Misinformation, Radicalization and Hate Through the Lens of Users

Manoel Horta Ribeiro [*1] [2], **Virgílio A. F. Almeida**[1], **Wagner Meira Jr.**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil

[2]School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

manoel.hortaribeiro@epfl.ch, {virgilio,meira}@dcc.ufmg.br

Abstract. *The popularization of Online Social Networks has changed the dynamics of content creation and consumption. In this setting, society has witnessed an amplification in phenomena such as misinformation and hate speech. This dissertation studies these issues through the lens of users. In three case studies in social networks, we: (i) provide insight on how the perception of what is misinformation is altered by political opinion; (ii) propose a methodology to study hate speech on a user-level, showing that the network structure of users can improve the detection of the phenomenon; (iii) characterize user radicalization in far-right channels on YouTube through time, showing a growing migration towards the consumption of extreme content in the platform.*

1. Introduction

In recent years, the information ecosystem has been deeply transformed. Users increasingly consume more news-pieces and opinion content in social media [Gottfried and Shearer 2016]; business models for traditional media organizations evolved [Newman 2011], and their importance as gatekeepers diminished in favor of alternative sources [Lianne and Simmonds 2013]. On darker corners of the internet, fringe websites, like *4chan*, and subreddits, like */r/TheDonald*, have great influence over which memes and news are shared in large social networks, such as Twitter [Zannettou et al. 2018b, Zannettou et al. 2018a], and often promote harassment campaigns and hateful narratives [Nagle 2017]. On social media, in the feeds of websites like Facebook and Twitter, the information users are exposed to is selected through recommendation algorithms [Liao and Fu 2013]. These black-boxes, tuned to optimize engagement, were accused of separating users from news and opinions they disagreed with [Pariser 2011].

Overall, we can identify two troublesome phenomena that were amplified by this new information environment: the dissemination of hateful content and of misinformation (or *fake news*). As we discuss later, part of the challenges we propose to approach have to do with the subjectivity of such concepts. Yet, we broadly define them:

[*]Work done while at UFMG.

- **Fake news** is a recently popularized term which refers to fabricated or excessively biased news created with the intention to manipulate, deceive or (in case of satire) entertain users [Tandoc Jr et al. 2018].
- **Hate speech** is speech that targets a group, or individuals as members of a group, causing or intending harm, often as actions beyond the speech itself. It is often expressed publicly or directly at members of the group, in a context where violent response is possible [Sellars 2016].

Although it is often hard to pinpoint exactly what constitutes either of these phenomena, their societal impact is significant. Fringe ideologies like White Supremacy got their voices amplified through online movements such as the Alt-right [ADL 2019], motivating terrorist attacks such as the one in Christchurch, NZ [Mann et al. 2019]. In the U.S., during the 2016 presidential election, researchers estimate that the average American "read and remember on the order of one, or perhaps several fake news articles" [Allcott and Gentzkow 2017]. Worryingly, researchers and the media indicate that these false pieces of information were partially driven by an orchestrated effort to promote political turmoil [Ferrara et al. 2016, Zannettou et al. 2019].

In this scenario, ways of mitigating the diffusion of hate speech and fake news are clearly necessary. Yet, there are several challenges involved with that task. Moderating this content is hard due to the sheer amount of images and comments produced every day by users in Online Social Networks [Schmidt and Wiegand 2017], due to the inherent friction with values such as freedom of expression [Rainie et al. 2017], and due to the hardness to determine what exactly is fake or what is hateful [Davidson et al. 2017], which may differ, for example, in different cultures [NW et al. 2015]. These difficulties have been obstacles for both mass-hired human moderators [Julia Angwin 2017], and for attempts to use automated techniques to characterize and detect these issues. For example, hate speech detection models fail to differentiate between harmful speech and offensive speech [Davidson et al. 2017], and several fake news detection models capture only stylistic cues, often not sufficient to tell whether something is fake [Shu et al. 2017]. Moreover, the dissemination of such information happens in an "adversarial" environment, in which agents may be spreading content to further a certain political agenda, which means that borderline cases will be exploited [Zannettou et al. 2019]. These challenges, which are largely shared in the task of tackling both hate speech and misinformation, make it logical to study these phenomena together. We argue that the development of methods to characterize and detect hate speech will further our understanding of fake-news (and vice-versa), as a big part of the challenge with dealing with both these phenomena is to deal with the elusiveness of their definition.

