2nd Multidisciplinary International Symposium on Disinformation in Open Online Media (MISDOOM)

Extended Abstracts

October 26–27, 2020

https://2020.misdoom.org

Contents

Echo Chambers? Filter Bubbles? A Critical Review

Extended Abstract

The related concepts of "echo chambers" (Sunstein, 2001a) and "filter bubbles" (Pariser, 2011) are widely used in the scholarly as well as popular literature to refer to a form of communicative dysfunction, especially in online search and social media, where participants become trapped in "information cocoons" (Zuiderveen Borgesius *et al*., 2016: 1) that expose them only to highly one-sided, politically partisan information and thus prevent them from maintaining a balanced information diet. Such dysfunction is said to result especially as users' personal communicative choices are channelled and restricted by the recommendation algorithms of search engines and social media platforms. Proponents of this perspective envisage severe consequences for the further functioning of democratic systems, as such information cocoons prevent citizens from basing their contributions to public debate and their electoral choices on sufficiently comprehensive information (Sunstein, 2001b). Echo chamber and filter bubble effects are therefore also seen as instrumental in the success of the 2016 Brexit referendum and the election of Donald Trump as US President.

However, a series of recent studies have challenged this perspective. Several studies of the search engine results that are provided by *Google Search* as well as *Google News*, in the US and Germany, for a variety of political topics, have demonstrated independently from each other that there is very little evidence for a comprehensive personalisation of search results (Haim *et al*., 2018; Krafft *et al*., 2018; Nechushtai & Lewis, 2019): in spite of their highly divergent personal interest profiles and ideological positions, different users are generally directed to the same news outlets and articles, to the point that one of these studies even criticises the lack of diversity in the search results encountered and calls for *greater* algorithmic personalisation (Nechushtai & Lewis, 2019: 302). If there is an information cocoon in search, then, it is a national or even global cocoon that encapsulates the full political spectrum and produces similar results for users from all walks of life – but such an all-encompassing cocoon cannot then produce the democratic dysfunction that the proponents of the filter bubble idea envisage; rather, this consistency in search engine results maintains a unified public sphere.

Similarly, evidence for filter bubbles in social media remains limited and unconvincing. Several studies point to the existence of homophily and clustering in social media, as users preferentially engage with others who share similar interests and views (e.g. Vaccari *et al*., 2016; Smith & Graham, 2019). But their methodological choices – for instance, a focus only on selected Twitter hashtags or Facebook pages, or a limitation to studying only the most active participants – often prevent them from determining whether the participants in such clusters are indeed insulated from counterattitudinal perspectives, or whether they also continue to encounter information that disagrees with them. Meanwhile, studies using different approaches find that social media support *both* "echo chambers" and "open forums" (Williams *et al*., 2015); that the interest-based clusters that exist in the Twittersphere nonetheless remain thoroughly interconnected (Bruns *et al*. 2017); and that social media users are even deeply frustrated with the volume of politically divergent content they are exposed to (Duggan & Smith, 2016). If so, social media echo chambers remain highly porous and permeable, and therefore cannot produce the severely deleterious effects that are commonly attributed to them.

This paper conducts a critical review of this evidence, and also points to the fact that the moral panic associated with echo chambers and filter bubbles is in fact incompatible with concerns about other contemporary communicative dysfunctions. For instance, the mis- and disinformation campaigns described by the problematic label "fake news" (cf. Jack, 2017) fundamentally exploit the absence of information cocoons: as they disseminate their content, they build on the substantial *interconnectedness* of diverse clusters and interest groups in social media, and the reach of their campaigns would be considerably more limited if users did indeed exist only in hermetically sealed echo chambers. Ultimately, the paper therefore concludes that the concern about echo chambers and filter bubbles is misplaced, and prevents us from addressing a far more critical challenge: the growing polarisation of political communication, especially online. Importantly, such polarisation is itself intensified by the fact that the *absence* of echo chambers and filter bubbles enables opposing political partisans to monitor and attack each other's perspectives on a continuous basis (cf. Garrett *et al*., 2013).

References

Bruns, A., Moon, B., Münch, F., & Sadkowsky, T. (2017). The Australian Twittersphere in 2016: Mapping the Follower/Followee Network. *Social Media + Society*, *3*(4), 1–15. https://doi.org/10.1177/2056305117748162

Duggan, M., & Smith, A. (2016). *The Political Environment on Social Media*. Retrieved from Pew Research Center Website: http://assets.pewresearch.org/wp-content/uploads/sites/14/2016/10/24160747/PI_2016.10.25_Politics-and-Social-Media_FINAL.pdf

Garrett, R. K., Carnahan, D., & Lynch, E. K. (2013). A Turn toward Avoidance? Selective Exposure to Online Political Information, 2004–2008. *Political Behavior*, *35*(1), 113–134. https://doi.org/10.1007/s11109-011-9185-6

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the Filter Bubble? Effects of Personalization on the Diversity of Google News. *Digital Journalism*, *6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Jack, C. (2017). *Lexicon of Lies: Terms for Problematic Information*. Retrieved from Data & Society Research Institute Website: https://datasociety.net/pubs/oh/DataAndSociety_LexiconofLies.pdf

Krafft, T. D., Gamer, M., & Zweig, K. A. (2018). *Wer sieht was? Personalisierung, Regionalisierung und die Frage nach der Filterblase in Googles Suchmaschine*. Retrieved from Algorithm Watch Website: https://www.blm.de/files/pdf2/bericht-datenspende---wer-sieht-was-auf-google.pdf

Nechushtai, E., & Lewis, S. C. (2019). What Kind of News Gatekeepers Do We Want Machines to Be? Filter Bubbles, Fragmentation, and the Normative Dimensions of Algorithmic Recommendations. *Computers in Human Behavior*, (90), 298–307. https://doi.org/10.1016/j.chb.2018.07.043

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. London: Penguin.

Smith, N., & Graham, T. (2019). Mapping the Anti-Vaccination Movement on Facebook. *Information, Communication & Society*, *22*(9), 1310–1327. https://doi.org/10.1080/1369118X.2017.1418406

Sunstein, C. R. (2001a). *Echo Chambers: Bush v. Gore, Impeachment, and Beyond*. Princeton, N.J.: Princeton University Press.

Sunstein, C. R. (2001b). *Republic.com*. Princeton, N.J.: Princeton University Press.

Vaccari, C., Valeriani, A., Barberà, P., Jost, J. T., Nagler, J., & Tucker, J. A. (2016). Of Echo Chambers and Contrarian Clubs: Exposure to Political Disagreement among German and Italian Users of Twitter. *Social Media + Society*, *2*(3), 1–24. https://doi.org/10.1177/2056305116664221

Williams, H. T. P., McMurray, J. R., Kurz, T., & Lambert, F. H. (2015). Network Analysis Reveals Open Forums and Echo Chambers in Social Media Discussions of Climate Change. *Global Environmental Change*, *32*, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Topic Analysis for Misinformation on Medical Internet Fora

Eline Honig*, Jitske van der Vlugt*, Anne Dirkson[0000−0002−4332−0296], and Suzan Verberne[0000−0002−9609−9505]

Leiden Institute of Advanced Computer Science, 2333CA Leiden, the Netherlands

Medical misinformation on the web can have grave consequences for public health if citizens base their medical decisions on these faulty claims. One well-reported example is the growing trend not to vaccinate [7] [10] which may be partially caused by the online spread of misinformation about the risks of these vaccines [8]. Similarly, cardiologists have reported cases of patients refusing statin medication based on misinformation about its risks and thereby increasing their chances of a heart attack [4].

Internet forums are especially conducive to the dissemination of misinformation. Communities tend to form around users that are similar and this leads to a high level of implicit trust in the information spread by other members [2]. In order to effectively combat the spread of misinformation on medical forums, it is essential to understand about which medical topics the most misinformation is being spread.

To answer this question, we have investigated a subset of MedHelp[1], a platform for health-related discussions. The subset of 1558 comments was selected by [5] based on keywords known to be associated with medical misinformation and manually labelled for misinformative content.

We modelled the topics discussed in this data set using Non-Negative Matrix Factorization (NMF) [6]. The amount of topics was determined using the topic coherence, measured with TC-W2V [9]. Topic labels were assigned manually by exploring the words with the highest weights and the top-ranked (i.e. most relevant) messages per topic. Each message was assigned to the topic for which it has the highest score, unless no score was above 0.03 in which case it was assigned no topic.

We found that these medical forums mainly discussed five topics: breast cancer, the link between autism and vaccines, allergies for mint, filth in belly buttons, and anxiety. Of these, breast cancer and the link between autism and vaccines contained the most misinformation, namely 28.9% and 27.5% misinformative content respectively (see Fig. 1).

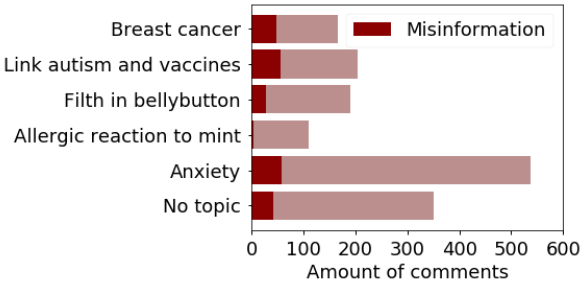


Fig. 1: Amount of comments per topic

Misinformation about breast cancer mostly concerned alternative treatment or unproven causes of cancer such as certain shampoo or deodorant ingredients. Moreover, particularly for this topic, many messages appear to be spread by bots.

Misinformation about vaccines and autism was mainly about the high risks of vaccination and often referred to books or famous people, such as the book *What your doctor may not tell you about Children's vaccinations* [1] or the actress Jenny McCarthy [3].

Although further investigation is necessary, this work provides a first indication of the medical topics about which misinformation is being spread online. This knowledge may inform public health departments what to focus on in order to more effectively combat medical misinformation.

* These authors contributed equally to this paper. This work was done in the context of the Pre-University program of Leiden University.

[1] www.medhelp.org

References

1. Cave, S., Mitchell, D.R.: What your doctor may not tell you about children's vaccinations. Grand Central Publishing, New York City, USA (2001)
2. De Choudhury, M., Sundaram, H., John, A., Seligmann, D.D., Kelliher, A.: "Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media? ArXiv (jun 2010), http://arxiv.org/abs/1006.1702
3. Einbinder, N.: How former 'The View' host Jenny McCarthy became the face of the anti-vaxx movement. Business Insider (2019), https://www.businessinsider.nl/jenny-mccarthy-became-the-face-of-the-anti-vaxx-movement-2019-4/
4. Hill, J.A.: Medical misinformation: vet the message! PACE - Pacing and Clinical Electrophysiology **42**(3), 299–300 (2019). https://doi.org/10.1111/pace.13616
5. Kinsora, A., Barron, K., Mei, Q., Vydiswaran, V.G.: Creating a Labeled Dataset for Medical Misinformation in Health Forums. In: Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017. pp. 456–461. Institute of Electrical and Electronics Engineers Inc. (2017). https://doi.org/10.1109/ICHI.2017.93
6. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999). https://doi.org/10.1038/44565
7. van Lier, E., Oomen, P., Giesbers, H., van Vliet, J., Drijfhout, I., Zonnenbert-Hoff, I., de Melker, H.: Vaccinatiegraad en jaarverslag Rijksvaccinatieprogramma Nederland 2018 — RIVM. Tech. rep., Rijksinstituut voor Volksgezondheid en Milieu RIVM (2018), https://www.rivm.nl/publicaties/vaccinatiegraad-en-jaarverslag-rijksvaccinatieprogramma-nederland-2018
8. Myers, M.G., Pineda, D.: Misinformation about Vaccines. In: Stanberry, L.R., Barrett, A.D.T. (eds.) Vaccines for Biodefense and Emerging and Neglected Diseases, chap. 17, pp. 255–270. Academic Press (2009)
9. O'Callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications **42**(13), 5645–5657 (2015). https://doi.org/10.1016/j.eswa.2015.02.055
10. Patel, M., Lee, A.D., Clemmons, N.S., Redd, S.B., Poser, S., Blog, D., Zucker, J.R., Leung, J., Link-Gelles, R., Pham, H., Arciuolo, R.J., Rausch-Phung, E., Bankamp, B., Rota, P.A., Weinbaum, C.M., Gastañaduy, P.A.: National Update on Measles Cases and Outbreaks — United States, January 1–October 1, 2019. MMWR. Morbidity and Mortality Weekly Report **68**(40), 893–896 (2019). https://doi.org/10.15585/mmwr.mm6840e2, http://www.cdc.gov/mmwr/volumes/68/wr/mm6840e2.htm?s_cid=mm6840e2_w

GRACE GAMBIZA: **Masters Student, Communication Studies Department, University of Johannesburg, Johannesburg, South Africa**

TOPIC: Fake News and Social Media Regulation in Zimbabwe: A Case Study Of the January 2019 National #Shutdown.

ABSTRACT

The question of so-called “fake news” and social media regulation has become a central concern for governments worldwide, the private sector, media regulators and – gradually – media scholars. The ubiquitous presence of “fake news” on social media, on a national and global scale, has provoked a flurry of government-sponsored and private sector sponsored regulations to deal with it. For instance, Facebook CEO, Mark Zuckerberg, recently proposed that governments should help Facebook to regulate social media and weed out fake news. This surprising plea suggests that social media companies may be losing the battle against “fake news”. Most governments worldwide have not needed prompting.The internet blackout of January 2019 in Zimbabwe, which forms the backbone of this study, caps off a string of internet disruptions on the continent. On 21 December 2018, the Sudanese government blocked internet access to popular social media sites in an attempt to quell nationwide protests triggered by economic instability and price hikes. Gabon experienced an internet shutdown on 7 January 2019 in the wake of an attempted military coup. A few days later, the Democratic Republic of Congo (DRC) saw widespread disruption of internet connectivity following the 30 December 2018 elections. Togo, Sierra Leone, Cameroon and Chad are among other African countries that faced internet substantive restrictions in 2018 alone. But in all this busy “regulatory” activity, three questions stand out for media scholars. The first one concerns matters of definition. What really is fake news? Despite the very public discourse about “fake news”, there is still no accepted criterion of defining fake news or an industry standard for noticing and recognising fake news. There is as yet no standard or universally agreed definition, amongst media scholars of the concept. The second question is about power. Who decides what should be regulated, and how? The third and last question is about digital platforms regulation. Whereas scholarship on digital regulation (in the traditional sense of new media) is widely available, scholarship on social media regulation is only in its nascent stages signalling a dearth of studies systematically engaging the issue of social media regulation. Owing to the fact that fake news (and social media itself) are recent innovations, and because the controversies are coming thick and fast, media scholars who deal with issues of media regulation, media freedom and freedom of expression have not yet fully woken up to the implications of the clamour to regulate social media. Is social media regulation, to weed out fake news, a good thing or a bad thing? This leads to an ancillary question. What does the clamour to regulate social media reveal, if anything, about the would-be regulators and about the nature of fake news? This study is an attempt to grapple with these questions, with an emphasis on the nexus of “fake news” and social media regulation.This exploratory study is qualitative. It utilises an interpretive approach to analyse thematic issues raised from purposively selected key informants from government, civil society, media-policy making circles in Zimbabwe. Semi- structured interviews will be conducted with the key informant’s to explore the nexus of fake news and social media regulation in Zimbabwe using the January 2019 National # Shutdown as a backdrop to understand exactly how “fake news” figures in the media regulation matrix.

What's in a name: Defining "fake news" from the audience's perspective

Nicoleta Corbu, National University of Political Studies and Public Administration, Romania; nicoleta.corbu@comunicare.ro

Divina Frau-Meigs, Savoir Devenir, France;

divina.meigs@orange.fr

Denis Teyssou, Agence France-Press, France;

Denis.TEYSSOU@afp.com

Alina Bârgăoanu, National University of Political Studies and Public Administration, Romania; alina.bargaoanu@comunicare.ro

The last decade has dramatically changed not only the media landscape, but also the news consumption patterns, due to the many technological developments and automation practices. Additionally, in the wake of the Cambridge Analytica scandal, the online disinformation became a focal point of concern for both academics and policy makers, and "fake news" a buzzword indiscriminately used in many social contexts. Consequently, an entire academic literature trying to define and classify the many facets of the so-called fake news phenomenon has flourished. Most of these papers propose a normative perspective, which is mainly focused on establishing boundaries between concepts (such as mis-, dis-, and mal-information), defining genres, establishing criteria for inclusion and exclusion into the large umbrella of the term, proposing alternative labels to clarify the field of investigation, etc. However, little is still known about what ordinary people understand by "fake news", even though the term is broadly used, to the point of becoming anecdotal.

In this context, we propose a comparative qualitative study conducted in four countries (France, Romania, Spain, Sweden), focusing on making sense of people's understanding of the term, as well as the perceived incidence and perceived effects of such news in various country specific contexts. By means of 8 focus groups with educated people (2 by country), which include highschool teachers as a specific group, we will gather insightful information about what could

be called the audience's perspective on fake news. This is ongoing research (one pilot focus group has been conducted in Romania) in the framework of the Youchek! Project, developed within the Preparatory Action on Media Literacy for All, who's goal is "to raise awareness of the public at large and the education and media literacy community about disinformation and *fake news* as a threat to democracy". Preliminary data show that people classify many concepts under the "fake news" label, such as "half-truth", "framing", "hyper-partisan news", "information taken out of the context", etc. They generally evaluate the phenomenon as quite worrisome and discuss the many negative effects they estimate this type of content have at individual and social levels.

A Large-Scale Linguistic Analysis of News and Fiction in Satirical News

Britta C. Brugman

Vrije Universiteit Amsterdam

Christian Burgers

Vrije Universiteit Amsterdam / University of Amsterdam

Camiel J. Beukeboom & Elly A. Konijn

Vrije Universiteit Amsterdam

Author note

Britta C. Brugman, Department of Communication Science, Vrije Universiteit Amsterdam (the Netherlands); Christian Burgers, Department of Communication Science, Vrije Universiteit Amsterdam (the Netherlands) / Amsterdam School of Communication Research (ASCoR), University of Amsterdam (the Netherlands); Camiel J. Beukeboom, Department of Communication Science, Vrije Universiteit Amsterdam (the Netherlands); Elly A. Konijn, Department of Communication Science, Vrije Universiteit Amsterdam (the Netherlands).

This work is part of the research program *Contemporary Political Satire: Medium, Language, and Impact of Satiric News* with project number 276-45-005, which is financed by the Dutch Research Council (NWO).

Correspondence concerning this article should be addressed to Britta C. Brugman, Department of Communication Science, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands. E-mail: b.c.brugman@vu.nl.

A Large-Scale Linguistic Analysis of News and Fiction in Satirical News

In the past two decades, satirical news (e.g., *The Onion, Borowitz Report*) has become an increasingly popular source of learning about current affairs (Becker & Bode, 2018). At the same time, satirical news can be seen as a special form of fake news. While it is not created with the intention to deceive the public, it does contain deliberately fabricated content (Egelhofer & Lecheler, 2019). This tension between real and fake in satirical news is often linked to *discursive integration*, which refers to the idea that regular news and fiction have integrated in satirical news to such an extent that both genres have become inseparable (Baym, 2005). This study adds to the fake news literature by examining how exactly regular news and fiction are discursively integrated in online written satirical news.

Because studying linguistic register has been shown to be a valid and reliable way of identifying genres in collections of texts, we analyzed linguistic register through computer-automated textual analysis. We first collected 96,280 articles published in the calendar year 2018 from 16 satirical-news websites, 19 regular-news website, and one online fiction archive, which resulted in a corpus that consisted of 65,620,540 words. We next identified four register dimensions that typified the included genres: (1) involved vs. informational discourse, (2) reported speech discourse, (3) predictive discourse, and (4) precise discourse.

Our analyses revealed that satirical news only differed from the regular news and fiction categories in the dimension of involved vs. informational discourse. That is, fiction was characterized by more linguistic features that signaled involved discourse (e.g., *first and second person pronouns, present tense verbs*) than regular news, and satirical news scored in between both genres. This study therefore suggests that discursive integration in satirical news is only clearly reflected in one register dimension. Future research may further investigate whether taking into account the register dimension of involved vs. informational discourse improves our ability to identity whether news is satirical.

References

Baym, G. (2005). The Daily Show: Discursive integration and the reinvention of political journalism. *Political Communication, 22*, 259-276. https://doi.org/10.1080/10584600591006492

Becker, A. B., & Bode, L. (2018). Satire as a source for learning? The differential impact of news versus satire exposure on net neutrality knowledge gain. *Information, Communication & Society, 21*, 612-625. https://doi.org/10.1080/1369118X.2017.1301517

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: a framework and research agenda. *Annals of the International Communication Association, 43*, 97-116. https://doi.org/10.1080%2f23808985.2019.1602782

Human detection of social bots

Thao Ngo, Magdalena Wischnewski, & Rebecca Bernemann

Research Training Group "User-Centered Social Media", University of Duisburg-Essen, Germany

With the emergence of social networking sites (SNS) users can easily generate and share content online with a wider audience. However, SNS are increasingly populated by automated user account, commonly identified as social bots, which are designed to mimic human behavior online (Ross et al., 2019). While social bots can distribute useful information (e.g., weather forecast), they have previously also been used to manipulate online. For instance, concerns have been raised about critical bot engagement in the U.K.'s Brexit referendum in 2016 (Bastos & Mercea, 2017), the U.S. Presidential election in 2016 (Bessi & Ferrara, 2016) as well as the 2017 French general election (Ferrara, 2017). Hence, social bots can interfere, to some extent, public opinion formation and thus, might endanger the integrity of political elections (Bessi & Ferrara, 2017). The detectability of social bots in online communication rises complex problems: While some are easily identifiable, others cannot be detected by the mere eye. Multiple automated solutions to detect social bot activity have been developed (e.g., Santia, Mujib, & Williams, 2019). However, given that it is difficult to verify the accuracy of the automated solutions, the actual efficacy of these solutions remains ambiguous.

