considering the polarized nature of the climate change discourse, this examination will shed light on the typical pathways through which different types of information propagate.

Ultimately, our work contributes to understanding the dynamics on and between various social media platforms, which can be used in disinformation detection, analyzing self-radicalization, and network economics.

References

Assenmacher, D., Clever, L., Pohl, J., Trautmann, H., & Grimme, C. (2020). A two-phase framework for detecting manipulation campaigns in social media. Proc. of the Int. Conf. on Human-Computer Interaction, 201–214.

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on twitter. Decision Support Sys tems, 160.

Kim, N., Duffy, A., Edson Tandoc, J., & Ling, R. (2022). All news is not the same: Di vergent effects of news platforms on civic and political participation. International Journal of Communication, 16.

Krafft, P. M., & Donovan, J. (2020). Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign. Political Communication, 37 (2), 194–214. https://doi.org/10.1080/10584609.2019.1686094

Lukito, J. (2020). Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017. Political Communication, 37 (2), 238–255. https://doi.org/10.1080/10584609. 2019.1661889

Ng, K., Horawalavithana, S., & Iamnitchi, A. (2021). Multi-platform Information Op erations: Twitter, Facebook and YouTube against the White Helmets. Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media. https://doi.org/10.36190/2021.36

Ng, L., Cruickshank, I., & Carley, K. M. (2022). Cross-platform information spread during the January 6th capitol riots. Social Network Analysis and Mining, 12. https://doi.org/10.1007/s13278-022-00937-1

Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M., & Ferrara, E. (2020). Charting the

Landscape of Online Cryptocurrency Manipulation. IEEE Access, 8, 113230–113245. https://doi.org/10.1109/ACCESS.2020.3003370

Wilson, T., & Starbird, K. (2021). Cross-platform Information Operations: Mobilizing Narratives & Building Resilience through both ’Big’ & ’Alt’ Tech. Proceedings of the ACM on Human-Computer Interaction, 5. https://doi.org/10.1145/3476086

Zhang, Y., Sharma, K., & Liu, Y. (2023). Capturing Cross-Platform Interaction for Iden tifying Coordinated Accounts of Misinformation Campaigns (J. Kamps, L. Goeu riot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, & A. Caputo, Eds.). Advances in Information Retrieval, 694–702.

Hijacking #Pride: How right-wing actors in Germany tried to piggyback on the pride-movement to spread patriotic and anti-queer narratives on TikTok (Extended Abstract)

Svenja Boberg*, Saïd Unger*, Marcus Bösch**, Johanna Klapproth*, Christian Stöcker**, Thorsten Quandt*

*Department of Online Communication, University of Muenster
**Department Information, HAW Hamburg

Abstract

This study examines the hijacking tactics and dynamics on TikTok through a case study of the #Stolzmonat campaign. Despite the platform's limited accessibility and searchability, a mixed-methods approach was used to collect and analyze 810 TikToks related to the campaign. The findings reveal a strong association between #Stolzmonat and the German right wing populist party AfD, as well as right-wing extremist accounts. Co

occurring hashtags indicate a government-opposing network with connections to other stirring issues like climate change. Qualitative analysis highlights the involvement of right-wing initiators, queer accounts as part of a significant counter-movement, amplifier accounts on both sites, and ties to Russian propaganda accounts and conspiracy communities. This case study serves as an initial exploration of hijacking on TikTok and suggests the potential for further research on the involvement of malicious communities in social media hypes.

Extended Abstract

New avenues of participation have emerged in the digital era, providing marginalized groups with opportunities for engagement and articulation. With its rapidly growing user numbers (Silberling, 2021), TikTok's content-centric approach tailored to producing and consuming videos and popcultural references introduces prospects for hashtag activism like #pridemonth but also provides a fertile ground for the spread of disinformation and conspiracy theories (Basch et al., 2021; Shang et al., 2021).

Viral content or social media hypes (Vasterman, 2005) like #pridemonth develop momentum, the hashtag has garnered 18,2 billion views according to TikTok, but also face backlash from emerging counter-movements. For example, right-wing actors in Germany have tried to piggyback on the pride-movement and spread patriotic and anti

queer narratives via the German translation #Stolzmonat on TikTok in June 2023. This practice of hijacking hashtags (Bradshaw, 2022; Jackson & Foucault Welles, 2015) is well-researched on Twitter with its core elements agenda surfing, reframing and derailment.

This study aims to investigate how the elements and dynamics of hijacking are employed on TikTok, using the example of the #Stolzmonat campaign. Accessibility to data and the searchability of the platform is challenging, so we use a mixed-methods design to identify relevant content and actors. As a first step a collection of 810 TikToks related to the #Stolzmonat search query was gathered, including account information, number of likes and comments, video descriptions, and timestamps. A combination of automated content analysis, network analysis and time-based analysis was conducted to identify related topics and communities and to trace the dynamics of the hashtag

campaign. Additionally, qualitative content analysis was performed on the ten most active accounts and viral TikToks to get a deeper understanding of the used features and actors involved.