2. Description of Work

In this dissertation, we propose automated methodologies to characterize and detect hate speech and fake news by aggregating data at the *user level*. We argue that: *(i)* these two steps —characterization and detection— are essential parts in the larger task of mitigating these phenomena; and *(ii)* adopting a user-centric perspective —one which considers users as the central unit of study— allows to better address the aforementioned challenges. Throughout the chapters, we show examples of how focusing on users allowed us to better understand the nuances of these ill-defined but high-impact social phenomena;

to develop detection methods better suited for the real world; and, lastly, to study more complicated processes, such as the alleged radicalization of users which is happening on YouTube [Lewis 2018] —amidst plenty of hateful and fake content.

Before further describing the achievements of this work, we use hate speech detection as a motivating example for our user-centric approach. Consider the tweet: `Timesup, yall getting w should have happened long ago`. Which was in reply to another tweet that mentioned the holocaust. Although the tweet, whose author's profile contained white-supremacy imagery, incited violence, it is hard to conceive how this could be detected as hateful with only textual features. Furthermore, the lack of hate-related words makes it difficult for this kind of tweet to even be sampled in text-based approaches.

Fortunately, the data in posts, tweets or messages are not the only signals we may use to study hate speech in Online Social Networks. Most often, these signals are linked to a profile representing a person or organization. Considering this profile, we could use plenty of other information to try to determine if the *user* is engaging in hateful behavior: other tweets, their network of friends and retweets, their activity patterns. The case can be made that this wider context is sometimes *needed* to define hate speech, such as in the example, where the abuse was made clear by the neo-nazi signs in the user's profile.

This motivates our user-centered approach, which is able to take into consideration all this extra information. Characterizing and detecting hateful *users* shares much of the benefits of detecting hateful content and presents plenty of opportunities to explore a richer feature space. Furthermore, on a practical hate speech guideline enforcement process, containing humans in the loop, its is natural that content needs to be surrounded with user context [1].

The aforementioned example inspired one of the three case studies we present in the dissertation—one case study per chapter. They go as follows.

In Chapter 2 of the dissertation we explore the connections between political polarization and the spread of fake news. We investigate how polarization may create distinct narratives on what misinformation actually is. We perform our study based on two datasets collected from Twitter. The first dataset contains tweets about US politics in general, from which we compute the political leaning of each user towards the Republican and Democratic Party. In the second dataset, we collect tweets and URLs that co-occurred with "fake news" related keywords and hashtags, such as #FakeNews and #AlternativeFact, as well as reactions towards such tweets and URLs. We then analyze the relationship between polarization and what is perceived as misinformation, and whether users are designating information that they disagree as fake. Our results show an increase in the polarization of users and URLs (in terms of their associated political viewpoints) for information labeled with fake-news keywords and hashtags, when compared to information not labeled as "fake news". We discuss the impact of our findings on the challenges of tracking "fake news" in the ongoing battle against misinformation. This touches on the

[1] This is present, for example, in YouTube's [Google 2019] and Twitter's [Twitter 2019] hateful conduct guidelines. To quote directly from Twitter Rules: *Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive (...)*

already mentioned topic of the ill-definition of what is hate or fake. The content of this chapter was partially published as a workshop paper in DS+J workshop at KDD 2017. An extended version is currently under consideration for a journal.

> Ribeiro, Manoel Horta, et al. "Everything I Disagree With is #FakeNews: Correlating Political Polarization and Spread of Misinformation" arXiv:1706.05924 (2017).

In Chapter 3 of the dissertation, we explore —as mentioned in the motivating example— how it may be helpful to characterize and detect hateful *users*, rather than hateful *content*. We develop and employ a robust methodology to collect and annotate hateful users which does not depend directly on lexicon and where the users are annotated given their entire profile. This results in a sample of Twitter's retweet graph containing $100,386$ users, out of which $4,972$ were annotated. We also collect the users who were banned in the three months that followed the data collection. We show that hateful users differ from normal ones in terms of their activity patterns, word usage and as well as network structure. We obtain similar results comparing the neighbors of hateful vs. neighbors of normal users and also suspended users vs. active users, increasing the robustness of our analysis. We observe that hateful users are densely connected, and thus formulate the hate speech detection problem as a task of semi-supervised learning over a graph, exploiting the network of connections on Twitter. We find that a node embedding algorithm, which exploits the graph structure, outperforms content-based approaches for the detection of both hateful (95% AUC vs 88% AUC) and suspended users (93% AUC vs 88% AUC). The content of this chapter was published as a poster paper in the 12th AIII International Conference on Web and Social Media (ICWSM 2018). An extended version with more detailed explanation of results can also be found on arXiv.