Apart from these technical challenges, we note that comparatively little attention has been paid to the users' understanding of social bots and their strategies to detect them. Therefore, this work in progress aims to (1) to investigate what users know about social bots, to what extent they are aware of their existence, and, ultimately, which strategies they apply to detect them. Because prior knowledge on this is limited, we intend to answer these questions through a qualitative approach. Furthermore, we want to (2) gain first insights into how well users detect social bots online. We specifically focus on the domain of political online communication, employing the psychological theory of motivated reasoning which argues that people pursue different goals when drawing inferences, that are either accuracy-goals or directional goals (Kunda, 1990). We hypothesize that bot detection would be biased in order to support political attitudes, resulting in overconfidence and "blindness" towards bots that promote favored content, and skepticism and rejection towards bots promoting opposing views. To test this, we would like to conduct a quasi-experimental study, varying the two factors 1) opinion/stance and 2) "botability".

CHARACTERISTICS OF DISINFORMATION CAMPAIGNS ON TWITTER DURING THE BRAZILIAN 2018 PRESIDENTIAL ELECTION

The 2018 Brazilian presidential election happened amidst several controversies, especially surrounding Jair Bolsonaro, the representative from the Social Liberal Party (PSL). Filling his campaign with far-right views and polemic declarations about minorities and opponents, the candidate defeated the leftist Fernando Haddad from the Worker's Party (PT). Bolsonaro's campaign heavy use of social media was also connected to the spread of disinformation by his supporters (Machado et al., 2018). In this context, our proposal focuses on discussing part of the results of a two years research that started in 2018 about disinformation in political conversations on Twitter during the 2018 presidential campaign in Brazil. Our goal is to present a case study and offer some insights about disinformation campaigns in Latin America, which is currently an understudied context.

Disinformation has been discussed by the literature as the content that is created to deceive (Derakshan & Wardle, 2017 amongst others). A disinformation campaign is, roughly, the coordinated spread of disinformation as a means to an end, to influence the public opinion through social media. Disinformation campaigns are strongly connected to political propaganda, sometimes used as tools to promote political views (Bastos & Mercea, 2019). These campaigns often rely on trolls and botnets (Ong & Cabañes, 2018), political influencers and activists (Soares, Recuero & Zago, 2018), hyperpartisan outlets (Marwick & Lewis, 2017), and other strategies to coordinate and legitimate the spread of biased and manipulated content.

Based on a dataset of 10 million tweets collected through Social Feed Manager (Prom, 2016) through several keywords, we tried, through multiple approaches, to answer the research question: *What are the key characteristics of the disinformation campaigns aimed to influence the Brazilian 2018 election through political conversations on Twitter?* For this research question, we aligned our results on three aspects of the characteristics of the disinformation campaign: (a) content strategies; (b)legitimation strategies and (c) spread strategies. We selected the most viral disinformation campaigns from the original dataset and further worked with multiple methods. For the content and legitimation strategies, we worked with discourse analysis (Fairclough, 2001; Van Leeuwen & Wodak, 1999) and content analysis (Krippendorf, 2014); for the spread strategies we worked on with social network analysis (Wasserman & Faust, 1994) and content analysis.

Our main results are:

- **Content Strategies:** We found that the majority of the most successful disinformation campaigns were from the right-wing (which increased in time), associated with hyperpartisan outlets and they rely partially on truthful content rather than entirely fabricated information. These campaigns often attacked mainstream media using hyperpartisan outlets as alternative real news. These strategies combined may be used to create an environment of mistrust and help the spread of disinformation. We also found that the majority of this type of content focused on the demonization of the Worker's Party and the left, as well as the electoral democratic system. This helped set a polarized context for the conversations, which seemed to be key for disinformation campaigns' success (Soares, Recuero & Zago, 2019).
- **Legitimation Strategies:** Disinformation campaigns often used links to hyperpartisan websites and retweets from authorities as to the main source of legitimation. They also used framming their stories as a duality between good and evil (mythopoesis, according to Van Leeuwen and Wodak, 1999). These strategies connected to polarization seemed to increase disinformation circulation and decrease traditional news circulation.
- **Spread Strategies:** The majority of disinformation campaigns in our case study would to start in small botnets that retweet/mention each other (Recuero & Gruzd, 2019). They often also mention authorities and other users as a "phishing strategy" to gain visibility. However, it seems that legitimation by the authority was key for virality (Soares, Recuero & Zago, 2018 and Recuero, Soares & Zago, 2019). The content was also frequently framed as "urgent" or "bombastic" to encourage users to retweet it. Finally, we found that the polarization and the circulation of hyperpartisan information also plays a key role, isolating clusters around the same political position creating a good environment for disinformation campaigns, as they offered "alternatives" to the "manipulation" of traditional outlets.

References:

Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and User-Generated Hyperpartisan News. Social Science Computer Review, 37(1), 38–54. https://doi.org/10.1177/0894439317734157

Derakhshan, H. Wardle, C. (2017). Information Disorder: Definitions. In *Proceedings of Understanding and Addressing the Disinformation Ecosystem*. Annemberg: University of Pennsylvania, 5-12.

Fairclough, N. (2001). *Discurso e mudança social*. Brasília: Editora UnB.

Krippendorf, K. 2014. *Content Analysis. An Introduction to Its Methodology* (3rd ed). California, CA Sage Publications.

Machado, C. Kira, B. Hirsch, G. Marchal, N. Kollanyi, B. Howard, P.. Lederer, T. Barash, V. (2018). News and Political Information Consumption in Brazil: Mapping the First Round of the 2018 Brazilian Presidential Election on Twitter. *Data Memo* 2018.4. Oxford, UK: Project on Computational Propaganda.

Marwick, A. & Lewis, R. (2017). Media, Manipulation and Disinformation Online. Available at: https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

Ong, J.C., & Cabanes, J.V. (2018). Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines. doi: 10.7275/2cq4-5396

Prom, C. (2016). *Social Feed Manager*. George Washington University Libraries. Zenodo. https://doi.org/10.5281/zenodo.597278.

Recuero, R. & Gruzd, A. 2019. Cascatas de Fake News Políticas: um estudo de caso no Twitter. *Galáxia* (São Paulo) n.41, pp.31-47. Epub May 23. Doi: 10.1590/1982-25542019239035.

Recuero, R, Zago, G, Soares, F. (2019) Using Social Network Analysis and Social Capital to Identify User Roles on Polarized Political Conversations on Twitter. SOCIAL MEDIA + SOCIETY, v. 1, p. 1-20, 2019.

Soares, F. B. Recuero, R. & Zago, G. 2018. Influencers in Polarized Political Networks on Twitter. In *Proceedings of the International Conference on Social Media & Society*, Copenhagen, Denmark (SMSociety). doi: 10.1145/3217804.3217909.

Soares, F. B., Recuero, R. Zago, G. 2019. Asymmetric Polarization on Twitter and the 2018 Brazilian Presidential Elections. In *Proceedings of the 10th International Conference on Social Media & Society*, Toronto, Canada (SMSociety). doi: 10.1145/3328529.3328546.

Van Leeuwen, T. Wodak, R. (1999). Legitimizing Immigration Control: A Discourse-Historical Analysis. *Discourse Studies*, 1(1), 83–118. DOI: 10.1177/1461445699001001005.

Wasserman, S., and Faust, K. 1994. Social Network Analysis. Cambridge*: Cambridge University Press*

Deepfakes in Brazil and the role of digital culture

Our proposal is an ongoing work in process which we aim to discuss the concepts of deepfake in communication and media studies. The majority of the literature on deepfake are very new - 2018/2019 - and they are from fields such as Law (Chesney & Citron, 2018), Computer Sciences and Information Technologies (Yang et al, 2019). These discussions focus on the relations between deepfakes and ethics or methods of detection of deepfake. Our proposal is to expand the discussion to the field of digital media and communication as an important category for understanding mediated processes and practices of digital culture users. Besides the theoretical discussion of deepfake concept we´ll also analyze deepfakes produced in brazilian context mainly done by Brunno Sartori, a media producer who became famous in brazilian internet due to his deepfake videos that combine pop culture, brazilian audiovisual - such as telenovelas genre - and politics. From what we've mapped so far, we can discuss that these specific deepfake can be addressed through four (4) kinds of approaches that maybe could result in a categorization of types:

1) *Audiovisual Culture* -The characteristics identified in the deepfakes produced in Brazil refer to the need to understand these phenomena contextualized in a technocultural environment (Fischer, 2013) in which technical images of different formats emerge that operate by convergence and dispersion (Kilpp, 2012) on social platforms of different scopes (from YouTube to WhatsApp), understanding software as "cultural" insofar as it operates as a layer that permeates all areas of contemporary society (Manovich, 2011). The images are in a constant process of remixing, both in the overlapping of references (a politician's face in a pop song video) and in the use of media authoring software as a cultural transcoding characteristic of the so called new media (Manovich, 2001). The audiovisual images of deepfakes, thus, demonstrate the coalescence of audiovisual techniques, media aesthetics and senses, insofar as they operate on the frontier of humor and political criticism.

2) *Brazilian Digital Cultural Practices* - The relations between memes, virals and deepfakes in the context of brazilian digital culture - This processes and practices of creating and sharing deepfakes are very close to the ones that have been studied in brazilian meme studies such as shown in studies done by Chagas (2018) and Vieira (2019) among others . These memes contents present elements of humour, pop culture and relations between brazilian political context, and the videos are sometimes discussed and appropriated by activists, fan activists (Amaral et al 2015) and others. In this approach, deepfake is discussed in a historic and cultural perspective that shows that those videos are the continuity of a long tradition of fake personas that can be traced before the internet forums (Donath, 1999) but also have specific characteristics that differ from Twitter fakes for instance (Amaral & Santos, 2012). Besides that, Sartori recently became an influencer on deepfake making and has appeared in TV Shows and other media. His celebrization process has outgrown the internet field and has brought the discussion of his role, not only in the past and the next elections, but in the political scene as a whole due to his criticism of the far-right tendencies that have been affecting Brazil and many other countries.

3) *Ecossystem of cyberevents and news/information flows* - Deepfakes are also part of a complex ecossystem of cyberevents (HENN, 2014), fake news and also are part of information and misinformation flows and circulations that include alternative media channels, digital platforms and the blurred lines between journalism and entertainment. Understanding cyber-event as the event whose procedurality already contains the weaving of digital networks, this concept has in the idea of the inaugural singularity the event as the driving force of semiosis. This semiosis has the vigor of the multiplicity of meanings that are produced by the encoding and framing of journalism. Thus, cyber-event is understood as a semiosis of altered or explosive flow, depending on online communication. Studies of this aspect analyze cases that circulate in the networks and in the press, trying to understand how they are meant by journalism and its consumers.

4) Digital labor behind deepfakess - The value of deepfakes is realized based on their circulation value, including meanings (Silverstone, 2002), on the different digital platforms, considered as means of production and means of communication (Williams, 2011). Deepfakes circulate with digital labor of clickworkers, often on microwork platforms such as Amazon Mechanical Turk. According to Ong and

Cabañes (2019), there are labor arrangements behind troll accounts and deep fakes production. This is especially true in the Philippines, known to be a center of ghost work behind artificial intelligence and deepfakes (Roberts 2019, Gray & Suri 2019). It is a crowdwork that does the labor of circulating deepfakes. The Brazilian case of Brunno Sartori reveals another face of the digital labor behind deepfakes. One of the gifts is the creation of a video with deepfake at the consumer's choice. Sartori also offers to participate in a group on Telegram with the promise of interactions and content production. This helps us understand platforms as means of production and communication and digital labor behind deepfake as crowdwork.

Case Study of Kristiansand Quran Burning: A Cross-Platform Analysis of Spill-Over Effects

Anna-Katharina Jung[1], Jennifer Fromm[1], Kari Anne Røysland[2], Gautam Kishore Shahi[1], Kim Henrik Gronert[3]
[1]University of Duisburg-Essen, Germany, [2]University of Agder, Norway, [3]Municipality of Kristiansand, Norway

The appearance and rise of social media platforms restructured and diversified the process of information diffusion. While priorly the dissemination of information was limited to traditional media outlets managed by gatekeeping journalists, nowadays information can be produced and shared by everyone with online access. Oftentimes, this leads to the emergence of different frames of one and the same incident. In addition, social media posts might be picked up and cited in traditional press coverage. In communication science, this phenomenon is described as the *spill-over effect* (Mathes and Pfetsch, 1991). The case of the recent Quran burning in Kristiansand demonstrates the relevance of further research about the diffusion of media frames throughout different types of online media platforms. On 16.11.2019, Lars Thorsen - the leader of the Norwegian organization *Stop the Islamification of Norway* (SIAN) - attempted to burn the Quran on the main square of the city of Kristiansand. Several persons who physically attacked Lars Thorsen to stop him from burning the Quran were arrested by the police. About 300 people witnessed the incident and shortly afterwards videos of the Quran burning and the attacks circulated on the net. Furthermore, the incident was heavily discussed in both newspapers and social media. While many members of the Muslim community described the attackers as *defenders of the Islam*, other actors rather created anti-Islamist narratives. These different media frames resulted in tensions within the Norwegian society and considerable problems for the Norwegian government. Although the media landscape is diverse, previous research on information diffusion often focused only on one specific social media platform or traditional news outlets (Jung et al., 2018).With our research-in-progress, we aim to advance the research on information diffusion by analyzing cross-platform spill-over effects of media frames. Hence, we aim to answer the following research questions:

RQ1: To what extent do media frames spill-over from social media (Twitter, YouTube) to traditional media (newspapers) and vice versa?

RQ2: How do media frames develop or change throughout the spill-over process?

To analyze the information diffusion about the Quran burning incident in Kristiansand across different types of online media, we collected 2.324 Norwegian tweets, 110 Norwegian/English YouTube videos, and 114 Norwegian online newspaper articles. We collected the data from 12.11.2019 – 30.11.2019 using the following keywords: *SIAN*, *Arne Tumyr* (member of SIAN), *Lars Thorsen* (leader of SIAN), and *koranbrenning* (Quran burning). In the first step of the analysis, we identified reference spill-overs in each dataset. A tweet including a link to an online newspaper article, for example, would represent a spill-over from an online newspaper to Twitter. In a second step, we will manually analyze the media frames of all tweets, videos, and articles including a reference spill-over using the coding scheme developed by Card et al. (2015). This way, we will be able to examine whether media frames change when they spill-over to a different platform. In a third step, we aim to conduct a social network analysis to visualize the media frame spill-overs between the different types of online media. Our research will deliver a new methodological approach to analyze cross-platform information diffusion and will deepen our understanding of frame spill-overs. Our research will provide important insights for authorities to effectively counteract fact-distorting frames that can lead to tensions in society.

Card D, Boydstun AE, Gross JH, et al. (2015) The media frames corpus: Annotations of frames across issues. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference* 2: 438–444. DOI: 10.3115/v1/p15-2072.

Jung AK, Ross B, Neuberger C, et al. (2018) Information diffusion between twitter and online media. *International Conference on Information Systems 2018, ICIS 2018* (November).

Mathes R and Pfetsch B (1991) The Role of the Alternative Press in the Agenda-Building Process: Spill-over Effects and Media Opinion Leadership. *European Journal of Communication* 6(1): 33–62. DOI: 10.1177/0267323191006001003.

Information Spread by Search Engines vs. Word-of-Mouth

Alon Sela[1,2,3], Shlomo Havlin[2], Louis Shekhtman[2], Irad Ben-Gal[3]

1. Department of Industrial Engineering, Ariel University, Israel, alonse@ariel.ac.il
2. Department of Physics, Bar-Ilan-University, Israel
3. Department of Industrial Engineering, Tel Aviv University, Israel

The spread of information in society has a significant political, social, and economic impact. The following work [1] compares two fundamental methods of modern information spread: (1) word-of-mouth (WOM), where opinions spread through social connections and (2) spread through web pages and search engines (WEB), where opinions are published on the internet and are then read by others who use search engines to find fitting results to their queries.

In both methods, there are multiple opinions available and the user, after considering a limited number of opinions (due to having limited time) eventually choses a single one. In both methods, choice of an opinion is also based on the opinion of others; that is, there exists a "social influence" effect, such that the probability that an agent will adopt an opinion is proportional to the number of times the agent has been exposed to any specific opinion.

Simulations of these two different mechanisms predict that the opinions in a large population will be less diverse when a population solely relies on WEB to search for information, compared to WOM. These results can be seen in Figure 1 (top), which shows the final opinion states in a population following temporal spread. In WOM, more opinions (colors) co-exist together. The simulations of the WEB spread dynamics include both the principle of the PageRank algorithm together with the observed users' tendency to click on different results returned from the Google search engine, also known as the SERP (Search Engine Result Probability) function. The simulative predictions were confirmed by an experimental work in which two groups of users were asked to answer similar questions. The first group was requested to answer the questions by asking their social circle (WOM), while the second was asked to search the questions on Google (WEB). The experiment confirmed that populations that search information only by WOM have a more diverse set of opinions as compared to populations solely using WEB methods.

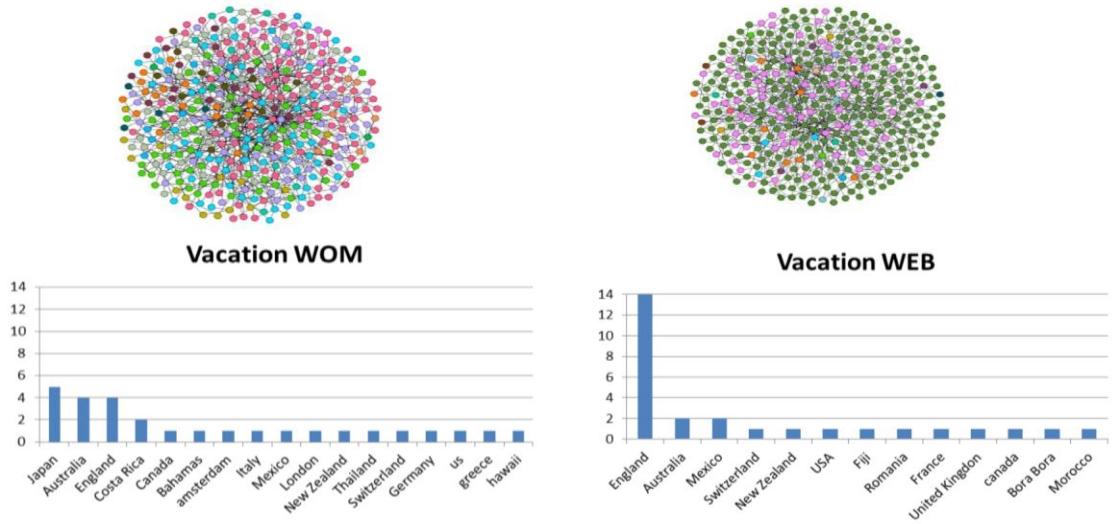


Figure 1: Diversity of information spread by word of mouth (WOM) vs. search engines (WEB). Upper figures, different colors represent different opinions. Lower, an example from the empirical study – "name the best place to go on vacation."

[1] Sela, Alon, Louis Shekhtman, Shlomo Havlin, and Irad Ben-Gal. "Comparing the diversity of information by word-of-mouth vs. web spread." *EPL (Europhysics Letters)* 114, no. 5 (2016): 58003

Information Spread – Intensive vs On-Going Campaigns

Alon Sela[1,2,3], Goldenberg Dmitri[3], Erez Shmueli[3], Irad Ben-Gal[3]

1. Department of Industrial Engineering, Ariel University, Israel, alonse@ariel.ac.il
2. Department of Physics, Bar-Ilan-University, Israel
3. Department of Industrial Engineering, Tel Aviv University, Israel

The spread of information is an important topic with major implications to politics, social and commercial sciences. The following work introduces the topic of *Scheduling Seeding* [1-4] and its main studied aspects as well as its application.

One can split the spread of political, commercial or ideological messages according to four types of spreading mechanisms. The first type is the use of mass media. Messages in this type are broadcasted to large populations, usually aiming to a defined or broad segment, with similar socio-demographic or political attributes. In this type there is no personal customization.

The second type is Social Media spread. These campaigns tend to be more "personalized" and can spread messages (true or fake) directly and solely to selected influential users. Selected users are carefully chosen from the social network to harness these opinion leaders to promote the agenda in a focused and personalized segment. The simplest measure of influence is the number of friends (or followers) that any possible node has. In more advanced methods of influencers detection, the PageRank or the Eigenvalue centrality measures can be a good proxy for influence.

The *Scheduling Seeding* methods advances beyond the simple definition of influential users' detection. It assumes that some users can be highly influential at one period, but less at another. An example of this effect can be seen in an election day. In the American political system, some states are "sure" republicans while others are "sure" democrat. The real fight is on the debating voices, that did not yet decide who to vote for. As the election day progresses, different users can become more important. For example, a user located at the center of a groups of voters that are still debating between the candidates is more important than a user located within a group of users that have already voted. The following talk presents the main aspects of the Scheduling Seeding approach and will try to inspect its relevance to different disinformation scenario including bots' campaigns, political struggles and commercial campaigns. We will present the main ideas behind the algorithmic method that were used to select influential users at the exact point of time. The scheduling seeding methods assume limited resource (time or money). These resources need to be used to reach a larger fraction of adoption, for any given budget.

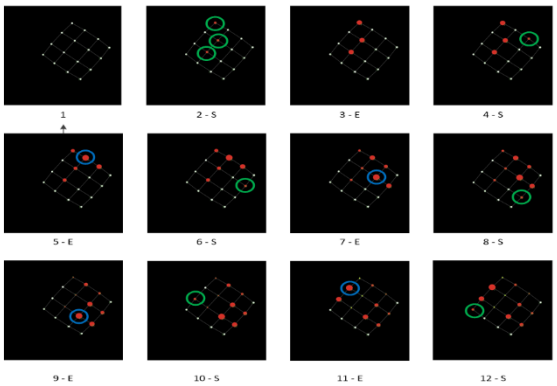


Figure 1: Seeding and Evolution steps in a 4 by 4 mesh. Threshold = 3; Oblivion = 5. Steps 1-12 represent the changes in which S represent a seeding of a node (green circle), and E represents a node that changes due to natural evolution (blue circle). The frame number represent the time of occurrence.