The automated content analysis revealed a strong association between the #Stolzmonat campaign and the German right-wing populist political party AfD and right-wing extremist accounts. The network analysis of co-occurring hashtags also showed a government-opposing network with spillover into other topics such as Covid, energy crisis, and climate change. The campaign peaked in the beginning of June, but remained at a constant level over the course of the month. However, a significant queer countermovement within the countermovement was also observed, which could develop a far larger amplitude, but then immediately faded. The qualitative content analysis highlighted the role of initiators, queer accounts, and amplifier accounts among the ten most active users. While queer content garnered high engagement, TikToks promoting #Stolzmonat featured peculiar slideshows with patriotic motifs, deviating from typical TikTok design elements and showing ties to Russian propaganda accounts and conspiracy communities.

The case study of #Stolzmonat exemplifies typical hijacking tactics and can be seen as a first attempt to capture macrostructures with specific domain knowledge on TikTok. We propose further iteration of this mixed-methods design in the sense of a snowballing approach to identify relevant networks of malicious communities and their involvement in social media hypes.

References

Basch, C. H., Meleo-Erwin, Z., Fera, J., Jaime, C., & Basch, C. E. (2021). A global pandemic in the time of viral memes: COVID-19 vaccine misinformation and disinformation on TikTok. Human Vaccines & Immunotherapeutics, 17(8), 2373–2377. https://doi.org/10.1080/21645515.2021.1894896

Bradshaw, A. S. (2022). #DoctorsSpeakUp: Exploration of Hashtag Hijacking by Anti Vaccine Advocates and the Influence of Scientific Counterpublics on Twitter. Health Communication, 1–11. https://doi.org/10.1080/10410236.2022.2058159

Jackson, S. J., & Foucault Welles, B. (2015). Hijacking #myNYPD: Social media dissent and networked counterpublics. Journal of Communication, 65(6), 932– 952. https://doi.org/10/f77hf2

Shang, L., Kou, Z., Zhang, Y., & Wang, D. (2021). A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. 2021 IEEE International Conference on Big Data (Big Data), 899–908. https://doi.org/10.1109/BigData52589.2021.9671928

Silberling, A. (2021, September 27). TikTok reached 1 billion monthly active users. TechCrunch. https://techcrunch.com/2021/09/27/tiktok-reached-1-billion monthly-active-users/

Vasterman, P. L. M. (2005). Media-Hype: Self-Reinforcing News Waves, Journalistic Standards and the Construction of Social Problems. European Journal of Communication, 20(4), 508–530. https://doi.org/10.1177/0267323105058254

Writing Style Affects Partisanship and Persuasiveness Ratings (Extended Abstract)

Allison Nguyen, Assistant Professor of Psychology, Illinois State University
Tom Roberts, Assistant Professor of Computational Linguistics, Utrecht University
Pranav Anand, Professor of Linguistics, UC Santa Cruz
Jean Fox Tree, Professor of Psychology, UC Santa Cruz

Hyperpartisan communication, that which is "… openly ideological, extremely biased … and attacks the other side's point of view, often at the expense of facts" (Rae 2021, p. 1118), presents significant social and political challenges. Hyperpartisan communication is frequently misleading and subjectively framed, as opposed to fabricated outright (Pennycook & Rand, 2019) and is more common online than explicit fake news (Faris et al. 2017; Rojecki & Meraz, 2016). Because hyperpartisan language is endemic to online spaces, it is important to understand how it differs from nonhyperpartisan language, and the effect it has on language users, particularly when it comes to spreading and sharing information.

In previous corpus studies, we (Nguyen, et al., 2022) demonstrated that in comparison to neutral language, hyperpartisan language in internet forums makes key use of communicative elements indicative of spontaneity. In this study, we investigate whether text with these linguistic features is perceived as hyperpartisan, and whether this has correlates with partisanship and persuasiveness. Spontaneous linguistic devices foster a sense of closeness among conversational participants (Rubin & Greene, 1992), which is important because this sense of closeness can create the conditions ripe for sharing misinformation. For example, the first and second person pronouns *I* and *you* highlight the co-construction of the communication. Swear words (e.g. *hell, fuck*) and discourse markers (e.g. *you know, like*) occur more frequently in conversations with close interlocutors, like friends, than in socially distant contexts, like professional settings (Dewaele, 2016; Fox Tree, 2007; Wang et al., 2014). In Reddit forums, the use of first and second person pronouns, swear words, and discourse markers were all found more often in hyperpartisan communication than in nonhyperpartisan communication (Nguyen et al., 2022). Writing-specific affordances, such as punctuation, can also be leveraged to foster conversational closeness. For instance, exclamation points were also more common with closer addressees (Rubin & Greene, 1992); accordingly, exclamation points were more common in hyperpartisan communication (Nguyen et al., 2022).