[1] Sela, Alon, Irad Ben-Gal, Alex Sandy Pentland, and Erez Shmueli. "Improving information spread through a scheduled seeding approach." In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 629-632. 2015.

[2] Goldenberg, Dmitri, Alon Sela, and Erez Shmueli. "Timing matters: Influence maximization in social networks through scheduled seeding." IEEE Transactions on Computational Social Systems 5, no. 3 (2018): 621-638.

[3] Sela, Alon, Dmitri Goldenberg, Irad Ben-Gal, and Erez Shmueli. "Active viral marketing: Incorporating continuous active seeding efforts into the diffusion model." Expert Systems with Applications 107 (2018): 45-60.

[4] Sela, Alon, Dmitri Goldenberg, Erez Shmueli, and Irad Ben-Gal. "Scheduled seeding for latent viral marketing." In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 642-643. IEEE, 2016

Monetizing Disinformation in the Attention Economy:
the case of genetically modified organisms (GMOs)
***European Management Journal* Special Issue "The Dark Side of Social Media"**
Authors: Camille D. Ryan[1], Andrew Schaul, Ryan Butner, and John Swarthout

Camille D. Ryan
Monsanto Company
700 Chesterfield Parkway West, Chesterfield, MO, USA 63017 camille.d.ryan@monsanto.com

Andrew Schaul
Bayer Crop Science
800 N. Lindbergh Blvd, St. Louis, MO, 63122 andrew.schaul@bayer.com

Ryan Butner
Pacific Northwest National Laboratory
902 Battelle Blvd, Richland, WA 99354 ryan.butner@pnnl.gov

John Swarthout
Bayer Crop Science
700 Chesterfield Parkway West, Chesterfield, MO, USA 63017 john.swarthout@bayer.com

Abstract:
The ubiquity of social media has created both opportunities and challenges for businesses and societies. For product brands, ideas, or campaigns to gain traction on social media platforms, they need to capture attention. This is often accomplished by creating and disseminating compelling information, even disinformation, on these platforms. Strategies that drive this attention economy are often not obvious. The monetization of disinformation is explored here through a case study on genetically modified organisms (GMOs) and the analysis of a dataset of 94,993 unique online articles. When combined these methods allow for the evaluation and exploration of various tactics that contribute to the evolving GMO narrative and their potential application to other topics. Preliminary results suggest that a small group of alternative health and pro-conspiracy sites received more totals engagements on social media than sites commonly regarded as media outlets on the topic of GMOs. Other externalities observed include continued social and political controversy that surround the GMO topic as well as the growth of additional product and marketing approaches such as "non-GMO" verification.

Keywords: social media, disinformation, attention economy, GMOs, genetic engineering

This presentation will build on previous work published in a special issue of *European Management Journal* on "The Dark Side of Social Media" in December of 2019 (available open access here). It represents the next in a series of scholarly pieces that we will tackle on the issue of disinformation and how it impacts business and societies. The methodology, analytics, and data form the foundation for this presentation and forthcoming articles that will explore regulatory and opportunity costs associated with disinformation as well as less visible public health and socioeconomic costs that come with negative public opinion, shelving of innovations, and increased regulatory barriers. Disinformation has implications across all sectors and societies.

[1] Corresponding author

Infowars-activity on Twitter: Exploring gatewatching, shareworthiness and social bots

Magdalena Wischnewski[1], Axel Bruns[2], Tim Graham[2], Tobias Keller[2], Dan Angus[2], Eshan Dehghan[2], Brenda Moon[2]

[1] University of Duisburg-Essen | [2] Queensland University of Technology

This is a work in progress contribution which investigates the interplay of gatewatching, shareworthiness and automated communication (social bots) on Alex Jones' Alt Right alternative media-outlet Infowars. We developed the following overarching research questions:

1) From all the content that was published on Infowars in the determined timeframe which contributions were also shared on Twitter (gatewatching, shareworthiness)?
2) What were the characteristics of that content that made it into Twitter?
3) How was the content shared on Twitter concerning different communication strategies?
4) Who was sharing that content on Twitter? What insights can we generate from account descriptions? Can we meaningfully cluster accounts?
5) Can we make connections from what is shared (and what not) to who shared it? Can we find a confirmation bias in news stories that made it into Twitter?

To this end, we analyzed all original tweets[1] that were shared on Twitter which contained an URL to an Infowars site between September 23 – 30, 2019. This collection resulted in 8024 tweets from 1365 individual accounts.

To answer RQ 1, we compared the URLs in our tweet collection with URLs published on Infowars during that time which we retrieved from GDELT. GDELT is a global database that monitors the world's broadcasts, print media, and online news. To answer RQs 2 and 3, we manually coded the content of the news articles and tweets according to inductively and deductively derived categories. In order to gage insights on accounts themselves (RQ4), we scraped the profile descriptions of all accounts that published an original tweet in the determined timeframe. Additionally, we are planning to conduct a cluster analysis to classify different account categories. Concerning the likelihood of automated communication, we ran all account IDs through botometer in addition to manual sighting of accounts. Finally, to answer RQ5, we want to connect results from shared content to the respective account. In doing so, we hope to gain deeper insights into sharing behavior of partisan alternative information on social media.

[1] With original tweet we mean any message that is shared on Twitter that is not also a reply or a retweet (RT).

Language and Hate:

Mechanisms of Dangerous Speech in German Politicians Facebook-Communication

Leonie Heims[1], Carina Strauss[1], Marcel Hansek[1], Yu Chen Yang[1], Tim Schatto-Eckrodt[1][https://orcid.org/0000-0003-1658-4373], & Lena Frischlich[1][https://orcid.org/0000-0001-5039-5301]

[1]Westphalian Wilhelms-University Muenster, Germany

Abstract. Scholars have shown that communication before the outbreaks of mass violence is often characterized by a rise in fear-inducing and divisive discourses attacking religious or ethnic minorities. Fear-mongering and divisive discourses targeting minorities are also typical elements of far-right politicians' rhetoric. Hence, the current study examined whether *dangerous speech* as typically observed in the context of mass atrocities can also be found in the Facebook communication of German politicians from far-right as well as other political parties. Combining characteristics identified in communication and linguistic research and using a quantitative content analysis of politicians posts around a violently escalated protest in the German city of Chemnitz, we demonstrate that the far-right party "Alternative for Germany" (AFD) used more and linguistically distinct forms of dangerous speech in their Facebook communication. We discuss the results in terms of their implications for the measurement of dangerous speech and the violence incitement through current far right communicators.

Keywords: Dangerous speech, Political Communication, Linguistics, Communication, Far right parties, Social media, Quantitative content analysis.

2

Extended Abstract

Political communication is easy to find on Facebook. When the tone of public debates becomes harsher, the baselines of what is socially accepted can slowly shift without the critical audience noticing it. Concerns are high that the audience's hate or even violence towards a member of another group will increase if so-called dangerous speech is used in such public debates.

Dangerous speech refers to rises in fear-inducing and divisive rhetoric observed before outbreaks of mass violence [1]. Typical elements are dehumanization, guilt attribution, or threat construction [2]. Such violent political speech [3] is characterized by specific linguistic elements like metaphors, diminutional suffixes, or swearwords. The current study adds to this literature by combining the dangerous speech approach and specific linguistic means as indicators in order to describe noxoius political communication in Germany.

Using a violently escalated right-wing protest in the German city Chemnitz 2018 as use case, we conducted a content analyses of Facebook-textposts from German parties or members of German parties discussing the events. Here, we showed that far-right politicians (associated with the "Alternative für Deutschland", AfD) used significantly more dangerous speech than the other parties when debating Chemnitz. In addition, they used other linguistic means to transmit their noxious messages. Furthermore, single individual used more dangerous speech than official party pages. Overall, our results shed light on noxious online communication by far-right politicians and provide a new perspective onto the development and events of the 2018 Chemnitz protests.

References

1. Benesch, S. (2012). Dangerous speech: A proposal to prevent group violence. Whitepaper. Online available https://worldpolicy.org/wp-content/uploads/2016/01/Dangerous-Speech-Guidelines-Benesch-January-2012.pdf. Last access: 2020/01/31
2. Leader Maynard, J., Benesch, S., Benesch, S., & American University. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, *9*(3), 70–95. https://doi.org/10.5038/1911-9933.9.3.1317
3. Scharloth, J. (2018). Sprachliche Gewalt und soziale Ordnung: Metainvektive Debatten als Medium der Politik [Verbal violence and social order: Meta-invective debates as policial mean]. In F. Klinker, J. Scharloth, & J. Szczęk (Eds.), *Sprachliche Gewalt: Formen und Effekte von Pejorisierung, verbaler Aggression und Hassrede [Verbal violence: Forms and effects of pejoratives, verbal aggression and hate speech]* (pp. 7–28). J.B. Metzler. https://doi.org/10.1007/978-3-476-04543-0_1

Disinformation about climate change on Twitter

Victor CHOMEL[a,b], David CHAVALARIAS[a,b], Maziyar PANAHI[a]

[a] *Institut des Systèmes Complexes Paris Ile-de-France (ISC-PIF), CNRS, Paris, France*
[b] *Ecole des Hautes Etudes en Sciences Sociales (EHESS), Paris, France*

Keyword: Computational Social Science, Disinformation, Misinformation, Social Networks, Community Dynamics, Climate change, Information propagation

Abstract

Live social media debates increasingly accompany burning issues like environment. Which is why we have decided to analyse misinformation and disinformation about climate change on social networks. The first step was to disclosing the organization of climate-related communities. We worked on more than 30 million tweets from 5 million Twitter users, all on the topic of climate change. This allowed us to map the relative influence of communities and interactions among them. We wanted to show the dynamics of communities that initiate the climate change debate, whether it is on the climate change skeptics side or on the pro-climate side.

We discovered radically different structures in these two communities. These differences in structures imply different mechanisms for the propagation of information, and therefore disinformation. The climate change skeptics community represents only one-quarter of the accounts in our sample. However, it compensates for its lower importance with a core of very proactive accounts (about one hundred thousand tweets/retweets each). This core group communicates on all themes at once and really drives the skeptics dynamic. This profusion sometimes leads to contradictions in the messages as we will show in the presentation. Basing their argument on false science, they seek to equate the two communities to better blur the debate.

Conversely, among accounts convinced of the anthropogenic origin of global warming, even if some people are at the heart of the community, we do not observe a core capable of tweeting on all areas at the same time.

We managed to highlight the echo chambers in the different communities by analyzing the propagation of specifically targeted tweets. Defining echo chambers also allow us to understand how to escape from them and see what kind of information has managed to spread beyond a filter bubble.

In a second step, instead of looking at a fixed landscape, we sought to study its evolution over time. We wanted to highlight the recruitment strategies of the communities. Those convinced of the anthropogenic origin of global warming mainly take advantage of new events such as bushfires in Australia to raise awareness among new users. Not really activists, these new accounts are often attached to a particular topic. This gateway gradually leads them to other topics and engage them in the community.

The climate skeptic community has different rhetoric. By commenting on pro-climate publications, they gain visibility. However, it is mainly through themes out of the climate topic that they succeed in gaining followers. One of the recruitment levers is precisely to divide people on social inequalities. Through calling attention to the wealth of the Democratic Party leadership, they politicize the issue and polarize the debate.

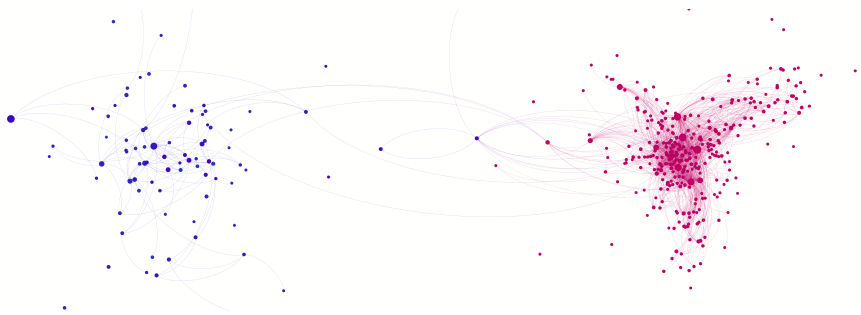


Figure 1: Retweet graph restricted to the accounts present from the beginning to the end of the study. The climatoskeptic accounts are in red and the proclimates in blue. As we can see, the climatoskeptic community is denser with a well-defined core.

Understanding the Evolution of State-Backed Disinformation Operations on Twitter through Network Analysis

State-backed disinformation is a potential threat to contemporary democracy (Collins et al., 2018). Such an all-encompassing phenomenon must be approached from a bird's eye view in order to understand its magnitude and structure. Temporal network analysis has the power to unlock the rich patterns and strategies at play in disinformation operations (Kriel & Pavliuc, 2019). My current research focuses on using temporal network analysis in order to understand strategies of disinformation operations.

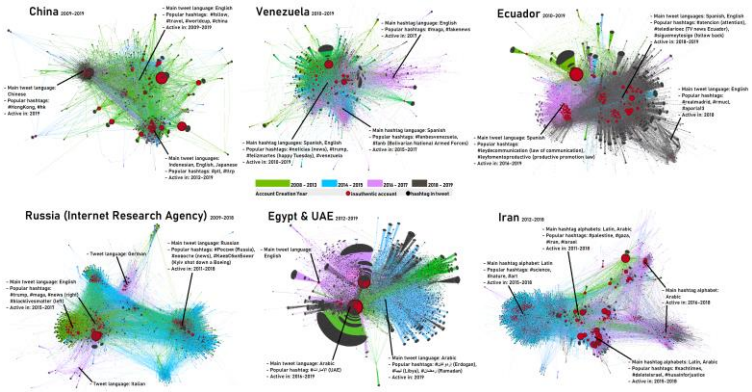


Figure 1 (Pavliuc, 2020)

Method

Twitter has released over 10GB of data on state-backed disinformation operations on their platform since 2018 (Gadde & Roth, 2018), and researchers have approached the datasets from different angles with different research questions (@DFRLab, 2018; Farkas & Bastos, 2018; Stewart et al., 2018). The most recent iteration of my research was funded by a Mozilla Foundation Open Source Support Award. It explores the similarities and differences between the evolving structures of six state-backed disinformation operations on Twitter, by visualizing each as a network, side by side. The static visualizations can be viewed in Figure 1, and the videos which showcase their evolution can be viewed in a Medium article (Pavliuc, 2020).

The state-backed disinformation operations I analyzed have been found by Twitter to have originated in Iran, China, Russia, Ecuador, Venezuela, and Egypt/UAE. These datasets were chosen due to their size, and noteworthiness in the news today. Each of these datasets underwent the same cleaning and preparation process, which included: extracting hashtags and metadata, visualizing inauthentic account-to-hashtag relationships in Gephi (a network visualization software (Bastian et al., 2009)), and unfurling the networks to play over time. The final step was crucial, as networks that do not account for the element of time can lead observers to false conclusions (Keim et al., 2008).

Findings and Future Research

All six Twitter datasets became active around the turn of the last decade, and have shifted languages, structures, and hashtags. Some countries focused on the languages of their own countries (Egypt/UAE, Ecuador), while the rest (Russian IRA, Venezuela, Iran, China) pushed beyond their national languages and began tweeting in other languages, such as English and/or Indonesian. Most datasets began with low amounts of tweeting, and graduated to deploying multiple bursts of hashtag use (when a large amount of hashtags are used at once for a period of time). Often, operations began by tweeting innocuous hashtags in order to build their presence, such as #followme,#felizmartes ('happy Tuesday'), #noticias ('attention'), or #news before graduating to more political hashtags in loud bursts of tweets such as #deleteisrael, #blacklivesmatter, #maga, #HongKong, and #إدراغودان ('Erdogan').Venezuela, Ecuador, and Iran deployed newer accounts in their hashtag bursts while Russia, China, and Egypt/UAE deployed old 'sleeper' inauthentic accounts (which may potentially have been purchased) alongside newer accounts. The countries that chose to create new accounts for their hashtag bursts may have been saving their older, more established, accounts for future opportunities in their disinformation operations.

Analyzing six disinformation operations equally as temporal networks and viewing them side by side allows for similarities and differences in patterns and strategies to emerge. Future research will focus on developing a taxonomy of disinformation network structures that can be used to categorize disinformation operations in the future. Additionally, future research will focus on designing simplified versions of the networks so they can be efficiently communicated to decision makers and disinformation practitioners, with the goal of increasing their knowledge on the evolution of state-backed disinformation strategies.

Bibliography

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the Third International ICWSM Conference*, 2.

Collins, D., Efford, C., Elliott, J., Farrelly, P., Hart, S., Knight, J., Lucas, I., O'Hara, B., Pow, R., Stevens, J., & Watling, G. (2018). Disinformation and "fake news." *House of Commons: Digital, Culture, Media and Sport Committee*, 89.

@DFRLab. (2018, October 17). #TrollTracker: Twitter's Troll Farm Archives. *DFRLab*. https://medium.com/dfrlab/trolltracker-twitters-troll-farm-archives-8be6dd793eb2

Farkas, J., & Bastos, M. (2018). IRA Propaganda on Twitter: Stoking Antagonism and Tweeting Local News. *Proceedings of the 9th International Conference on Social Media and Society - SMSociety '18*, 281–285. https://doi.org/10.1145/3217804.3217929

Gadde, V., & Roth, Y. (2018, October 17). *Enabling further research of information operations on Twitter*. https://blog.twitter.com/official/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization: Human-Centered Issues and Perspectives* (pp. 154–175). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7

Kriel, C., & Pavliuc, A. (2019). Reverse Engineering Russian Internet Research Agency Tactics through Network Analysis. *NATO Defense Strategic Communications*, *6*, 30.

Pavliuc, A. (2020, January 29). *Watch six decade-long disinformation operations unfold in six minutes*. Medium. https://medium.com/swlh/watch-six-decade-long-disinformation-operations-unfold-in-six-minutes-5f69a7e75fb3

Stewart, L. G., Arif, A., & Starbird, K. (2018). Examining Trolls and Polarization with a Retweet Network. *MIS2*, 6.

TOWARDS AUTOMATIC DETECTION OF PROPAGANDA TECHNIQUES IN NEWS ARTICLES

Daria Sinitsyna
College of Arts and Sciences
Syracuse University
dasinits@syr.edu

Lu Xiao
Associate Professor
School of Information Studies
Syracuse University
lxiao04@syr.edu

Bo Zhang
College of Arts and Sciences
Syracuse University
bzhang49@syr.edu

Yimin Xiao
School of Information Studies
Syracuse University
yxiao39@syr.edu

Guoxing Yao
College of Engineering & Computer Science
Syracuse University
gyao02@syr.edu

February 3, 2020

Nowadays, the place of the online media in the world is crucial. Not only is it the fastest way of providing the general public with information, but it also is one of the most persuasive tools in politics, economics and journalism. This resource is constantly used to manipulate data and built a specific narrative that interferes with social interactions and proper democratic processes. To manipulate the audience and promote a definite agenda, subjective propaganda is used in online media, namely, in online news articles. The issue with its use is that propaganda presents information selectively and in an emotional way. For that reason, detecting propaganda techniques in news articles is an important task that could promote participation and openness.

In our research, we use the same dataset as in the work of Da San Martino et al. (2019) Fine-Grained Analysis of Propaganda in News Articles. It includes 451 news articles from 48 news outlets, both propagandistic and non-propagandistic. The data was annotated by 6 annotators from A Data Pro. Our preliminary analysis of the annotated fragment of data showed that some of the annotations were duplicated (for example, Name calling and Labeling were separately reported while being the same propaganda technique). For that reason, we preprocessed the dataset, deleted duplicating rows, corrected the names of the techniques and trained 3 annotators to check the remaining data. They have additionally checked all of the propaganda techniques in the first 500 data points. The Fleiss' kappa between our three annotators is 0.75, while the Fleiss' kappa between them and the original annotation is 0.81. Considering that this agreement is high, we accepted this dataset and continued with the examination to then identify the context of each propaganda technique. Currently, no context of an identified technique is annotated in the text. We believe, context would be helpful for propaganda detection as it would provide more linguistic information about the surroundings of each phrase.

As a starting point to explore which parts of the text offer useful and sufficient context for identifying a propaganda technique, the sentence that contains the labelled technique and the sentence immediate before it are extracted. Then, the annotators are manually checking the extracted contexts and either add or delete part of it. With the manually annotated context information, we envision that the performance of neural network-based methods for the fine-grained propaganda detection will improve. We plan to try several vectorization approaches as well as state-of-the-art models as BERT.

Our research is expected to shed light on the various way a specific narrative is built in the online media through the use of propaganda, disinformation and other types of manipulative acts in the news content.

URL of the dataset will be provided upon acceptance of the paper.

MissDoom

Alexandre Leroux, Matteo Gagliolo

Université Libre de Bruxelles

February 1, 2020

Belgium, and Europe in general, experienced an increased attention of public opinion to immigration in the last few years, accompanied by outbursts of activity on online social networks, which ranging from solidarity with migrants and refugees, to xeno- and Islamophobia. These digital traces enable to study citizens opinion and attitudes about refugees and migrants. Combining textual information with Facebook network features allows us to piece together a topography of the disparate citizen narratives about migration. Our ongoing research aims to investigate those opinions and their evolution between January 2014 and December 2018, as discussed on Facebook.

The corpus consists of 24.8 million comments written on a set of 15 000 Facebook pages related to migration, between January 2014 and December 2018, and collected via the platform API. Out of the 6000 pages reporting geographic data, 83% report being located in Belgium. For this work we will focus on a third of comment which are expressed on page identified as French speaking.

From those public pages we distinguish three group: pro-migration, anti migration, neutral. In order to identify groups of pages, we create a network based on between-page activities, such as page A commenting on a post on page B; pages are then clustered by means of the *Leiden* community detection algorithm.

As we are interested only in migration related discourse, we diminish the amount of noise from the dataset through string matching. We implement results from a previous work on the same corpus to keep comments containing terms with the highest lexical and semantic similarity to "refugees" and "migrants". Those similarities are computed from words embedding applied to the corpus year by year.

In order to observes discursive patterns over times, we split the collections of comments in four-month intervals and define fifteen micro co-occurrence networks between lemmatized terms. Graphs edges strength are computed from between-word similarity measured from the period word-comment matrix.

We identify themes through the clustering of each micro networks; community of words is then themes-labelled. The last step in the analysis look upon junction and disjunction between months: between categories of discourse. We compare the words distribution between clusters by intervals connecting parallels topics in a macro network. The analysis of this macro network should allow us to observe the impact of punctual events as well as identifying trends in the public discourses during 2014 and 2019.

Judith Möller, University of Amsterdam & Michael Beam, Kent State University

Spiral of noise: towards a new theoretical framework to understand the effects of biased information

Current research into filter bubbles offers competing results. On the one hand, there is clear evidence of groups of users that make prolific use of algorithmic filter systems and artificial intelligence to exchange like-minded (mis)information and avoid the mainstream media (Jamieson, 2018). These users are trapped inside a so-called filter bubble, which has been linked to increased political polarization (Törnberg, 2018). On the other hand, there is a growing body of research that finds no evidence of filter bubbles on the aggregate level in the population (Flaxman, Goel, & Rao, 2016). We argue that these findings are not contradictory but a consequence of studying filter bubbles in the wrong place. Social media communities locked into communication bubbles are thriving, but in smaller and specific segments of the population and niche audiences as opposed to the population. These filter bubbles thrive among extreme political ideologies such as followers of extreme right-wing populist ideology.

Spiral of silence theory (Noelle-Neumann, 1974) assumes that extreme minority opinions are suppressed out of fear of isolation. Crucial in this process is the concept of the "quasi statistical organ," used to identify the majority opinion. We argue that echo chambers found on social media potentially impair this organ. Cloistered online information environments bias the perception of public opinion, to the effect that those who hold radical points of view feel surrounded by others agreeing with them (Matthes, Rios Morrison & Schemer, 2010). As a consequence, they feel safe to speak out.

We posit the reverse mechanism from those proposed in the spiral of silence theory to explain how the biased and filtered nature of information within filter bubbles affects the perceptions of those trapped inside. Since these points of view are often novel and newsworthy, they are prone to be picked up by the mainstream media, which increases the visibility of those voices even further (Benkler, Farris, Robers, & Zuckerman, 2017). This means, the spiral of silence is reversed due to algorithmic filter systems. Instead of being muted, are being amplified into a spiral of noise.

References

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, *80*(S1), 298-320.

Jamieson, K. H. (2018). *Cyberwar: how Russian hackers and trolls helped elect a president: what we don't, can't, and do know*. Oxford University Press.

Matthes, J., Rios Morrison, K., & Schemer, C. (2010). A spiral of silence for some: Attitude certainty and the expression of political minority opinions. *Communication Research*, *37*(6), 774-800.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of communication*, *24*(2), 43-51.

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, *13*(9).

Understanding polarizing content distribution on social media

As political polarization increases in contemporary democratic societies, research interest is growing in the wide range of alternative media sources that have recently come to prominence, and the extent to which these sources promote narratives of division and polarization in the societies which they operate. Such reporting has attracted a variety of labels (such as 'fake', 'junk' and 'hyperpartisan'), but the content they produce is largely similar: sensationalised stories about immigrants, religious or ethnic minorities, political opponents or people of differing sexualities which have the potential to generate mistrust and sow discord.

We are at the beginnings of understanding how this content operates, and what its effects are. In this paper, we contribute to the field by picking up the particular question of divisive and polarizing content distribution. A common theme in the literature is that this type of content, perhaps leveraging biases in the ranking systems of social media platforms, generates more engagement than that created by mainstream media, and has greater potential to 'go viral'. But research testing this claim is scarce; and more importantly, we know little about where this content generates its engagement from.

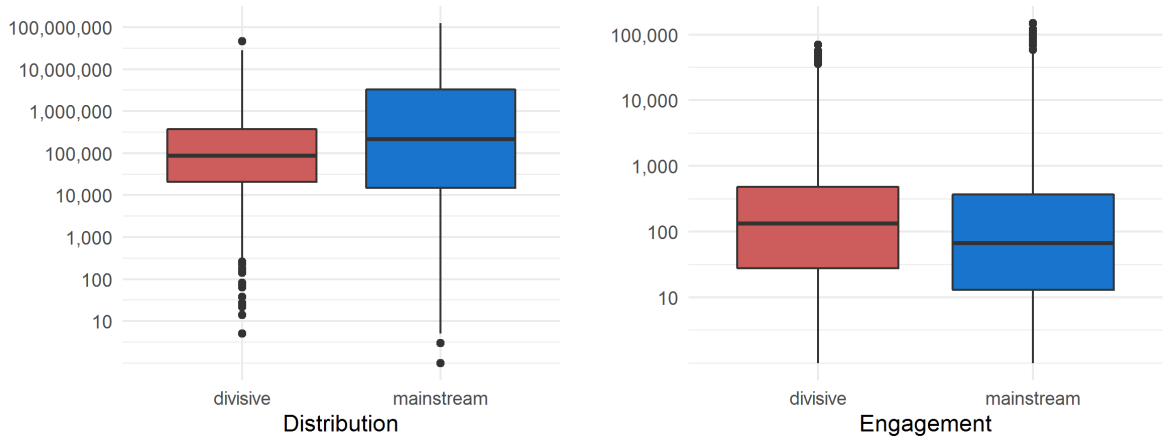


Figure 1: Preliminary Results. *Left panel shows the distribution of divisive content in our data, right panel shows engagement.*

Based on a unique dataset of hand coded polarizing content produced by UK based alternative media, we seek to answer two major questions. First, we study the engagement generated by divisive content. We examine whether it differs significantly from that generated by mainstream content, and also ask whether engagement levels are consistent across different types of polarizing content (from rhetoric about immigrants and minorities to narratives of state and climate conspiracy). Second, we explore the types of venues sharing this content, looking at their makeup, motivation, and whether they consistently amplify the same groups and voices. Our study covers Facebook, Twitter and Reddit.

Figure 1 above provides some preliminary results from the study. Our data appear to show that mainstream content is more shared than divisive and polarizing content. However, when it is shared, divisive content generates more engagement.

The ‘deepfake porn’ phenomenon went viral in November 2017 when AI-manipulated pornography was uploaded to discussion website, *Reddit* (Cole, 2017a). A month later, over 80,000 people had shared their own ‘deepfake’ porn on the site. Even though 96% of deepfakes are pornographic (Ajder et al, 2019), public attention typically centers on ‘political’ deepfakes. While powerful institutions clamour to address politically oriented deepfakes, their pornographic counterparts have become part of the scenery in cyberspace. This article explores the relationship between ‘political’ and pornographic deepfakes, finding that they operate in similar ways to silence critical speech. Using source material from *Twitter*, this article identifies three processes that have shaped the rise of deepfake porn: Its exclusion from news media reporting, its normalization on porn sites, and its automated distribution online. By exploring the continuities between pornographic and ‘political’ deepfakes, this article demonstrates that policy responses to all deepfakes must be informed by their roots in pornography.

Deepfake porn is a new phenomenon that fits into a long history of ‘audio-visual manipulation’ (Paris & Donovan, 2019), and scholars grappling with the ethics of deepfake porn are coming to a range of conclusions (See Öhman 2019, Popova 2019, Newton & Stanfill 2018). While deepfake technology can operate as a subversive source of creativity and pleasure within pornography, it causes severe harm when created non-consensually (Siegemund-Broka, 2013; Citron, 2019). This article defines non-consensual deepfake porn as an invasion of sexual privacy (Citron, 2018; Henry et al, 2018) and an example of image-based sexual abuse (Powell et al, 2018; McGlynn & Rackley, 2017: 536).

The sample created for this research was drawn from the *ASC Wharton* database, which gathers a random 1% sampling of daily tweets through *Twitter’*s public stream. A systematic search identified 3,595 English Language tweets published from 2012 – 2019 that reference deepfake porn and synonymous terms. *Twitter*’s close ties to the pornography industry make it an interesting site from which to chart the rise of deepfake porn (Cole, 2017b; Bell, 2017). Although *Github* and *Reddit* were essential to the creation and distribution of deepfakes (Winter & Salter, 2019), *Twitter* data tells a longer story about the rise of fake porn. This article finds that same factors driving deepfake porn also drive ‘political’ deepfakes. This article recommends that technological solutions and regulatory policies address the underlying inequalities that shape both pornographic and non-pornographic content.

References

Ajder, Henry, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. 2019. *The State of Deepfakes: Landscape, Threats, and Impact*.

Bell, Karissa. "Twitter just Updated its Policy on Revenge Porn and Non-Consensual Nudity 'to Better Protect Victims'." Mashable., last modified /10/27, accessed Oct 28, 2019, https://mashable.com/2017/10/27/twitter-revenge-porn/.

Citron, Danielle Keats. House Permanent Select Committee on Intelligence. 2019. *The National Security Challenge of Artificial Intelligence, Manipulated Media, and 'Deep Fakes.* June 13,.

———. . 2018. *Sexual Privacy*. Rochester, NY. https://papers.ssrn.com/abstract=3233805.

Cole, Samantha. 2017a. *AI-Assisted Fake Porn is here and We're all Fucked*. Vice. https://www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn.

———. 2017b. *Porn is Still Allowed on Twitter*. Vice. https://www.vice.com/en_us/article/xwavkq/porn-allowed-on-twitter-adult-content-ban-new-guidelines-pornhub-patreon.

———. 2018. *We are Truly Fucked: Everyone is Making AI-Generated Fake Porn Now*. Vice. https://www.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley.

Hasan H. R. and Salah. K. 2019. *Combating Deepfake Videos using Blockchain and Smart Contracts*. Vol. 7.

McGlynn, Clare, Erika Rackley, and Ruth Houghton. 2017. "Beyond 'revenge Porn': The Continuum of Image-Based Sexual Abuse." *Feminist Legal Studies* 25 (1): 25-46.

Newton, Olivia B. and Mel Stanfill. 2019. "My NSFW Video has Partial Occlusion: Deepfakes and the Technological Production of Non-Consensual Pornography." *Porn Studies*: 1-17.

Öhman, Carl. 2019. "Introducing the Pervert's Dilemma: A Contribution to the Critique of Deepfake Pornography." *Ethics and Information Technology*. doi:10.1007/s10676-019-09522-1.

Paris, B., & Donovan, J. 2019. "Deepfakes and Cheap Fakes." *Data & Society Research Institute*. https://datasociety.net/output/deepfakes-and-cheap-fakes/.

Popova, Milena. 2019. "Reading out of context: pornographic deepfakes, celebrity and intimacy." *Porn Studies Journal.*

Powell, Anastasia, Flynn, Asher and Henry, Nicola. "AI can Now Create Fake Porn, Making Revenge Porn Even More Complicated." The Conversation., accessed Jan 17, 2020, http://theconversation.com/ai-can-now-create-fake-porn-making-revenge-porn-even-more-complicated-92267.

Siegemund-Broka, Austin. "'Storage Wars' Star Brandi Passante Wins 'Stalker Porn' Lawsuit." The Hollywood Reporter., last modified /07/01, accessed Oct 28, 2019, https://www.hollywoodreporter.com/thr-esq/storage-wars-star-brandi-passante-578047.

Westerlund, Mika. 2019. "The Emergence of Deepfake Technology: A Review." *Technology Innovation Management Review* 9: 39-52.

Winter, Rachel and Anastasia Salter. 2019. "DeepFakes: Uncovering Hardcore Open Source on GitHub." *Porn Studies*: 1-16. doi:10.1080/23268743.2019.1642794. https://doi.org/10.1080/23268743.2019.1642794.

Experimental Research in Progress; Beliefs in Conspiracy Theories on the Web

Arnout B. Boot, Katinka Dijkstra, and Rolf A. Zwaan

DPECS, Erasmus School of Social and Behavioral Sciences,
Erasmus University Rotterdam, the Netherlands

The rise of social media and the Web have introduced virtually unlimited communication and unprecedented access to information. Within two decades, they have become an integral part of society, affecting human psychology, beliefs, and behavior. For instance, Web users have access to a wide array of informative sources (e.g., Wikipedia, Google's search engine, YouTube, social-media platforms, news aggregation/discussion sites, forums) that can influence their knowledge and beliefs. Additionally, Web users (partially) control the type of information they encounter, and in some occasions explicitly search for information that confirms or supports prior beliefs.

Confirmation bias has led to a segregation within online communities, reflecting *echo chambers* or *social-media bubbles,* in which reports of confirmative and partisan (mis)information is encouraged. Conspiracy theorists are a notorious example. They have online communities with a strong ingroup-outgroup dynamic, in which they experience an ideologically common-ground to confirm and reciprocate their personal beliefs (e.g., flat-earth society). Social media as well as the Web have provided conspiracy theorist an accessible public platform to widely spread their beliefs. Considering the ubiquity of conspiracy theories on the Web, it is vital to understand how novel readers process these types of information.

In an online experiment, novel readers will be introduced to a conspiracy theory on a custom-build website. The reader will have (ostensibly) naturalistic browsing control (e.g., options to read more or less) and will be incrementally exposed to information about the conspiracy. Participants in this experiment will not be aware of behavioral measurements during the task. Instead, they will be under the impression that they are browsing on a real social-media platform. Furthermore, we aim to implement an implicit measure that indicates whether the participant accepts and fully (or partially) beliefs the conspiracy theory. Afterwards, the participant fills in a questionnaire about the credibility of individual components of the story. In addition, we aim to compare different types of media-education interventions to improve media literacy and intellectual skepticism in the participants. After the task, participants will be debriefed about the experimental manipulations and measurements.

We would appreciate to present our work (in progress) at the MISDOOM symposium. With a background in the field of Cognitive Psychology we focus on the Web-user's information processing, reasoning, decision-making and behavior. Our study could yield new insight into the formation of beliefs in conspiracy theories, and more generally, the acceptance of false information. In addition, our research might have valuable implications about the development of educational methods to improve media literacy and skepticism.

Entertaining far-right propaganda on Instagram: User reactions to eudaimonic posts

Tobias Kleineidam[1], Lina Gunstmann[1], Anna Schonebeck[1], Tim Schatto-Eckrodt[1][https://orcid.org/0000-0003-1658-4373], and Lena Frischlich[1][https://orcid.org/0000-0001-5039-5301]

[1] Institute for Communication Science, Westphalian Wilhelms-University Muenster

Abstract.
Social media platforms such as Instagram have become a powerful tool for extremists aiming at inspiring new followers via propaganda disguised as entertainment. Entertainment can be distinguished in two different processes of experiencing entertainment – the hedonic and the eudaimonic process of entertainment experience. The first one is characterized by emotions of fun and amusement and the latter is linked to feelings of inspiration and meaning, enhancing the recipient's self-transcendence. The current study investigates how recipients react to these different types of entertainment in extremist propaganda. Assuming that especially eudaimonic entertainment might be of interest for extremists, the experimental study compared the reactions of Instagram-users (N = 153) who saw either hedonic or eudaimonic right-wing extremist propaganda versus a neutral post. Results showed that eudaimonic posts triggered both eudaimonic affect as well as eudaimonic cognitions in the recipients. Furthermore, users were more willing to amplify eudaimonic posts by liking and sharing them and were more willing to follow the far-right account who had posted them. Our findings offer new insights into the dynamics of the darker side of eudaimonic entertainment and enhance our understanding of (far-right) propaganda in the present days.

Keywords: Propaganda, Instagram, Dark inspiration, Eudaimonic entertainment.

2

Extended abstract

In the wake of the digitalization, extremists have discovered the Internet as a powerful tool for information disseminate [1] Often extremist propaganda is disguised as entertainment, particularly on social media platforms like Instagram. Using entertainment research as framework, the current study investigates how recipients react to entertaining extremist propaganda on Instagram and what kind of propaganda evokes the strongest reactions. Several studies reveal two different processes of entertainment experiences, one being hedonic, characterized by emotions of fun, excitement and amusement, the other being eudaimonic entertainment. Eudaimonic entertainment is linked to feelings of inspiration, meaning and morality, enhancing the recipient's self-transcendence [2]. Particularly the latter might be of interest for extremists aiming at inspiring new followers.

To test this assumption, our experimental study ($N = 153$) compared the reactions of Instagram-users who saw either hedonic or eudaimonic right-wing extremist Instagram propaganda versus a neutral post, answering the question whether different types of entertainment in extremist propaganda caused different responses to extremist ideologies. After each post, the participants' "state of inspiration" (i.e. their eudaimonic affect and cognitions) was measured. Additionally, we asked about recipients' behavioral intentions as regard to the post and the account which had posted it.

Results showed that eudaimonic posts triggered both eudaimonic affect as well as eudaimonic cognitions in the recipients. Furthermore, users were more willing to amplify eudaimonic posts by liking and sharing them and were more willing to follow the accordant who had posted them.

Overall, our results can help to clarify the dynamics of the darker side of eudaimonic entertainment, especially in comparison to its hedonic counterpart. Research on this matter enhances our understanding of (far-right) propaganda in the present days and might even help to find ways to protect social-media users from its bad influence.

References

1. Frischlich, L.: Propaganda3: Einblicke in die Inszenierung und Wirkung von Online-Propaganda auf der Makro-Meso-Mikro Ebene. In B. Zywietz (Hrsg.), Fake-News, Hashtags & Social Bots: Neue Methoden der populistischen Propaganda (Propaganda). Springer Fachmedien VS (2018).
2. Oliver, M. B., Raney, A. A., Slater, M. D., Appel, M., Hartmann, T., Bartsch, A., … Das, E. Self-transcendent media experiences: Taking meaningful media to a higher level. Journal of Communication, 380–389. https://doi.org/doi:10.1093/joc/jqx020 (2018).

Disinformation based on surveillance and the disappearance of privacy: the use of personal data in the direction of false information and the impact on reducing individual autonomy

MILENA FISCHER
Country: Brazil. City: Porto Alegre. State: Rio Grande do Sul.
Contact: milfischer@gmail.com.

Journalist with 22 years of experience in press, book publishing and communications management – today, I she is also a Law student at Fundação Escola Superior do Ministério Público do Rio Grande do Sul, in Brazil. Milena has been honored with visiting scholar grant by the Freedom Forum (USA), which promotes freedom of the press, and also with a grant from Inter-American Development Bank to participate in youth leadership meeting in Israel. Today, runs her own digital content agency.

Brazil is among the examples of countries in which the election process of a President was marked by the wide spread of fake news and the emergence of disinformation networks, which gained space on social media platforms such as Twitter, Facebook and WhatsApp. After episodes like this, as well as the 2016 US elections or the Brexit, these companies began to be questioned by public officials.

The first question that arises concerns accountability: why do public authorities allow private companies to establish their own guidelines, when these have become spaces for public debate to the point of influencing - or even determining - the fate of democracies, at the expense of targeting information and disinformation to people vulnerable to them? To what extent can the freedom to exercise an economic activity justify intervention in the course of democracy and people's private lives - without them even knowing that they are being manipulated? If in many countries there is ample regulation in sectors such as infrastructure and health, there is no reason why, today, public authorities do not exert greater action on these companies and work together on rules and laws that limit the use of data for disinformation purposes. If someone is a constant target of misinformation and fake news, he has his power of choice limited by the imposition of a dystopian reality.

This unregulated and intentional manipulation has contributed to weakening democratic institutions. The prerogative today for the fundamental right to privacy is the right to data protection. Without it, there is no need to talk about privacy or autonomy, since the power of self-determination becomes conditioned to the previous manipulation of the person's information and data.

In a dialogue between freedom of expression and the future of democracy, it is necessary to seek answers to the question: how far can the State and the Law intervene in the transit of data and information in order to safeguard the privacy of the person so that they have autonomy to make choices in a democracy? The thesis we pursue is that of state discomfort. In a technological scenario in which capitalism is fueled by web surveillance and the data that people make accessible (and those that they don't even know are being accessed), it is unreasonable to expect large private companies, with the power to process data, to self-regulate. It is necessary for the State to be active in order to stop anti-democratic practices - without impeding technological development.

And this must be a transnational effort, because data protection laws are national but in the online environment, data does not respect borders. Cross-border cooperation and agreements to deliver effective data protection are essential to democracies.

Russia's Recycling of Strategic Narratives in Epistemologic Truth-setting in the Baltics

Monika Hanley[1]

[1]Fulbright Researcher, Technical and Scientific Development Branch, NATO Strategic Communications Centre of Excellence, Latvia

Abstract

The spread of disinformation via social media for the purpose of election disruption and public division is not a new phenomenon, though more attention has been paid to it in recent years. The hand of Russian state-backed social media can be found in thousands of profiles, in English and non-English messages and across almost all social media platforms.

Consumption and reaction to information on Twitter occurs more quickly than verification of truth or accuracy. Information is spread faster and farther than ever before, with very minimal effort. This ever-present threat to the quality of information and societal resiliency has become more pressing as ease of spreading disinformation grows. Russian-originating disinformation in the Baltic information sphere began over 70 years ago and continued throughout the Soviet period. Many narratives promulgated during this time are rebranded and used today, albeit spread much faster with the help of social media.

This paper critically examines the tweet as an emerging epistemic category in the Baltic states and the impact that these old narratives have in reaching new audiences via social media.