We leverage these findings in two experiments testing how swear words, questions (information-seeking and rhetorical), and writing-specific features (e.g. ??? or !!! and ellipses) affected ratings of hyperpartisanship and persuasiveness. Participants read naturally occurring political text with and without these elements and rated them for partisanship and persuasiveness (Experiment 1). We then tested how the addition or subtraction of elements affected ratings of the same statements in a within-items design (Experiment 2). We predict that elements emblematic of spontaneous communication will be interpreted as hyperpartisan, and that communication with these elements will be rated as more persuasive. We also predict that the presence of hyperpartisan language will result in higher ratings of partisanship compared to the same stimuli with the hyperpartisan elements removed.

The role of spontaneous communication elements in creating persuasive language has not yet been tested. The current work is important for media literacy training – being able to show people what to be aware of might help them inoculate themselves. Understanding what can create hyperpartisan spaces can help people recognize and avoid falling into them, and open up a dialogue about hyperpartisanship.

Selected References

Dewaele, J. M. (2017). Self-reported frequency of swearing in English: Do situational, psychological and sociobiographical variables have similar effects on first and foreign language users?. Journal of Multilingual and Multicultural Development, 38(4), 330-345.

Faris, Robert M., Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. (2017). Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election. Berkman Klein Center for Internet & Society Research Paper.

Fox Tree, J. E. (2007). Folk notions of um and uh, you know, and like. Text & Talk, 27–3, 297–31. Fox Tree, J. E. (2015). Discourse markers in writing. Discourse Studies, 17(1), 64–82.

Fox Tree, J. E., Mayer, S. A., & Betts, T. E. (2011). Grounding in instant messaging. Journal of Educational Computing Research, 45(4) 455-475.

Nguyen, A., Roberts, T., Anand, P., & Fox Tree, J. E. (2022). Look, dude: How hyperpartisan and non hyperpartisan speech differ in online commentary. Discourse & Society, 33(3), 371-390.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences USA, 116(7), 2521-2526. http://dx.doi.org/10.1073/pnas.1806781116.

Rae, M. (2021). Hyperpartisan news: Rethinking the media for populist politics. New Media & Society, 23(5), 1117-1132.

Rojecki, A. & Meraz, S. (2016). Rumors and factitious informational blends: The role of the web in speculative politics. New Media & Society, 18(1), 25-43, https://doi.org/10.1177/1461444814535724

Rubin, D. L., & Greene, K. (1992). Gender-typical style in written language. Research in the Teaching of English, 7-40.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014, February). Cursing in english on twitter. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (pp. 415-425).

Resisting Interventions: An agent based model of the effect of tie-dissolution on the diffusion of disinformation and prebunking interventions (Extended Abstract)

Saïd Unger[*], Johanna Klapproth[*], Janina Pohl[**],
Svenja Boberg[*], Christian Grimme[**], Thorsten Quandt[*] first name.last name@uni-muenster.de
[*] Online Communication, University of Münster
[**] Computational Social Science and Systems Analysis, University of Münster

1 Abstract

Disinformation campaigns on social media significantly threaten democratic societies by undermining trust and influencing public debates. Scholars have focused on studying countermeasures, like prophylactic prebunking interventions and post-exposure debunking, to contain the spread of manipulative content online. The formation of echo chambers and confirmation bias contributes to selective information consumption, but concerns about their influence may be overstated. In our study, we implement an agent-based model to explore the impact of users dissolving existing ties within their social network in cases of non-compliance, thus mimicking the creation of isolated communities with a homogeneous opinion. Preliminary results indicate that excluding dissenting voices reduces disinformation diffusion and isolates intervention-resistant subgroups. Future research will explore additional tie dissolution scenarios to expand the understanding of the dynamics of disinformation diffusion in online networks.

Keywords: information diffusion; social media; agent-based modeling; social network analysis; disinformation

2 Extended Abstract

The strategic spread of disinformation campaigns on social media platforms to achieve political goals present a significant threat to democratic societies. Disinformation, described as misleading or false content to cause harm (Wardle, 2017), can influence public debates and threaten to destabilize societies by undermining trust in political, social, and media systems (Bennett & Livingston, 2018).

Due to the harmful effects of disinformation campaigns, scholars intensively research the effectiveness of countermeasures to contain the spread of manipulative content in online environments. Existing research suggests that interventions can be distinguished into