Approaches and Methods

By analysing three categories of narratives identified in the Baltic Twitter information environment, (anti-NATO sentiment, information to sow discord between ethnic Russian populations and Baltic governments, rewriting of history) we are able to view the predominant and unwavering narratives deemed to be at the forefront of the Russian disinformation strategy against and about the Baltic states. The article then concludes with practical context to truth setting and historical resilience in the Baltics. Following these categories is an analysis of the success of the messages, i.e. which have been most impactful in destabilizing the local truth environment. An analysis on Baltic resiliency against these messages will also be explored, along with the potential successes emerging.

From this, we can discern how beliefs are being built in the Baltic states and if the messages have or do not have an impact on this truth foundation. We can also see how old narratives from the Soviet period have gained new traction in the Twittersphere.

Using Contextual Features to Detect Online Influence Campaigns

Meysam Alizadeh[1], Jacob N. Shapiro[1], Cody Buntain[2], Joshua A. Tucker[2]

[1] *Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ 08544, USA*
[2] *Department of Politics, New York University, New York, NY 10012, USA*

We study a platform-agnostic method of using public activity to detect coordinated influence operations on social media. This is different from bot detection task because (1) not all participating accounts are bots, and (2) the focus is on the coordinated and networked nature of their behavior, as opposed to their individual behavior. Our approach classifies the post-URL pair based on more than 1,000 human-interpretable features derived solely from content without relying on historical or friendship network data. For example, we classify a given tweet without using user's previous tweets or her following/follower network. We test this method on Twitter data on Chinese (2,660 accounts, 1.9M tweets), Russian (3,722 accounts, 3.7M tweets), and Venezuelan (594 accounts, 1.5M tweets) troll activity targeting the United States, as well as the Reddit dataset of Russian influence efforts (944 accounts, 15K posts).

To assess variability over time in how well content-based features distinguish such influence operations from random samples of American users and politically-engaged American users (23M and 21M tweets respectively) we train and test classifiers on a monthly basis for each campaign across four out-of-sample prediction tasks: (1) within-month 50/50 train/test split; (2) train in *t*-1, test on all posts *t*; (3) train in *t-1*, test on posts by previously unused accounts in *t*; and (4) train on accounts identified at one point by Twitter and test on activity by accounts identified at a later date. Content-based features perform well across period, country, platform, and experimental design (Table 1). Average monthly F1 scores on task (1) range from 0.82 on the Russian Reddit campaign to 0.99 on the Venezuelan Twitter operations. F1 scores for Russian troll activity ranges from 0.85 in task (1) to 0.74 in task (4). Regression results show that the observed drops in predictive performance in some months can be explained by low activity of trolls (i.e. small test size) and major political events which lead to significant changes in trolls tactics and routines.

Table 1: Average and standard deviation of monthly F1 scores across campaigns and platforms.

Operation	Platform	Within-Month 50/50 Train/Test[a]	Train on $t-1$ Test on t[b]	Train on $t-1$ Test on New Users in t[c]	Within-Month Cross-Release
China	Twitter	0.89 (0.08)	0.93 (0.04)	0.89 (0.12)	NA[d]
Russia	Twitter	0.85 (0.13)	0.81 (0.07)	0.81 (0.13)	0.74 (0.12)
Russia	Reddit	0.82 (0.07)	0.82 (0.09)	0.74 (0.15)	NA[d]
Venezuela	Twitter	0.99 (0.03)	0.99 (0.002)	0.92 (0.15)	0.49 (0.07)

[a] We report the results for the case of training on half of users and testing on the other half.
[b] All features related to account creation date (the only user-level feature we have) have been excluded in training.
[c] We calculate the average and standard deviation over those months in which there are at least 1,000 trolls tweets or 500 trolls Reddit posts in test set.
[d] Not Applicable. There is only one official data release for Chinese campaign on Twitter and Russian campaign on Reddit as of December 1, 2019.

Important features vary by month, operation, platform, and experiment (i.e. Tests 1-4). However, in general, meta-content features such as top hashtags used or top users mentioned by trolls plus features related to the age of accounts are frequently among the most important features. In addition, analyzing the dynamics of feature importance over time provides insights about trolls tactics. Finally, our false negative experiment results reveal that in Test 1, inducing a 1% false negative users leads to 0.01 reduction in average monthly F1 score. However, it caused a 0.02 and 0.006 increase for test 2 and 3 respectively, presumably by reducing over-fitting. We also see a 0.04 reduction in average monthly F1 score for Test 4. False positive results, effect of training size and testing time-window, reasons for why Venezuelan campaign is so easy to detect, monthly important features, active learning simulation, and policy implications will be discussed.

Neta Kligler-Vilenchik, Hebrew University of Jerusalem

Title: Information Verification Practices among Political Talk Groups on WhatsApp

Recent years show a growing concern about the spread of mis- and disinformation on social media (e.g., Alcott, Gentzkow & Yu, 2019; Tandoc, Jim & Ling, 2018). While some social media companies have been suggesting technologically-oriented solutions, the current project focuses on the human factor in preventing the spread of misinformation, by examining a case study of a group that has cultivated advanced practices for information verification and fact-checking.

Employing a mixed-methods approach combining quantitative and qualitative analysis, the paper examines a large WhatsApp group devoted to political talk[1]. The group, called "The Workers", was opened by a journalist, and at the time of research consisted of around 90 participants, with an average of 270 messages posted daily. Group participants are diverse in terms of political leaning, yet all are highly interested in politics—to the extent that they are willing to pay a nominal fee for group participation. This participation fee, and the fact that the group discusses politics, lends it a semi-public nature. This ongoing project focuses on identifying the groups' mechanisms for information verification, and for avoiding the spread of misinformation. We have finished the quantitative analysis, as described in this abstract, and are currently conducting qualitative analysis.

For quantitative analysis, we collected all the contents posted to the group during 2016[2], creating a corpus of 101,351 messages. On this corpus, we conducted manual coding, with the aim of identifying *instances of information verification*, defined as cases in which *two or more participants interact around contradicting or casting doubt about a piece of verifiable information*. Each information verification instance was then coded[3] for several variables, particularly whether the case consisted of casting doubt or of contradiction, and whether the case included a discussion of source credibility.

We identified 795 instances of information verification in the corpus, with an average of 2.2 a day. Out of these, the overwhelming majority (91%) of cases consisted of casting doubt about a piece of information, while contradicting a piece of information was relatively rare (9%). Slightly over 10% of cases included a discussion of source credibility. The group administrator played a very significant role in this process, taking part in almost half (47%) of information verification cases.

In the next, qualitative part of the research, we aim to identify the mechanisms through which the group propagates information verification as a salient group norm. By analysing the communicative processes enacted in information verification, we aim to show how group participants collectively construct a discussion environment in which participants know they can and will be held accountable for information they share.

The project thus points at the potential of groups on social media to collectively construct norms and practices of information verification, which help them to avoid the spread of misinformation.

For a related publication see: Kligler-Vilenchik, N. & Tenenboim, O. (2020). Sustained journalist-audience reciprocity in a meso news-space: The case of a journalistic WhatsApp group. *New Media & Society, 22*(2), 264-282.

[1] The research has been approved by the IRB; The group participants and the group administrator have consented to the research.

[2] We ceased data collection after 2016 because at that time the group administrator stopped charging for participation in the group, which arguably shifted its semi-public nature.

[3] Inter-coder reliability among the four coders, on around 10% of the data, reached satisfactory levels (Krippendorff's alpha = .74 for doubt/contradiction, and .79 for discussion of source credibility).

GENDER FOCUS ON CONSTRUCTION OF NARRATIVE STRATEGIES FOR FIGHTING POLITICAL DISINFORMATION: THE BRAZILIAN CASE

Beliza Boniatti[1], Mariele de Almeida Hochmüller[2]

Digital disinformation strategies were strongly used on Brazilian presidential election in 2018. The result was the election of a far-right candidate - who has a racist and misogynist rhetoric - with more than 57 million votes, and with 52% of voting women endorsing him. The fact that politics is a field traditionally occupied by men, in which women were for centuries prohibited to vote and to run for office, reverberates gender inequality and reproduces the social construction that women should not be part of politics. The present study defends that social the construction of women is a central subject that must be fought in a collective and social way. Hence, fighting disinformation is an essential tool for the emancipation of women's critical and political thinking. Therefore, even though there must be a global action against disinformation, strategies focused on specific niches tend to be more effective.

The present ongoing study aims to find answers on how to fight disinformation of political themes with narrative strategies built from a gender perspective based on the brazilian context. Thus, following are the secondary objectives: a) to understand which points are related to the social construction of women that could serve as a basis for a niche narrative construction; b) to identify which fake news about politics most affect women; and c) to build a sociological basis for approaching non-hegemonic narrative strategies.

First, this study carried out a bibliographical investigation to understand the construction of gender in Western society. Second, by conducting a documentary research, this study identifies the key fake news that influenced the presidential elections in 2018 in Brazil. Additionally, it conducts qualitative semi-structured in-depth interviews with Brazilian women who voted for the far-right. Based on this information, this paper analyses the data from a historical-materialist perspective that allows the construction of a sociological base for future non-hegemonic narratives.

The main hipotesis of this study is that women feel more affected by fake news related to their social construction as woman or traditionally feminine values. Therefore, showing that a gender perspective is needed when fighting fake news spread by hegemonic narratives. The current state of this investigation is able to present the intermediate results and the draft of the sociological basis that links gender construccion with narrative strategies focused by niches.

[1] PhD Candidate in the Freie Universität Berlin in Social Communication Studies (bebodos@alumni.uv.es).
[2] PhD Student in the Universitat de Valencia in Gender and Equality Policies (dealhoch@alumni.uv.es).

Stance Classification for Rumour Verification in Social Media Conversations

Elena Kochkina[1,2], **Maria Liakata**[1,2], **Arkaitz Zubiaga**[3]
[1] University of Warwick, Coventry, United Kingdom
[2] Alan Turing Institute, London, United Kingdom
[3] Queen Mary University of London, London, United Kingdom

Due to the risks posed by the proliferation of unverified content online, there is a need to develop Machine Learning (ML) methods to assist with the verification of circulating rumours, statements unverified at the time of posting. Rumour verification can be formulated as a classification problem, where a model is trained to predict if a rumour is true, false or unverified, given posts discussing a rumour as the input. Previous research (Zhao et al., 2015) has shown that rumours attracting a lot of sceptical and denying reactions are more likely to be proven false later. Thus classifying the stance of posts towards rumours automatically is an important task that aids rumour verification. Thus, we propose a talk discussing the relation between the tasks of rumour stance and veracity classification in social media conversations and giving the overview of recent advances leveraging that relation based on our work in this domain and experience from organising a shared task.

The RumourEval shared task (Derczynski et al., 2017; Gorrell et al., 2019) was proposed to test the hypothesis regarding the synergy between stance and rumour veracity. RumourEval consists of 2 sub-tasks: (A) rumour stance classification and (B) rumour veracity classification, where the input is a collection of Twitter conversations discussing rumours related to news breaking events. In its first edition in 2017, the winning system of subtask B was the only system that used the predicted stance labels as features for their classifier. In the second edition of RumourEval in 2019 more systems utilised stance as a helpful feature to determine veracity. The winning system of subtask B, which outperformed baselines (winners of previous edition) and other competitors, used an ensemble of classifiers and stance extracted from subtask A.

Furthermore, we explored the incorporation of stance classification into rumour verification as an auxiliary task in a multitask learning set up, when a deep learning model was trained to perform several tasks simultaneously (Kochkina et al., 2018). The results show that the joint learning of two tasks from the verification pipeline outperforms a single-learning approach to rumour verification for RumourEval and larger PHEME dataset. The combination of three tasks (stance classification, detection and verification) leads to further improvements. Dungs et al. (2018), proposed a competitive approach using Hidden Markov Models with stance and tweets' times features for rumour verification. Recent work (Lillie et al., 2019) suggested that stance-based veracity works across languages and platforms. In conclusion we will outline open challenges that rumour verification models are facing, and share our view on how to tackle them.

References

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413.

Anders Edelbo Lillie, ITU Copenhagen, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405. International World Wide Web Conferences Steering Committee.

Prediction of Complaints of Hoax News in West Java using Chapman Kolmogorov's Equation and Markov Chain Stationary Distribution

Marsha Cahya Anggarwati[1], Firdaniza[2], Atje Setiawan Abdullah[3], Juli Rejito[4], Diah Chaerani[5], Annisa Nur Falah[6] and Budi Nurani Ruchjana[7*]

[1,2,5,6,7]Department of Mathematics, Universitas Padjadjaran
[3,4]Department of Computer Science Universitas Padjadjaran
Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363
West Java-Indonesia
budi.nurani@unpad.ac.id

Abstract.

The development of information technology is currently growing more rapidly and can not be separated from various negative impacts, one of them is the rise of hoax news spread through online media. Hoax news distribution that continues to increase can be detrimental to many parties, so it needs to be predicted how will be the level of hoax news complaints in the future. In this research the Chapman-Kolmogorov equation is used to determine the prediction of the next three days level of hoax news complaints in West Java and the stationary distribution of the Markov chain is used to determine the long-term prediction of hoax news complaints in West Java. The results of this research indicate that the level of hoax news complaints in West Java in the next three days and long term is in a state of decline, that means the number of hoax news complaints in West Java today is less than yesterday.

Keywords: Hoax, Chapman-Kolmogorov, stationary distribution of the Markov chain.

Filter bubbles and opinion polarization.

Why we may not even be close to having understood the complex link.

Michael Mäs and Marijn Keijzer

Extended Abstract

Political events such as the Brexit referendum, the election of Donald Trump, and the success of other populist politicians in democratic elections have sparked an intensive public and scholarly discussion about the effects of online communication technology on public debate and collective decision-making. One of the most prominent warnings is that personalization algorithms installed in online social networks, search engines, and online stores contribute to the formation of so-called "filter bubbles". These bubbles create echo chambers, isolating users from information that might challenge their views and exposing them to online content that is in line with their views, and, thus, reinforces their opinions. Experts, pundits, and scholars have warned that this contributes to opinion polarization, a dynamic where competing political camps develop increasingly opposing political views. Public attention is enormous. Newspapers regularly cover the topic (e.g. Chapin, 2018; Lapowsky, 2019); leading politicians echo the warning (Obama, 2017; Steinmeier, 2018); and various initiatives have been undertaken to fight filter bubbles and polarization (Bozdag & van den Hoven, 2015). Here, we summarize the key arguments underlying the hypothesis that personalization algorithms contribute to opinion polarization and reflect on existing scientific research. While we echo the warning that personalization might have serious effects on societal processes, we also point to gaps in the theoretical and empirical literature that need to be filled before one can draw conclusions about whether or not personalization is indeed responsible for increasing polarization. Unlike other recent contributions (Bruns, 2019), we do not argue that personalization is an innocent technology, but conclude that experts, politicians, and also scientists leap to conclusions when they propose that personalization is responsible for increased polarization. Accordingly, we call for more research on communication in online environments,

pointing to the potential of approaches that combine theoretical modeling with the emerging field of data science.

Our analysis is inspired by the complexity approach (Bar-Yam, 2003; Michael Mäs, 2018; Page, 2015) and builds on a rich literature in the field of opinion dynamics in social networks. This work departed in the 1950s in the social sciences and today profits from contributions from disciplines as diverse as physics, computer science, mathematics, economics, philosophy, sociology, political science, and complexity research (Flache, Mäs, et al., 2017; Friedkin & Johnsen, 2011; Mason, Conrey, & Smith, 2007). In this literature, formal models of social networks have been developed, where network nodes exert social influence on the opinions of their contacts. These models allow one to understand the rich and intricate opinion dynamics that arise from social influence and to identify the conditions under which repeated social influence fosters the formation of opinion consensus, the fragmentation of the network into multiple clusters with competing opinions, or even opinion polarization. Decades of modeling work with analytical and computational methods have demonstrated that even seemingly innocent changes in models' assumptions can have profound effects on the outcomes of social influence processes, which shows that drawing conclusions about real complex systems, such as online communication systems, requires a formal model that is informed by detailed empirical research. This model is not available, to date.

In a nutshell, we argue that the current public and scholarly debate about the personalization-polarization hypothesis has been paying too little attention to two important aspects. First, many contributions do not acknowledge the complexity of online social networks arising from repeated social influence between users. Complexity arises when a system consists of multiple micro-entities (users) that do not act in isolation but exert influence on each other (Bar-Yam, 2003; Page, 2015). In online social networks millions of users with a large number connections communicate with weak constraints on time and space, making these systems a very typical example of a complex system. Interdependency between users can generate chains of reaction such that even rare idiosyncratic events can have profound impact on the system as a whole (Macy & Tsvetkova, 2013; M. Mäs & Helbing, 2017). So far, most contributions to the public and scholarly debate about the personalization-polarization-hypothesis are based on informal theoretical arguments and anecdotal evidence, and thus fail to address system complexity. We do not argue that the conclusions drawn from these contributions are necessarily false, but we discuss findings from

complexity research that demonstrate how conclusions can change when a system's complexity is considered.

Second, we argue that contributions to the current debate tend to lean heavily on empirical and theoretical research on communication in offline worlds. We review insights from the opinion-dynamics literature showing that there may be differences between online and offline interaction that can critically alter opinion dynamics. In particular, we distinguish three levels of communication networks on which these differences typically reside: the individual level, the local level, and the global level.

The talk is organized as follows. First, we summarize the central theoretical, empirical, and political arguments underlying the scholarly and public debate about the effects of personalization on polarization. Next, we identify gaps in these debates, reviewing findings from the literature on opinion dynamics in social networks. In the concluding part, we sketch an agenda for future research, advocating an approach to data science that combines empirical research with rigorous theoretical modeling.

Figures used in the remainder of the paper

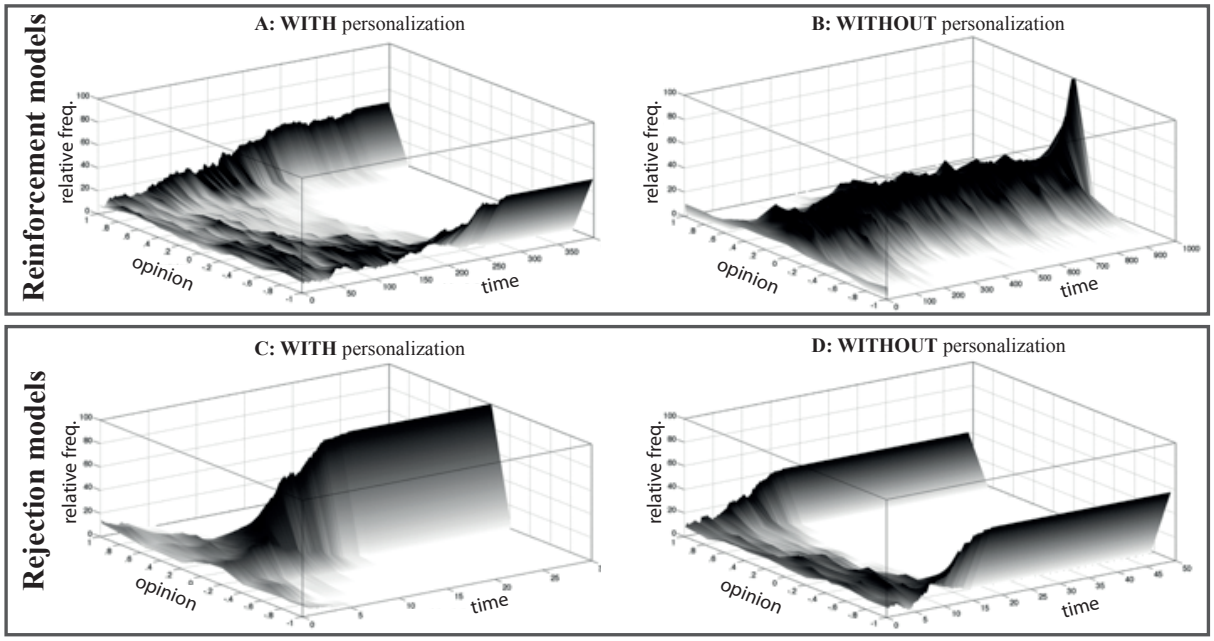


Figure 1: Contradicting predictions of reinforcement and rejection models for increases in web personalization

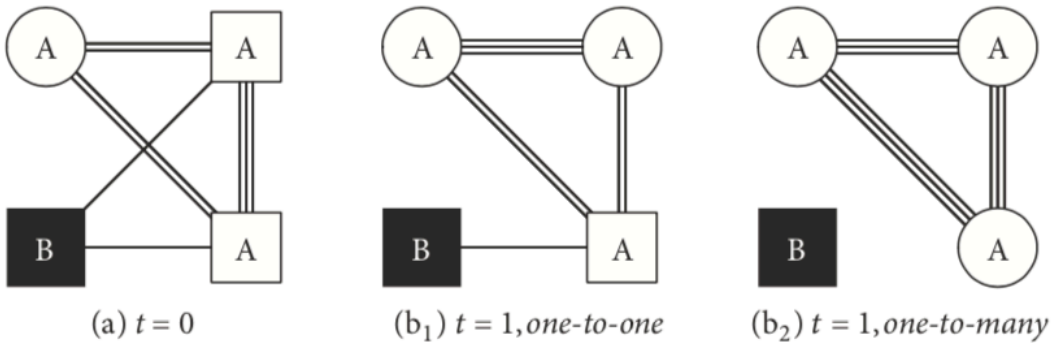


Figure 2: Illustration of the intuition that one-to-many communication fosters isolation. (Keijzer, Mäs, & Flache, 2018)

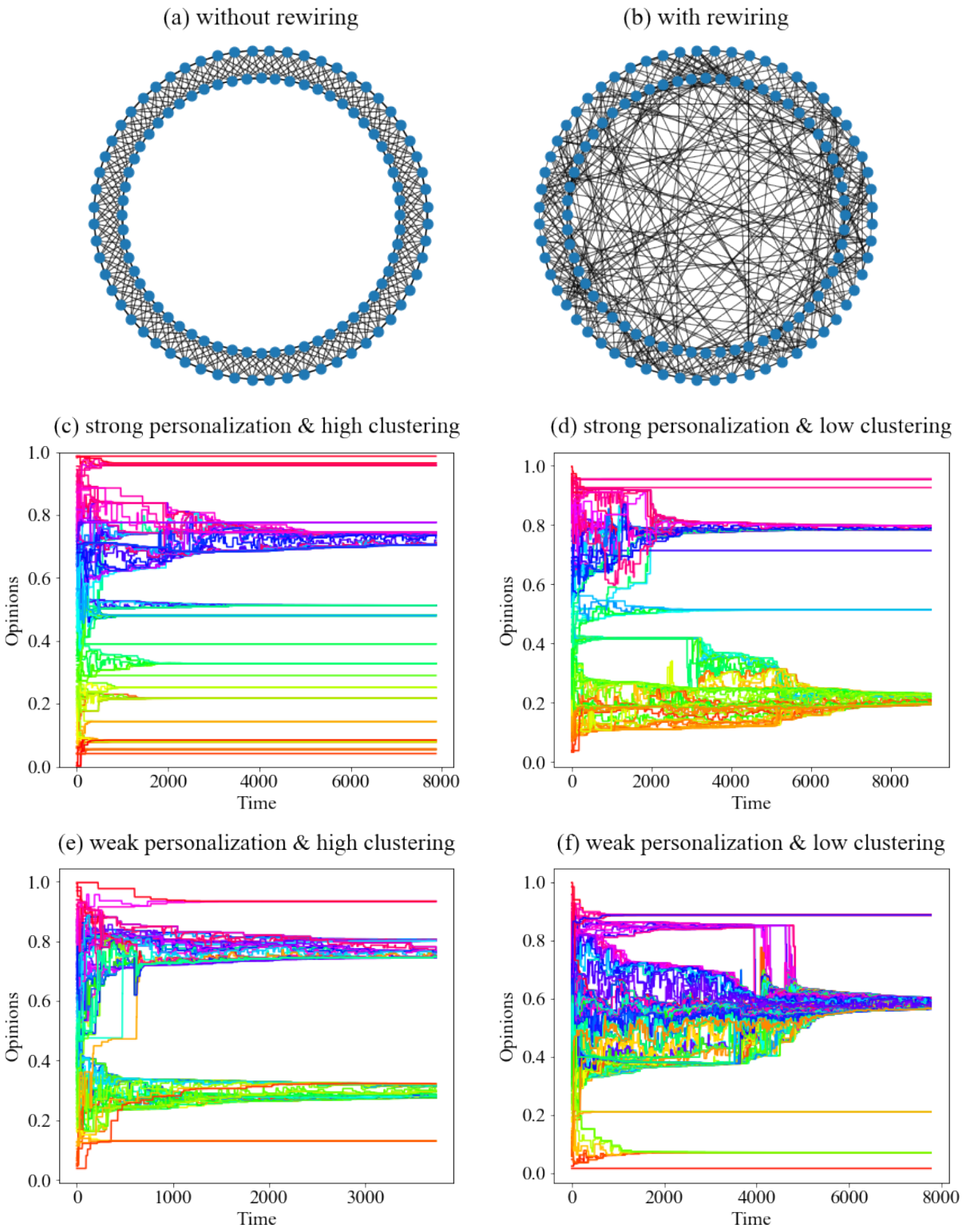


Figure 3: Effect of network clustering and personalization on opinion fragmentation.

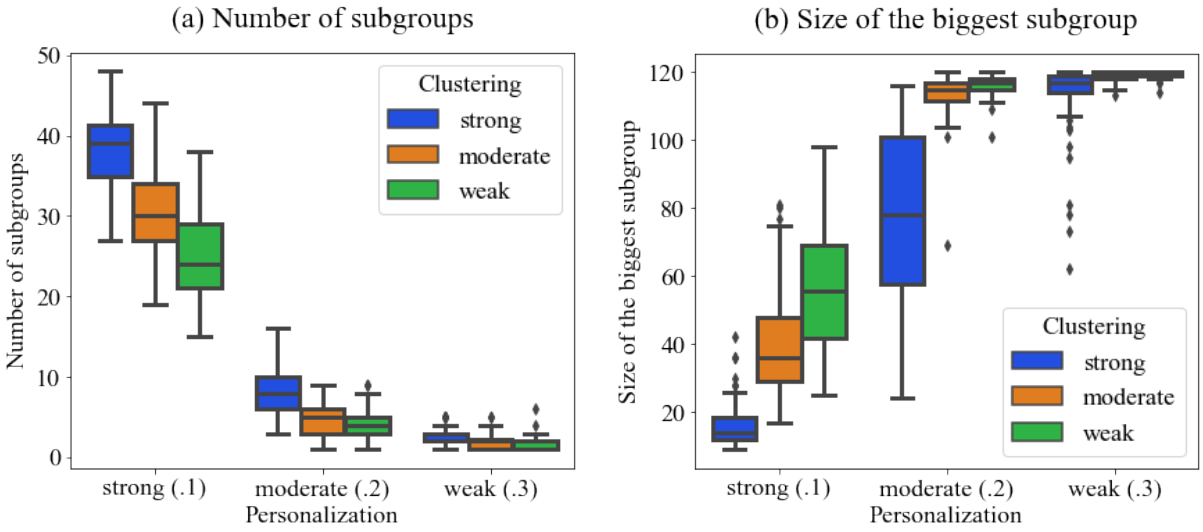


Figure 4: The effect of network clustering and personalization on opinion fragmentation measured by the number of subgroups and by the size of the biggest subgroup

1. Literature

Bar-Yam, Y. (2003). *Dynamics of complex systems*. Westview Press.

Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, *17*(4), 249–265. https://doi.org/10.1007/s10676-015-9380-y

Bruns, A. (2019). *Are filter bubbles real?* John Wiley & Sons. Retrieved from https://www.amazon.com/Filter-Bubbles-Real-Axel-Bruns/dp/1509536442

Chapin, S. (2018). Who's Living in a 'Bubble'? - The New York Times. Retrieved April 12, 2019, from https://www.nytimes.com/2018/12/11/magazine/whos-living-in-a-bubble.html

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: towards the next frontiers. *Jasss-the Journal of Artificial Societies and Social Simulation*, *20*(4). https://doi.org/10.18564/jasss.3521

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *JASSS*, *20*(4). https://doi.org/10.18564/jasss.3521

Friedkin, N. E., & Johnsen, E. C. (2011). *Social Influence Network Theory*. New York: Cambridge University Press.

Keijzer, M. A., Mäs, M., & Flache, A. (2018). Communication in online social networks fosters cultural isolation. *Complexity*, 1–20. https://doi.org/10.1155/2018/9502872

Lapowsky, I. (2019). How'd the Cohen Hearing Go? That Depends on Your Filter Bubble | WIRED. Retrieved April 12, 2019, from https://www.wired.com/story/cohen-hearing-filter-bubbles/

Macy, M. W., & Tsvetkova, M. (2013). The Signal Importance of Noise. *Sociological Methods & Research*, *44*(2), 306–328. https://doi.org/10.1177/0049124113508093

Mäs, M., & Helbing, D. (2017). Random Deviations Improve Micro–Macro Predictions: An Empirical Test. *Sociological Methods and Research*. https://doi.org/10.1177/0049124117729708

Mäs, Michael. (2018). The Complexity Perspective on the Sociological Micro-Macro-Problem. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3129362

Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, *11*(3), 279–300.

Obama, B. (2017). Farewell adress. Retrieved from https://obamawhitehouse.archives.gov/farewell

Page, S. E. (2015). What Sociologists Should Know About Complexity. *Annual Review of Sociology*, *41*(1), 21–41. https://doi.org/10.1146/annurev-soc-073014-112230

Steinmeier, F.-W. (2018). 2018 Christmas Message. Retrieved from https://www.bundespraesident.de/SharedDocs/Reden/EN/Frank-Walter-Steinmeier/Reden/2018/12/181225-Christmas-message.html

CLIFA - An Open Knowledge Base For Facts On Climate Change

Connecting Society and Experts to debunk myths and rumors around Climate Change.

Felipe Schaeffer Neves[2], Vinicius Woloszyn[1], Michael Wilmes[1], Sebastian Möller[1]
[1]Technische Universität Berlin, Germany; [2]Universitat de València, Spain
Corresponding Author: woloszyn@tu-berlin.de

There is an increasing politicization of climate science in an effort to delegitimize the scientific consensus about climate change, which is that human-caused global warming is happening. Recent studies have suggested that this is carried out by vested-interest groups through organized disinformation campaigns and fake news. These efforts can be hugely profitable for those involved and hugely harmful to society as they create a sense of confusion around climate change and divert the public's attention from the urgency of the issue. In this context, we present our work in progress for the development of an Open Knowledge Base for facts on climate change: CLIFA.

CLIFA is an open knowledge platform that will amass and store facts on climate change making them accessible to the public as well as the academic community. It will primarily work in two fronts: 1- by debunking fake news, rumours and general disinformation around climate science, and 2- by promoting knowledge amongst the public about climate change through empirical science and fact-checked evidence. In addition to these two main objectives, our platform aims to: 1- become a fertile ground for discussion around climate change and a bridge between the public and academics, and 2- build an open Knowledge Base that will allow anyone to quickly learn and share the current consensus from thousands of independent experts.

Our platform will rely on independent experts and fact-checking agencies around the world to gather its database which will be categorised and made freely available to all. Additionally, users will also be able to contribute thanks to an interactive feature which intends to incite a community environment. For example, when searching for a specific topic, people can either access the pre-existent search/results, or ask a new question in case they can not find the answer they were looking for. The new question will then be dealt with either by the experts or the fact-checking agencies who will verify and respond accordingly; the person who asked the question will then be notified.

In this, CLIFA will also support the opening of climate research to society by supporting the interdisciplinary provision of scientific facts and knowledge in a transdisciplinary context of use. Non-scientific actors are provided with a valid and situationally adapted response option in conjunction with the respective scientific communities relevant in the climate context. Conversely, the need for support by, but also the potential for misuse of, scientific facts in the political and social discourse becomes apparent in quasi real time, which also benefits self-reflection on the role of science itself. A new initiative led by the TU Berlin to establish a Berlin-Brandenburg based and supra-institutional "Evidence-Based and Action-Oriented Climate Solutions Lab" also follows this concept of societal embedded research practice.

In short, CLIFA will empower people to make evidence-based decisions and, in doing so, it can become a catalyst for discussion and comprehension of climate change, and an instrument for promoting a sustainable green transformation. Moreover, the rich Knowledge Base generated collaboratively can be used for further scientific purposes, e.g.: analysis of trends over time and popular myths on a particular topic. Thus, we hope to foment the emergence of practical solutions to the climate issue through a horizontal approach and in a collaborative environment.

On the design of a misinformation widget for messaging apps: bridging expert knowledge and automated news classification

David Arroyo Guardeño* and Sara Degli-Esposti^

* Leonardo Torres Quevedo *Information Technology and Physics Institute* (ITEFI), Spanish National Research Council (CSIC), E: *david.arroyo@csic.es*

^ *Institute of Goods and Public Policies* (IPP), Spanish National Research Council (CSIC), E: *sara.degli.esposti@csic.es*

Abstract

On one hand, decentralised systems that do not rely on the authority of a Trusted Third Parties posit the challenge of determining whether a piece of information is authentic. On the other hand, people consume more news and information coming from decentralised sources, such as social networks or messaging apps, than from centralised media such as newspapers or national television channels. Decentralisation and multiplication of types and sources of information erode our ability to discern the accurate from inaccurate information. Traditionally information quality and reliability was established based on the credibility and the reputation of the source. On social media platforms and across messaging service apps, such as Whatsapp or Telegram, attribution cannot be properly established. As a result, the curation of news data along the entire data life cycle becomes a difficult task. Clearly categorising news on the continuum from unintentionally inaccurate to intentionally misleading information remains problematic. Poor identification of non-genuine information is a serious issue that prevents the effective containment of false information.

Risks coming along these technological challenges are exacerbated in the case of vulnerable groups. Smartphone usability has increased user acceptance amongst both young and elderly people. While young people spend most time on social media platforms, elderly people tend to rely more on messaging apps, such as Whatsapp or Telegram, where users communicate with friends and family members. These relationships features strong ties and trust relationships that have an impact on the credibility of the information shared through those channels.

By focusing on the specific case of messaging Apps, in this talk we would like to examine how to tackle the problem of identifying false information through encrypted communication channels. Our reflection includes differences between the way false information propagate in messaging apps with respect to open social networks. Based on this analysis, we will explore the design implications for the construction of a misinformation widget guiding users in assessing the trustworthiness of various sources of information.

A critical aspect in the design of the widget is the identification of the best news classification tools and methodologies. To achieve this objective, on option is to rely on fact checking platforms and human experts to obtain feedback, which can be extended by leveraging the so-called wisdom of crowds and perform news curation as result of a collaborative effort among users and experts. Expert-based systems are accurate but costly and not scalable, while crowds-based systems can be biased by herding behaviour. To overcome these limitations, we can ponder the developing of automatic detection techniques by means of Natural Language Processing (NLP) and more advanced Machine Learning (ML) techniques. Nonetheless, the selection of adequate models and datasets for their tuning and training is itself a challenge.

Thus, we explore the option of adopting a so-called “human on the loop” approach, which integrates expert knowledge on fact checking and automatic detection of fake news and misinformation. Specifically, we propose a methodology that leverages fact-checking platforms to perform datasets labelling and the validation of the performance of NLP and ML tools for the automatic classification of information. We would like to obtain feedback on the feasibility of this proposal and explore opportunities of collaboration. This research is funded by the H2020 TRESCA project.

When media critics go on the offensive

Digital publicity and the populist attacks against journalists

David Cheruiyot
University of Groningen
d.k.cheruiyot@rug.nl

Abstract

Populism has today spawned cynicism of journalism, disinformation, anti-press rhetoric and even personal attacks on legacy news media and journalists. Political actors, radical right-wing movements, social commentators or generally, users of social networks, have popularised narratives disparaging the news media (Esser, Stepinska, & Hopmann, 2016; Figenschou & Ihlebæk, 2018; Mazzoleni, 2008), through delegitimising labels such as "fake news" or "liar press" or "enemy of the people[1]" (Beiler & Kiesler, 2018; Lee & Quealy, 2019; Trump, 2018). At the same time, it is acknowledged in journalism and media criticism studies that critical feedback on digital space is important in 'watching the watchdog' (Cooper, 2006). Media critics are argued to be critical for the oversight of the media, and possible alternatives to weak media accountability mechanisms such as press councils and ombudspersons. However, among the rational and constructive criticisms of social networks, journalists receive numerous offensive feedback marked by, among others, sexist, homophobic or vile remarks on social networks (Cheruiyot, 2018) that threaten the safety of journalists and press freedom (Gardiner, 2018; Löfgren Nilsson & Örnebring, 2016). Uncivil statements, personalized attacks and threats all fall in the category of pollutants of the public sphere however much they are argued to symbolize the functioning of free speech in democracies (Hayes, 2008; Reader, 2012; Santana, 2013). The aim of this study is therefore to examine how journalists of legacy news media describe and negotiate such offensive speech in digital spaces. The special focus is media criticism on social networks that journalists read in the process of news production or in seeking audience feedback. Theoretically, this paper draws from works on media criticism and metajournalistic discourse to interrogate the difficult terrain of evaluative feedback that journalists experience in the digital age. I draw the findings from in-depth interviews with 18 practising journalists in Kenya and South Africa who are active users of specifically both Twitter and Facebook. The analysis is focused on a.) how journalists describe the variety of offensive criticisms on social networks, and b.) their (re)actions to the offensive criticisms. The findings show that journalists employ a variety of discursive resistance against the offensive speech such as filtering (e.g.) blocking of uncivil users and rationalisation (reasoning with some critics). The findings are evaluated in the context of risks and challenges to professional journalism in a digital world.

Keywords: Digital media criticism, delegitimising labels, offensive speech, social networks

[1] In his regular tweets, American president Donald Trump has referred to sections of the US mainstream media as "enemies of the people".

References

Beiler, M., & Kiesler, J. (2018). "Lügenpresse! Lying press!" Is the Press Lying? In K. Otto & A. Köhler (Eds.), *Trust in Media and Journalism: Empirical Perspectives on Ethics, Norms, Impacts and Populism in Europe* (pp. 155-179). Wiesbaden: Springer Fachmedien Wiesbaden.

Cheruiyot, D. (2018). Popular criticism that matters: Journalists' perspectives of "quality" media critique. *Journalism Practice, 12*(8), 1008-1018. doi:10.1080/17512786.2018.1494511

Cooper, S. D. (2006). *Watching the watchdog: Bloggers as the Fifth Estate*. Spokane, Wash.: Marquette Books.

Esser, F., Stepinska, A., & Hopmann, D. N. (2016). Populism and the Media. Cross-national findings and perspectives. In T. Aalberg, F. Esser, C. Reinemann, J. Strömbäck, & C. H. d. Vreese (Eds.), *Populist political communication in Europe* (pp. 365-380). London: Routledge.

Figenschou, T. U., & Ihlebæk, K. A. (2018). Challenging Journalistic Authority: Media criticism in far-right alternative media. *JOURNALISM STUDIES, Advance online publication*, 1-17. doi:10.1080/1461670X.2018.1500868

Gardiner, B. (2018). "It's a terrible way to go to work:" what 70 million readers' comments on the Guardian revealed about hostility to women and minorities online. *Feminist Media Studies, 18*(4), 592-608. doi:10.1080/14680777.2018.1447334

Hayes, A. S. (2008). *Press critics are the fifth estate: Media watchdogs in America*. Westport, Conn.: Praeger.

Lee, J., & Quealy, K. (2019). The 551 people, places and things Donald Trump has insulted on Twitter: A complete list. *The Upshot*. Retrieved from https://www.nytimes.com/interactive/2016/01/28/upshot/donald-trump-twitter-insults.html

Löfgren Nilsson, M., & Örnebring, H. (2016). Journalism under threat: Intimidation and harassment of Swedish journalists. *Journalism Practice, 10*(7), 880-890. doi:10.1080/17512786.2016.1164614

Mazzoleni, G. (2008). Populism and the media. In D. Albertazzi & D. McDonnell (Eds.), *Twenty-first century populism: The spectre of western European democracy* (pp. 49-64). Basingstoke: Palgrave Macmillan.

Reader, B. (2012). Free press vs. Free speech? The rhetoric of "civility" in regard to anonymous online comments. *Journalism & Mass Communication Quarterly, 89*(3), 495-513. doi:10.1177/1077699012447923

Santana, A. D. (2013). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice, 8*(1), 18-33. doi:10.1080/17512786.2013.813194

Trump, D. J. r. (2018). There is great anger in our Country…. Retrieved from https://twitter.com/realDonaldTrump/status/1056879122348195841

"You said so!": Identifying inconsistencies in quotations in news.

Tommaso Caselli
Rijksuniversiteit Groningen
Groningen, NL
t.caselli@rug.nl

Roser Morante
Vrije Universiteit Amsterdam
Amsterdam, NL
r.morantevallejo@vu.nl

The growth of the Web has been accompanied by a flourishing of news outlets, ranging from on-line newspapers to information pages or accounts on social media such as Facebook or Twitter. Recent work has shown that on-line news consumption tends to reinforce polarisation mechanisms (Bail et al., 2018), even if users can be actually exposed to a higher diversity of news. [1] Although conclusive findings have yet to be found, it appears that the stronger the political partisanship of a news outlet, the higher the risks of misinformation and polarization effects (Fletcher and Jenkins, 2019; Rashkin et al., 2017).

As a strategy to mitigate the spread of misinformation, we propose to develop a new research agenda centered around a specific phenomenon that characterises news: inconsistencies. Specifically we aim at automatically detecting inconsistencies using Natural Language Processing (NLP). Consider the following examples [2] both attributed to Donald Trump:

1. "[...] I promise you, I will pay for the legal fees. I promise. I promise."

2. "I don't condone violence," Trump said on ABC. "I never said I was going to pay for fees."

The two statements by the same source about the "payment of legal fees" are contradictory, not because of inaccurate reporting by news outlets, but because the source (Trump) presents different perspectives on the same topic. This exemplifies a type of inconsistency.

Inconsistencies may emerge from two sources: directly from the speaker, or from the news provider(s) that reports the quote/statement. Our research focuses on (i) defining a typology of phenomena that can lead to inconsistent statements and (ii) determining how they can be detected automatically. Recent work on perspective discovery (Chen et al., 2019) is addressing related issues.

In this abstract, we focus on two phenomena: (i) *shift of stance detection*; and (ii) *shift of factuality classification*. Addressing the first would allow to measure the change in stance of a speaker in time with respect to a target topic. For instance, we can monitor the stances of targeted speakers (e.g. prominent politicians) about gay marriages over time using their quotations or reported statements. Addressing the second would allow us to measure whether the speaker has changed over time his/her commitments to the truthfulness of his/her statements, as clearly illustrated by examples 1 and 2.

To automatically detect these phenomena, NLP systems must be able to perform at least the following tasks: (i) identify speakers, quotations, and assign the quotations to the rightful speaker; (ii) determine the factuality profile of the events expressed in the quotations/statement; (iii) identify the stance of the speaker with respect to the topic of the quotations/statements ; (iv) being able to aggregate different mentions of the same speakers as well as statement about the same topic. Some of these tasks have already been framed as classification problems, such as attribution detection (Pareti et al., 2013), belief detection (Prabhakaran et al., 2015), stance detection (Sobhani et al., 2017), and coreference resolution, both at entity and at message levels. However, detecting inconsistencies requires more than performing these tasks. A systematic investigation is required in order to determine how the tasks can be integrated for inconsistency detection purposes. Additionally, benchmarks need to be created for evaluation purposes.

[1] http://bit.ly/2UPxTaU
[2] Examples are taken from http://bit.ly/2ORi5AM

References

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard Fletcher and Joy Jenkins. 2019. *Polarisation and the news media in Europe*, volume 2019. European Parliament.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

MISDOOM 2020, 2nd Multidisciplinary International Symposium on Disinformation in Open Online Media, April 20-22, Leiden, the Netherlands

Abstract for paper presentation:

Disinformation and the Russia-Ukraine Conflict: How Russian and Ukrainian news media cover fake news online

Dr Eugenia Kuznetsova, Kyiv School of Economics, Ukraine; Dr Ansgard Heinrich, Centre for Media and Journalism Studies, University of Groningen, the Netherlands

The ongoing armed conflict in Ukraine has left more than 10,000 people killed, up to 1.5 million displaced and affected lives within the entire country and region. Disinformation campaigns remain an essential element of the conflict, deepening the devastating effects of the war. Consequently, social and political environments are impacted by the spread of misinformation, disinformation and fake news.

In general, conflict-affected environments make local populations especially vulnerable to false information. Moments of political unrest, armed conflict or war cause severe disruptions to the lives of those involved. In these times, news and information become vital channels to reorganize and even safe lives on the ground. Yet, while news and information gain such crucial roles, conflict causes massive uncertainty. In such times where information is much needed (and often requested at great speed), we also witness that misinformation, disinformation and propaganda travel alongside what can be dubbed ‘quality information’.

This research paper aims to track three stories that were debunked as fake and that occured in conflict zones of Ukraine. We study how widely-used online news media outlets of Russia and Ukraine covered these three fakes, namely: 1) The fake story accusing Ukrainian Armed Forces of having shelled Vostochny district in Mariupol (a story that is till date widely believed to be true - OSCE reports do confirm however that the city was in fact shelled by Russia-backed forces, killing 29 civilians on the ground); 2) the fake story of Canada sending insurgents to fight in the Luhansk region; and 3) the fake story of the crucifixion of a boy in Sloviansk by Ukrainian soldiers. Content analysis will be used to understand how these three fake stories were dealt with in online news. By tracing the coverage of these fakes we want to discuss a) whether fakes surface in online news coverage, b) how news outlets that are thought to be influential in the formation of public opinion deal with these fake stories; and c) how these fakes are covered across news outlets and over time.

Our research is situated in the larger context of “information ecologies”. We are interested in the changing communicative patterns of the digital age and study how misinformation, disinformation and propaganda potentially influence public memory over time. While fake news is introduced via diverse information channels, disseminated and subsequently debunked, they may still persist in social consciousness, morphing and gaining new features. This research paper lays grounds for a larger investigation into the evolution of such “disinformation ecologies” and aims to enhance our understanding of long-term effects of misinformation, disinformation and propaganda.

Copycats and Hijackers: How malicious actors exploit social media hypes

Svenja Boberg & Thorsten Quandt
Department of Communication, University of Muenster

Social media offers the opportunity to put issues on the public agenda. Although traditional media still play a central role, Twitter in particular allows net-affine groups to make topics available for public discussion initially, before the raised attention is transferred to other parts of society. These social media hypes always follow the same pattern: Social topics arise in digital publics, often triggered by an event, followed by a massive circulation via social media, resulting in the traditional media taking up the issue. Groups of actors intervene in an influencing manner according to their goals, but are at the same time submitted to macro logics of diffusion.

In many cases, social media hypes immediately trigger harsh criticism and counter-movements try (sometimes automated or coordinated) to place their arguments just as prominently, to hijack hashtags or establish new frames of interpretation.
Attacks are conceivable at various levels - from individual messages in tweets or posts (e.g. by fake profiles), to coordinated behavior in social media threads or even automated social bots networks. One example is described by Grimme et al. (2017) who found an attack on the Twitter discussions on the TV debate in the run-up to the German federal elections. Here, Twitter profiles were created to dominate the hashtag "Kanzerduell" and to flush the pro AfD hashtags "HoeckeforKanzler" and "Verräterduell" into the debate. Examples of this can also be found in the international Twitter sphere, such as in #metoo, where right-wing groups tried to reinterpret the debate in the direction of sexual assaults by refugees, in order to push their own anti-migration agenda.

The present study aims to take a look at the hashtags #metoo in order to identify strategies to exploit the attention of this media hype. To this end, all tweets containing #metoo (N= 1,029,062) in the period from March 20, 2017 to March 19, 2018 drawn from the Twitter Decahose (a 10% sample from the worldwide Twitter stream) were analyzed via topic modeling and co-occurrence analysis. The timespan allowed us to detect differences over time, including counter-movements around the peak of the discussion in October 2017 and also in the long run after the attention had faded.

The preliminary results show that the phase of the first outbreak mainly attrackted counter-movements by misogynist who tried to emphasize the illegitimacy of the debate and to victimize men, who are allegedly under false suspicion. As the debate progressed, the so called “whataboutism” was replaced by further spins of the issue. New hashtags were introduced by right-wing parties to discredit refugees and link them to sexual crimes. In the further course of the study other Hashtags like #blacklivesmatter or others will be included to derive general dynamics and strategies to disrupt social media debates.

References:

Grimme, C., Assenmacher, D., Adam, L., Preuss, M., & Lütke Stockdiek, J.F.H. (2017). Bundestagswahl 2017: Social-Media-Angriff auf das #kanzlerduell? Report 2017.1, Project PropStop: 1-9. http://www.propstop.de/wp-content/uploads/2017/09/bundestagswahl-2017-social-media.pdf (14.02.2020)

Bot squads in Twitter political debates

Guido Caldarelli[1], Rocco De Nicola[1], Marinella Petrocchi[1,2], and Fabio Saracco[1]

[1]IMT Scuola Alti Studi Lucca, Piazza S. Francesco 19, 55100 Lucca, Italy name.surname@imtlucca.it
[2]Istituto di Informatica e Telematica, CNR, Pisa, Italy marinella.petrocchi@iit.cnr.it

February 14, 2020

Extended Abstract

Since a decade microblogging platforms, like Twitter, have become prominent sources of information [1], catching breaking news and anticipating more traditional media like radio and television [2]. Helped by the simple activity consisting of creating a text of 140 (now 280) characters, on Twitter we assist to the proliferation of social accounts governed - completely or in part - by pieces of software that automatically create, share, and like contents on the platform. Such software, also known as *social bots* - or simply *bots* - can be programmed to automatically post information about news of any kind and even to provide help during emergencies. As amplifiers of messages, bots can simply be considered as a mere technological instrument. Unfortunately, the online ecosystem is constantly threatened by malicious automated accounts, recently deemed responsible for tampering with online discussions about major political elections in western countries, including the 2016 US presidential elections, and the UK Brexit referendum [3–6].

Academicians make their best efforts to fight the never ending plague of malicious bots populating social networks. The literature offers a plethora of successful approaches, based, e.g., on profile- [7,8], network- [9–11], and posting-characteristics [12–14] of the accounts. However, the studies regarding detection of automated accounts rarely analyse their effective contribution in the social networks panorama. Indeed, while messages exchanged on social platforms contain a great amount of data, just a fraction of them carries crucial information for the description of the system, while the rest contributes to random noise. Thus, detecting the relevant (i.e., those not compatible with users' random activity) communication and interaction patterns is of utmost importance in order to understand which accounts, including bots, contribute to the effective dissemination of messages.

This work merges the application of the lightweight classifier for bot detection proposed by Cresci et al. in [7] with the analysis of complex networks via entropy-based null-models [15, 16]. Once we have cleaned the system from the random noise via the application of the null-model, we study the effects of social bots in retweeting a significant amount of messages on Twitter. The analysis is applied to a tweet corpus about migration in the Mediterranean Sea from North Africa to Italy.

This study has two main results: firstly, after cleaning the system from the random activity of users, we detect the main hubs of the network, i.e., the most effective accounts in significantly propagating their messages. We observe that those accounts have a number of bots among their followers (in the cleaned network) higher than average. Secondly, the strongest hubs in the network *share* a relatively high number of bots as followers, which most probably aim at further increasing the visibility of the hubs' messages via following and retweeting. Hereafter, we will refer to groups of bots that follow and retweet the same group of hubs with the term *bot squads*. To the best of our knowledge, the existence of formations of bots shared by a group of human-operated accounts has never been reported in the literature before.

References

[1] Kwak, H., Lee, C., Park, H. & Moon, S. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, 591–600 (ACM, New York, NY, USA, 2010). URL http://doi.acm.org/10.1145/1772690.1772751.

[2] Hu, M. *et al.* Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, 2751–2754 (ACM, New York, NY, USA, 2012). URL http://doi.acm.org/10.1145/2207676.2208672.

[3] Gangware, C. & Nemr, W. *Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age* (Park Advisors, 2019).

[4] Bovet, A. & Makse, H. A. Influence of fake news in Twitter during the 2016 US presidential election. *Nat. Commun.* **10** (2019). 1803.08491.

[5] Bastos, M. T. & Mercea, D. The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review* 0894439317734157 (2017).

[6] Ferrara, E. Manipulation and abuse on social media. *ACM SIGWEB Newsletter* 4 (2015).

[7] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. Fame for sale: efficient detection of fake twitter followers. *Decision Support Systems* **80**, 56–71 (2015).

[8] Badri Satya, P. R., Lee, K., Lee, D., Tran, T. & Zhang, J. J. Uncovering fake likers in online social networks. In *CIKM* (ACM, 2016).

[9] Yuan, S., Wu, X., Li, J. & Lu, A. Spectrum-based deep neural networks for fraud detection. In *CIKM* (ACM, 2017).

[10] Wang, B., Gong, N. Z. & Fu, H. GANG: detecting fraudulent users in online social networks via guilt-by-association on directed graphs. In *ICDM* (IEEE, 2017).

[11] Liu, S., Hooi, B. & Faloutsos, C. Holoscope: Topology-and-spike aware fraud detection. In *CIKM* (ACM, 2017).

[12] Giatsoglou, M. *et al.* ND-Sync: Detecting synchronized fraud activities. In *PAKDD* (Springer, 2015).

[13] Chavoshi, N., Hamooni, H. & Mueen, A. Debot: Twitter bot detection via warped correlation. In *ICDM*, 817–822 (2016).

[14] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing* **15**, 561–576 (2018).

[15] Squartini, T. & Garlaschelli, D. Maximum-entropy networks. Pattern detection, network reconstruction and graph combinatorics. *Springer* 116 (2017).

[16] Cimini, G. *et al.* The Statistical Physics of Real-World Networks. *Nat. Rev. Phys.* **1**, 58–71 (2018). URL http://www.nature.com/articles/s42254-018-0002-6http://arxiv.org/abs/1810.05095{%}0Ahttp://dx.doi.org/10.1038/s42254-018-0002-6. 1810.05095.

39 Kanishk Karan and John Gray. Memes on Pinterest gamify polarization in Canadian elections

Link to the paper:
https://medium.com/dfrlab/trudeaus-and-trudeaunts-memes-have-an-impact-during-canadian-

Abstract

The algorithm that determines Pinterest users' subsequent content appears to lead those same users down rabbit holes of increasingly hostile political memes after they click on a particularly partisan image. This is not dissimilar to discovery algorithms on other content platforms, which are designed to show a user more of the types content with which a user engages most. For the platform, this gamifies more of the user's time, but for the user, it drives at base user behavior. This piece illustrates how algorithmic confounding could increase the recommendation of homogenous partisan content.

Methodology

Our methodology relies on open-source techniques and methods to look for evidence that stands visibly and can be corroborated by the readers themselves. Using memes focused on Canadian Prime Minister Justin Trudeau, open-source analysis of how the platform's algorithm operates revealed how Pinterest may contribute to the dissemination of extremist memes as recommended popular items on its app. The DFRLab monitored Pinterest for the 15 days ahead of the Canadian election (held on October 21) until October 25 allowing for the study of electoral mis- and disinformation on the platform pre- and post-election. The results of the study showed that, by clicking on only a single hyperpartisan and often hostile meme, the platform would recommend other politically intense memes.

Results and discussions

As a visual medium, memes can act as particularly viral projections of political opinions. For that reason, they are often a particularly salient choice to propagate conspiracy theories and disinformation. When content recommendation systems cluster memes based on their political qualities, they risk compounding those harms.

Abstract for paper presentation:

Fighting Fake: Who's there to counter misinformation, disinformation and propaganda?

Dr Ansgard Heinrich, Centre for Media and Journalism Studies, University of Groningen, NL

This paper aims to explore the emerging practices to counter what has commonly been dubbed as 'fake news'. It sheds light on *who* participates in fighting fake in digitally networked spheres and sets out to characterize the many types of organizations and individuals assisting to debunk misinformation, disinformation and propaganda.

While current scholarship in media and journalism studies is paying much attention to issues of trust, credibility, and fake news (as thematic issues of journals such as *Digital Journalism* show), we are only beginning to understand how the circulation of false information impacts democratic societies – and what practices we can develop to take up the fight against fake. In addition, given the massive changes in information production and distribution over the last three decades, one cannot but attest that we are witnessing a profound disruption of our information systems. The provision of fact *as well as of* fake at the speed of light and at a level of penetration not seen before is part of this disruption. In essence, todays digital, networked, fast-paced information environments in which multiple actors participate in production and dissemination of information as well as m/disinformation, demands answers to questions such as: *Who* helps to dig for 'truth' and 'fake'? *What* strategies are employed to debunk misinformation, disinformation and propaganda and *how* do organizations or individuals try to insure the public does not only get a truthful picture, yet also (re)install trust in certain (online) information networks?

This paper develops a typology of actors involved in fighting fake in digital online spheres. These actors appear in many forms and carry various facets. They include fact-checking sites solely dedicated to debunking fake news (e.g. the Ukrainian StopFake.org operation) or educational fact-checking sites (e.g. the Dutch NieuwsCheckers platform at the University of Leiden). There are Artificial Intelligence tools created to reveal fake news (e.g. the AI tool FakerFact) or dedicated fake news sections operated by legacy news media such as the BBC or the German Tagesschau. These are just some examples of many. Throughout the paper, case studies will be used to discuss the different forms that fighting fake takes today. The paper aims to 1) profile the vast variety of fact-checking operations (including legacy news media, foundation-operated web endeavors, educational or activist platforms); 2) to characterize the different strategies employed by these fact-checkers; and to 3) understand their motives to participate in fighting fake.

As democratic societies find themselves at the dawn of a new era of information provision in which algorithms impact what people see online and where public trust appears to be no given for legacy news outlets, it appears of vital importance to develop strategies of resilience in the face of fake. This paper tries to contribute to that fight against fake by revealing the many original ways in which misinformation, disinformation and propaganda can be tackled in the digital age.

Improving the localization of hidden misinformation source in complex networks

Janusz A. Hołyst (1,2), Robert Paluch (1), Łukasz G. Gajewski (1), Krzysztof Suchecki (1) and Bolesław K. Szymański (3,4)

1. Faculty of Physics, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland
2. ITMO University, 49 Kronverkskiy av., 197101, Saint Petersburg, Russia
3. Social Cognitive Networks Academic Research Center, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180-3590, USA.
4. Społeczna Akademia Nauk, Henryka Sienkiewicza 9, 90-113 Łódź, Poland.

As the world becomes more and more interconnected, our everyday objects become part of the Internet of Things, and our lives go more and more in virtual reality, every piece of information, including misinformation, fake news and malware, can spread very fast. In order to counteract to these problems, new computer systems and algorithms, capable to track down these malicious signals have to be developed. The method of maximum likelihood estimation (MLE) solves the important case of this problem in which a limited set of nodes act as observers and report times at which the spread have reached them. Deriving the simple and closed form of estimator is possible with assumption of normal distributed time delays between nodes. While the MLE has been shown to be optimal on trees there are several challenges remaining on general graphs. One important issue is the complexity $O(N^{\alpha})$ where N is the size of the network and $3 \leq \alpha \leq 4$ depends on the network topology and the number of observers. Such a high complexity makes the method not applicable to large complex networks. We address this issue with a new approach in which observers with low quality information (i.e., with large spread encounter times) are ignored when likelihood is computed. Moreover, we limit the number of potential sources by using a gradient-like selection. Our Gradient Maximum Likelihood Algorithm (GMLA) reduces the complexity to $O(N^2 \log(N))$. We compare GMLA and MLE on Erdős–Rényi, Barabási-Albert and Gnutella networks. Although GMLA does not use information from all observers, as MLE does, it achieves better results for scale-free networks in quality of localization tests based on three measures: the accuracy, the rank of true source, and the distance error. The other issue we address here is of precision on general graphs. The original MLE approach assumes the information travels via a single, shortest path, which by this assumption is supposed to be the fastest way. We show that such assumption leads to the overestimation of propagation time in networks where multiple potential traversal paths exist. We propose a new method of source estimation based on maximum likelihood principle that takes into account the existence of multiple shortest paths – Equiprobable Paths (EPP) and Equiprobable Links (EPL). We compare these methods with MLE on Erdős–Rényi and Barabási-Albert networks as well as on a real communication network from the University of Rovira i Virgili. The results of tests show a vast improvement in precision of the method using our proposed adjustments with the difference between EPP and EPL being barely noticeable yet still consistently in favour of approach EPP.

References

1) R. Paluch, X. Lu, K. Suchecki, B.K. Szymański, J. A. Hołyst, *Fast and accurate detection of spread source in large complex networks*, Scientific reports 8 (1), 2508, 2018

2) Ł.G. Gajewski, K. Suchecki, J. A. Hołyst, *Multiple propagation paths enhance locating the source of diffusion in complex networks,* Physica A: 519, 34-41, 2019

42 Juliane von Reppert-Bismarck. Resilience to Disinformation: Gaming, TikTok, Twitch: where do European pre-teens get their news?

Full report: https://lie-detectors.org/wp-content/uploads/2019/09/JournalistsFindings_final.pdf

Young Europeans are growing up digitally highly active, politically intensely aware and largely out of sight of digital content moderators, fact-check systems and academic research. They don't consume news like their parents and teachers; they shun Facebook and Twitter, and their information-gathering happens largely in the uncharted territory of private chat groups and visual content platforms. Even as media literacy becomes a buzzword among researchers and policymakers concerned with preserving democratic institutions in an age of disinformation, the media habits of young people are increasingly diverging from those of adults. Measuring adult digital engagement with information and disinformation– predicated upon behaviour in open comment sections and on Twitter - often precludes insights into the habits and vulnerabilities of the next generation of voters. How can we measure young people's vulnerability to disinformation when we do not inhabit the same digital space?

European news-literacy project Lie Detectors has gathered data from two years face-to-face meetings between 120 journalists and 8,500 schoolchildren aged 10-15 in diverse school settings across 33 cities in Germany, Belgium and Austria. The snapshots gathered via schoolchildren's handwritten feedback on standard questionnaires after each point of contact show children are active across multiple platforms that change rapidly and attract loyalty from different age groups. Recurring anecdotal evidence from the journalists who visited them shows schoolchildren to be more familiar with platforms than with news sources; more visually literate than textually literate and vulnerable to disinformation regardless of socio-economic standing. They also show a teaching community that sees itself largely under-equipped to address themes of disinformation and source literacy in the classroom. This presentation by Lie Detectors will be based on the data analysis report "Tackling Disinformation: Journalists' Findings from the Classroom" (attached). It will highlight a unique set of quantitative as well as qualitative data from children, journalists and teachers. It will outline how this data collection approach has enabled powerful and highly visible evidence-based advocacy for media literacy.

Where relevant, we would also be glad to discuss with the gathered scientific community what the most valuable questions might be to ask children in the future in order to gain data that advance a focus on critical media literacy.

Early Fake News Detection on Twitter by analysing User Characteristics in a Tweet Propagation Path.

Konstantin Smirnov[1]

[1]Under supervision of Gerasimos Spanakis and Gerhard Weiss, Maastricht University, Department of Data Science and Knowledge Engineering.

February 13, 2020

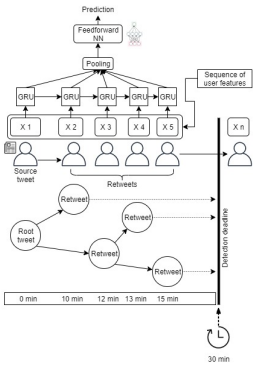
The development of social networks has given people not only a lot of opportunities, but also a lot of threats. Misinformation has shown itself to be a serious issue that can change the outcome of political events and cause damage to the economy. Therefore, detection of misinformation is most useful at an early stage of diffusion, in order to prevent harmful consequences by taking precautionary actions.

Most methods have focused on exploring the content of the news, however, content analysis has several drawbacks such as language dependency, domain-specific knowledge and common sense. Therefore, a more universal algorithm was proposed by Liu, Wu (2018), that utilizes user characteristics and their propagation paths inside a single retweet cascade. This type of data is always available at the beginning of news diffusion and can be used as a good baseline [1] [3]. As the news is further distributed, more features can be collected (such as user comments, replies, information about users who shared it, etc, in other words, more supplementary data that can improve results. However, as more time is allowed for detection, the more harmful a news piece becomes, given that its an intended false piece of information. Hence, the interest is to find a good approximation as early as possible.

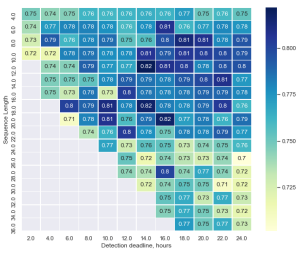
Early fake news detection research has used 2 main versions of datasets[1], covering the same news domain, that were collected by observing Tweets that shared the same URL links. However, the collected data shall emit reality as close as possible, to have robust results. It was observed that the the model proposed by Liu, Wu (2018) has shown different results on a dataset with a different data collection protocol (FakeNewsNet) [2], with less dense amount of connections. Therefore, an extension was proposed that would aggregate results from multiple sparse tweet cascades, giving out a label for a general news topic. The results have shown that it is possible to achieve 77% accuracy in less than 4 hours for entertainment news stories [2], reaching optimality at 79% after 16 hours of propagation, while for political news stories [3], the results were 82% accuracy after 16 hours, having similar early results but with more deviation. To our knowledge, we have been first to analyze early detection performance on data from different news domains. In addition, it was observed that different sets of features play an important role for early detection depending on the news domain [4].

In my presentation, I will discuss how does data quantity influences on early detection results, how does a different set of user and propagation features play their role depending on the news domain, and what changes can be done to the model proposed by Liu, Wu (2018) to improve results while having less dense cascades, keeping in mind the limitations that Twitter API gives and what data it could offer.

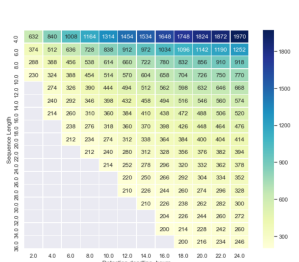
Politifact dataset: GRU Network and model input (a), accuracy measures and number of training samples (retweet cascades) for a corresponding configuration (b,c). Both figures (b,c) illustrate results for different detection deadlines (left to right), vs fixed retweet cascade size (up to down).



(a) Input data & model structure



(b) Accuracy measures.



(c) Training dataset size.

[1]Twitter15, Twitter16, Sina Weibo

[2]`gossipcop.com`

[3]`politifact.com`

[4]By comparison of feature importance, using a Random Forest Classifier

References

[1] Liu, Yang Wu, Yi-Fang. (2018). Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. *in Proceedings, The Thirty-Second AAAI Conferenceon Articial Intelligence (AAAI-18),2018*

[2] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. *arXiv preprint arXiv:1809.01286, 2018.*

[3] Twitter API. *https://developer.twitter.com/en/docs*

Bursting the Bubble

Jasper Schelling[1], Noortje van Eekelen[1], IJsbrand van Veelen[1], Maarten van Hees[1 1][0000−0002−4887−9371], and Peter van der Putten[1][0000−0002−6507−6896]

[1] ACED, Amsterdam, The Netherlands
{jasper,noortje,ijsbrand}@aced.site, vanheesmaarten@hotmail.com
https://www.aced.site
[2] LIACS, Leiden University,Leiden, The Netherlands
p.w.h.van.der.putten@liacs.leidenuniv.nl

The Bursting the Bubble initiative is a data-driven exploration of our networked news culture. Media consumption is controlled online by AI algorithms and feedback loops that have become a permanent part of social media and the hunt for clicks by news organisations. Though some claim definite evidence is lacking [1], it is generally thought this leads to polarization, echo chambers and filter bubbles [3]. In this project we seek to find ways to reverse the role of AI in this context: how can it be used to burst bubbles rather than create these, or to uncover framing and bias in publication and reporting.

We have built a large corpus of Dutch news articles (over 3.5M) across a diverse set of media, and collaborate with data scientists, designers and media professionals to carry out AI experiments and share results to evoke debate with professionals and the general public. The focus is not so much on fact-checking, but on unlocking so-called "Ideology Spaces". What we also want to show to the public is the way in which bias in current use of AI techniques contribute to polarization in the media, and how AI can also be used to counter this.

The project is a work in progress and focus so far has been mainly on data collection, but first experiments have been carried out in the areas of gamifying labelling for clickbait classification, as well as exploring the opportunities and limits of emotion classification. As a methodological underpinning, we use the Value Sensitive Design framework ([2]), a theoretically founded and iterative approach to designing technology that takes human values into account in a fundamental and comprehensive way.

References

1. Bruns, A.: Filter bubble. Internet Policy Review **8**(4) (2019)
2. Friedman, B., Kahn, P.H., Jr.: Value sensitive design: Theory and methods. Tech. rep. (2002)
3. Pariser, E.: The Filter Bubble: What the Internet Is Hiding from You. Penguin Group , The (2011)

Relevance-based Tweet Classification during Natural Disasters using BERT and User Centrality Measures

Mohamed Barbouch **Frank W. Takes** **Suzan Verberne**
m.barbouch@umail.leidenuniv.nl,{f.w.takes,s.verberne}@liacs.leidenuniv.nl
Leiden University (LIACS) The Netherlands

Natural disasters leave a disruptive impact on people and the environment[1]. To coordinate relief efforts during and after a disaster, intervention of crisis management organizations and cooperation in society are required. For a correct response, information about affected people and areas is crucial. This information is often implicitly present in the communication that takes place within the community. However, the communication lines can be broad and diverse making it difficult to infer correct and relevant pieces of information.

Social media is a rich source containing useful information for crisis response and management. However, it remains a challenge to extract the necessary information due to for example data abundance, noise, and use of informal language. Addressing this challenge with the use of an information processing pipeline such as filtering, classification, ranking, selection, and summarization can help extract the useful information [4].

In this research we aim to categorize social media messages into relevant and irrelevant classes. We apply supervised machine learning by performing multi-class text classification on Twitter data of eight natural disasters, including earthquakes, floods and storms. The original data we build upon contains several millions of tweets from various natural disasters, part of which have been labeled based on their relevance[2]. In this work, we reuse the datasets as well as the Tweet classification scheme used in [2]. The classes are given in Table 1.

These classes can be prioritized from most relevant to irrelevant. This prioritization can later help in giving preference to certain classes over others, e.g., by considering dead and injured people as more critical than broken infrastructure.

To classify the tweets, we approach the problem from two sides: by looking at the content of the tweets and by looking at the position of the users in a corresponding mention network. The main reason for doing so is that we are interested in both extracting reliable content as well as assessing the value of particular users in producing this content. Therefore, the approach utilizes elements from both text mining and social network analysis.

[1]World Health Organization: `http://apps.who.int/disasters/repo/7656.pdf`.
[2]Twitter Datasets from Crises: `https://crisisnlp.qcri.org/lrec2016/lrec2016.html`.

Table 1. The class names assigned in the labeled datasets.

1	Injured or dead people
2	Missing, trapped, or found people
3	Displaced people and evacuations
4	Infrastructure and utilities damage
5	Donation needs, offers or volunteering services
6	Caution and advice
7	Sympathy and emotional support
8	Other useful information
9	Not related or irrelevant

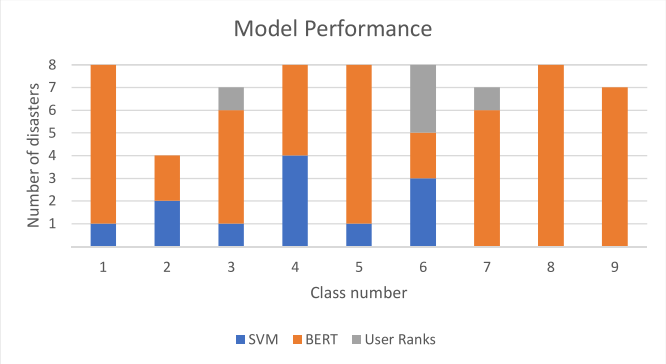


Figure 1. Number of cases in which each method was outperforming.

For the content classification, traditional classifiers (SVM, Naïve Bayes and Random Forest) with bag-of-words features are used as a baseline. We next investigate to what extent transfer learning with a large pre-trained language model can improve the text classification effectiveness. We use two BERT language models[3] from the HuggingFace Transformers library [5]: the $\text{RoBERTa}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{LARGE}}$ models[4].

To understand user influence, we extend the model with user-based network features to see whether these can provide additional information to improve the classification. The ranks are based on four user centralities: indegree, closeness, betweenness and eigenvector. For reliable values, we first determine the ranks using the (larger) unlabeled datasets, and then project them to the labeled tweets. The final model is a stacked ensemble of the best content-based model and user rank network features.

We find that BERT is outperforming traditional text classifiers when there is sufficient data to train. Against expectation, SVM did better on classes with few data. In addition, we find that for some categories of tweets, the influence of a user based on his or her social position, is of high relevance in particular classes of tweets. The relative performance between BERT, SVM and the user-based method is shown in Figure 1. In future work, performance improvements at disaster level can be achieved if the strengths of the three methods are utilized in one unified model. Overall, the proposed approach and corresponding tool can help to gain insights in relevant content and users on social media in case of crisis situations.

[3]BERT (Bidirectional Encoder Representations for Transformers), a language model proposed by Google in 2018 which has achieved state-of-the-art results on natural language processing benchmark datasets [1]
[4]RoBERTa (A Robustly Optimized BERT Pretraining Approach), an optimized version of BERT proposed by Facebook in 2019 [3].

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894* (2016).

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[4] Christian Reuter, Stefan Stieglitz, and Muhammad Imran. 2019. Social media in conflicts and crises. *Behaviour & Information Technology* (2019), 1–11.

[5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

46 Peter van Aelst, Sophie Morosoli, Edda Humprecht, Anna Staender and Frank Esser.

Resilience to Disinformation: An Experimental Study on the Spread of Online Disinformation

Online disinformation can be seen as a major threat to democracy. Recent events, such as election campaigns in different countries have demonstrated that disinformation spreads quickly on social media. However, empirical evidence regarding the spread of online disinformation and its effects on society is inconclusive, and little is known about the situation outside the U.S. However, some countries have been found to stand out as being stable, adaptive and resilient in times of social and technological transformation. In those countries, online disinformation is considered a minor problem at present because it is not widely disseminated. Thus, it can be assumed that structural conditions exist that create social resilience in the context of online disinformation. Against this background, we examine whether micro-level mechanisms work in similar manners across different countries in which social media is an important source of information. Social media is particularly suited for the dissemination of disinformation, due to the selective representation of opinions and algorithm-controlled news feeds that favor scandalous and popular content.

We use a survey embedded factorial experiment to study how the diffusion of disinformation on social media differs across and within six Western democracies (BE, CH, DE, FR, UK, US; N=6000). The experimental stimuli are Facebook posts on three different issues (climate change, immigration, vaccines) that contain false or doubtful information. We will manipulate the degree of sensationalism and the popularity of the posts to see how these features influence the individual motivation for disseminating (like, share, comment) these messages. The study will take into account both characteristics of the media environments (e.g. scale, degree of issue polarization) and individual characteristics of users (e.g. personality traits, social media behavior, attitudes towards issues) to better understand structural and personal conditions that foster resilience towards online disinformation. The data are gathered in the beginning of 2020 and we look forward to present first results at the Misdoom conference.

47 Marina Tulin, Jason Pridmore, Sara Degli Esposti and David Arroyo Guardeño. Trustworthy, Reliable and Engaging Scientific Communication Approaches (TRESCA): A Research Agenda

Post-truth, fake news and misinformation are contributing to a decline of trust in science. TRESCA is a Horizon 2020 project that focuses on understanding how we can use science communication to revert the effect of disinformation and social engineering. Using both qualitative (focus groups) and quantitative research methods (survey experiments), the TRESCA team will examine people's views on three topics related to digitalization: (1) misinformation and digital safety, (2) environ-mental health, and (3) automation and the future of work. Data collection will take place in seven European countries (Germany, Spain, the Netherlands, France, Italy, Poland and Hungary), which have been selected for their wide variety of public trust. The project will assess to what extent different science communication elements (e.g. graphic layout; links to sources; references to prestigious scientific institutions/media) affect people's willingness to believe the content they see and their readiness to act upon it. Personality traits and other individual characteristics will be included to identify people's predispositions toward (dis)trusting experts.

The ultimate aim is to develop a set of tools for improving science communication. These tools include: an animated science communication video; the prototype of a misinformation widget working on encrypted communication channels; and a Massive Open Online Course (MOOC) teaching scientists, journalists and policy makers how to improve their engagement in reliable and trustworthy science communication. The aim of the Misinformation Widget is to help people distinguish reliable from unreliable information by identifying potential biases or belief systems that make people vulnerable to disinformation and more prone to believe in conspiracy theories.

Building on previous research around public understanding of medical (i.e., anti-vaccination movements), environmental (i.e., climate change) or biological research (i.e., genetically modified organisms), TRESCA focuses on digitalization and the ways it is transforming people's interactions and everyday life. By prioritizing trust and specifically investigating the communication of social science research, this project will be able to examine how digitalization – specifically digital media – affects people's understanding and response to scientific research. TRESCA will also help the public and media experts understand how digital technologies are influencing everyday life and which technical tools we can use to establish online trustworthiness. In the process, TRESCA will help people distinguish trustworthy sources and contents from untrustworthy ones and support journalists and policy makers in learning how to better draw upon the communication of scientific research.

48 Anne Janssen.

A communication science perspective on the echo chamber debate

The world currently appears more polarized than ever: opinions are clashing about e.g., climate change, Brexit, and politics. According to echo chamber theory, today's online media environment is to blame for the polarization in society as it would have produced sets of isolated ideologically homogeneous echo chambers, in which similar opinions reinforce each other (e.g., Pariser, 2011; Sunstein, 2002). Echo chamber theory is based on both the online behaviour of the Internet user (a behavioural characteristic) and the function of algorithms (a structural characteristics): the online media environment gives individuals the unprecedented ability to selectively consume news and the algorithms curating the information environment is feared to have strengthened people's natural preference for selective exposure.

In this position paper, I argue that there are several reasons that prevent us from concluding whether the fear for echo chambers is justified. My position is based on an extensive literature synthesis covering three research domains relevant to the issue at hand: the human tendency for selective exposure (cf. the behavioural characteristic responsible for causing echo chambers), algorithmic influence on online consumption patterns (cf. the structural characteristic responsible for causing echo chambers), and online attitude polarization (cf. the feared result of echo chambers). I demonstrate that studies in the echo chamber literature show mixed findings, that they have only focused on selective exposure processes in the context of political news that has specific features not present in other contexts, and that research has had the tendency to treat the issue of echo chambers as a universal outcome that would affect all users to the same degree (Vaccari et al., 2016; Wollebæk et al., 2019). Based on the current state of the art, we can thus simply not conclude whether the widespread concern about echo chambers causing polarization is justified.

I will conclude my presentation with a discussion of the research gaps in the echo chamber literature, and what questions to ask and methods to use to close these gaps and provide a definitive answer to the echo chamber debate.

49 Thais Jorge and João Canavilhas. Will there be journalism after the fake news?

In an earthquake in Mexico (September 2017), an outbreak of fake news broke out during excavations to find survivors at Enrique Rebsamén School. There would be a 12-year-old girl – Frida Sofía – among the rubble. She would have moved her hand and asked for water. Like "Monchito", a character invented in 1985, Frida Sofía never existed: it was fake news. In Rio de Janeiro, councilwoman Marielle Franco, killed in an attack in 2018, also suffered from a wave of false information.

Journalistic practice went through different stages in its trajectory until liquid modernity (Bauman, 2000). Mutations affect the production of news, the profile of journalists and the relationship of these professionals with audiences (Adghirni and Pereira, 2011; Jorge, 2008). For some authors (Deuze and Witschge, 2017), journalism is moving from being a coherent industry to becoming a set of varied practices. For others (Kischinhevsky, 2016), the media are in constant reconfiguration, like any capitalist enterprise that intends to become viable. For Bonville et al. (2004), the transformations are part of a natural process of evolution, which goes through moments of stability and profound changes, influenced by cultural, social, legal, political, technical and economic aspects. Nowadays, beyond the desired space for free expression brought by the participation of the public in manufacturing the news, and the free expression of opinions, there were also large-scaled and orchestrated manipulations.

Foucault (1988) would say that what is happening now is "a discursive explosion". Fake News are fake stories that appear to be news stories and are broadcast on the Internet or in other media created to influence political or social opinions. We often see that part of this manipulation process is attributed to journalists. What will become of journalism? Some advocate its disappearance, others that a "new journalism", reinvented, must emerge.

In this paper we investigate what this new explosion of fake news represents for journalists. Journalists write news and their social contract is with the truth (Kovach and Rosentiel, 2004). In electoral periods mainly, does false news pose a threat to the credibility of good journalism, causing a disruption of the traditional work? Wolton (2010) advocates that, given the overabundance of information – Gaye Tuchman (1999) already told us about that – the journalist's main role in the future would be to legitimize the news. Lévy (2002) predicts that, soon, journalists will no longer be needed. The freedom of expression provided by cyberspace would lead institutions, companies and individuals to become their own media or "automedia". Kovach and Rosenstiel (2010) think that journalism should create other ways to incorporate public participation, to innovate in reports, which could be the "next journalism".

Through questionnaires, this ongoing project asks journalists from Portugal and Brazil what they think about fake news, and if they think journalism's credibility is threatened. It is also important to know what fake news is for a journalist, how they identify the different forms of fake news and why fake news is often attributed to journalists. Will there be journalism after the explosion of all kinds of misinformation and disinformation (including deep fake) that ravage reality? A pre-test was carried out in Brazil and the results published in a book in Portugal (Figueira and Santos, 2019). At this stage of the work, we are sending a form with 12 questions to Portuguese journalists to hear their ideas on the subject. Then, we will visit media organizations in Portugal and Brazil and do interviews with journalists. We will present in Misdoom an account of the state of research until the moment of the congress, as a way to contribute to the study of the phenomenon of fake news, social bots, filter bubbles, virality and disinformation as well.

50 Georgiana Udrea, Alina Bârgăoanu, Corbu Nicoleta and Gabriela Guiu. They can be fooled by fake news, but not me! Evidence of third person effect on people's ability to detect news

Fake news, already a buzzword in political communication research, is anything but a new phenomenon. Although both the concept and the practice of deceiving the public through falsified information are as old as humanity, it is the content and the scope of the term that have utterly changed in the new media ecosystem. The unprecedented contribution of digital, algorithmic and data-driven mass communication innovations, and the prevalence of the online platforms in the lives of citizens have dramatically reshaped the fairly traditional practices of disinformation.

While advances have been made in conceptualising fake news, there is relatively little empirical data on the various effects of this phenomenon. In this context, we aim at adding to the existing body of research on the effects of fake news, mainly with respect to what scholars commonly call the third person effect (TPE). Based on a survey on a national, diverse sample of adult Romanians (estimated N = 1000), we aim at contributing with empirical evidence to a better understanding of people's assessment of self, and others' ability to detect fake news. There are very few previous studies on the TPE related to people's ability to detect disinformation; additionally, there is still little evidence about the major predictors of the intensity of this effect. In our study, we aim at both confirming the TPE with regards to this particular topic, and identifying significant specific predictors, such as people's willingness (and habit) to fact-check online information, people's tendency to get trapped in echo-chambers, media consumption, or social comparison.

TPE regarding people's ability to detect fake news has profound implications on their willingness to fact-check information they come across on a daily basis. Their confidence in themselves and relative lack of confidence in others could lead them to consider that there are always the others who could be the "victims" of disinformation and need guidance and help in this respect. This type of attitude creates a vicious circle and may make the arguments about fighting disinformation through fact-checking futile. Our analysis attempts to refine and expand the scholarly discussion about the online disinformation phenomenon and its mechanisms. We aim at providing solid arguments for finding alternative ways to address the negative implications of the phenomenon, by better understanding the mechanisms explaining how and why ordinary people approach and consume fake news in the digital environment.

51 Silvia Majo-Vazquez, Mariluz Congosto, Tom Nicholls and Rasmus Kleis Nielsen.

The Role of Suspended Accounts in Political Discussion on Social Media: Analysis of the 2017 French, UK and German Elections

Information operations have been at the centre of attention of researchers, politicians and journalists. In this paper, we assess one specific type of these operations, namely the action of suspended users in the run up to elections. We analyse how suspended accounts on Twitter behaved during three of the most important elections in Europe in 2017. To this end, we borrow tools from computational social science e.g., network and text analysis and map the behaviour of over 8 million accounts identified as active in political discussion, from which almost 400,000 were suspended by the social platform.

We first use a web crawler (Author, 2018), to fetch the content of the top 9,999 distinct URLs linked by suspended and active users in each election. For each of the three countries, we build a corpus of news articles and use a tf-idf model, a standard scoring technique for keywording (Manning, Raghavan, and Schütze 2008, pp.116-121), to obtain a robust measure of the relative prominence of keywords in articles shared by the suspended accounts as opposed to the active ones.

Finally, to understand the underlying structure of the suspended users' operation, we map their behaviour by tracking all their retweets during the elections (N=1,866,670). We borrow tools from network science to identify where the amplification efforts by suspended accounts were located and the actors that most benefited from those efforts.

Our results show significant differences between the news content they shared and the patterns that shape the behaviour of suspended accounts in comparison to that of the active users. We provide evidence that suspended accounts which participated in the elections in France, the UK and Germany aimed at increasing the salience of specific political figures and divisive content related among others, to religion or immigration. We find that news content from legacy media as well as from right-wing digital-born outlets was significantly more shared by suspended accounts as opposed of content from fake news sources.

Overall, this paper contributes to the extant research from the field of information disorders (Wardle & Derakhshan, 2017) by providing novel evidence that the main goal of Twitter suspended accounts that actively participated in three major elections in Europe was spreading pro-party messages, sometimes at the extreme of the ideological dimension, and increasing the prominence of the content produced by news outlets covering divisive issues like immigration, terrorism and religion.

Our study has direct implications to assess the quality of public debate taking place on social platforms. It informs current debates on the extend to what social platforms mediate natural evolving political discussions or the interests of those who aim at amplifying divisive political issues and influence the public debate. So far, the research attention has been focused on assessing information operations at scale conducted by malicious or automated actors or unveiling the mechanisms behind the spread of mis- and dis-information online, so-called "fake news" (Neudert, Kollanyi, & Howard, 2017; Varol, Ferrara, Davis, Menczer, & Flammini, 2017). Yet, to the best of our knowledge, no previous research has looked at how users suspended by social platforms acted during elections.