> Ribeiro, Manoel Horta, et al. "Characterizing and detecting hateful users on twitter" Twelfth international AAAI conference on web and social media. 2018.

Lastly, in Chapter 4, we study the phenomena of user radicalization on YouTube. We analyze 331,849 videos posted on 350 channels, which we broadly classified into four types: Media, the Alt-lite, the Intellectual Dark Web (I.D.W.), and the Alt-right. According to a radicalization hypothesis widely discussed by NGOs and the media, channels in the I.D.W. and the Alt-lite serve as gateways to fringe far-right ideology, here represented by Alt-right channels. Processing 79M+ comments, we show that the three channel types indeed increasingly share the same user base; that users consistently migrate from milder to more extreme content; and that a large percentage of users who consume Alt-right content now consumed Alt-lite and I.D.W. content in the past. We also probe YouTube's recommendation algorithm, looking at more than 2mi video and channel recommendations between May/July 2019. We find that Alt-lite content is easily reachable from I.D.W. channels, while Alt-right videos are reachable only through channel recommendations. The content of this chapter was partially published as a full paper in ACM FAT* 2019. It was also one of the 100 most influential papers in 2019 (#3 in Computer Science) according to AltMetrics [2].

> Ribeiro, Manoel Horta, et al. "Auditing radicalization pathways on YouTube"Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.

[2] https://www.altmetric.com/top100/2019/

Each case study contributes to their more specific subject: misinformation, hate speech and user radicalization. Yet, altogether, they advance a central argument: that studying users rather than content itself is more productive to better understand (and eventually mitigate) ill-defined social phenomena such as hate speech and fake news.

Referências

ADL (2019). From Alt Right to Alt Lite: Naming the Hate.

Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Eleventh International AAAI Conference on Web and Social Media*.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The Rise of Social Bots. *Commun. ACM*, 59(7):96–104.

Google (2019). Hate speech policy.

Gottfried, J. and Shearer, E. (2016). News Use Across Social Media Platforms 2016. Technical report, Pew Research Center.

Julia Angwin, H. G. (2017). Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children.

Lewis, R. (2018). Alternative influence: Broadcasting the reactionary right on YouTube. Technical report, Data and Society.

Lianne, C.-F. and Simmonds, H. (2013). Redefining Gatekeeping Theory For A Digital Generation. *The McMaster Journal of Communication*, 8.

Liao, Q. V. and Fu, W.-T. (2013). Beyond the Filter Bubble: Interactive Effects of Perceived Threat and Topic Involvement on Selective Exposure to Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2359–2368, New York, NY, USA. ACM.

Mann, A., Nguyen, K., and Gregory, K. (2019). 'Emperor Cottrell': Accused Christchurch shooter had celebrated rise of the Australian far-right.

Nagle, A. (2017). *Kill All Normies: Online Culture Wars From 4Chan And Tumblr To Trump And The Alt-Right*. John Hunt Publishing.

Newman, N. (2011). Mainstream media and the distribution of news in the age of social media. Technical report.

NW, . L. S., 800Washington, S., and Inquiries, D. U.-.-. |. M.-.-. |. F.-.-. |. M. (2015). Global Support for Principle of Free Expression, but Opposition to Some Forms of Speech.

Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.

Rainie, H., Anderson, J. Q., and Albright, J. (2017). The future of free speech, trolls, anonymity and fake news online. Technical report, Pew Research Center Washington, DC.

Schmidt, A. and Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Sellars, A. (2016). Defining Hate Speech. SSRN Scholarly Paper ID 2882244, Social Science Research Network, Rochester, NY.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Tandoc Jr, E. C., Lim, Z. W., and Ling, R. (2018). Defining "Fake News". *Digital Journalism*, 6(2):137–153.

Twitter (2019). Hateful conduct policy.

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Suarez-Tangil, G. (2018a). On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, pages 188–202, New York, NY, USA. ACM. event-place: Boston, MA, USA.

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Blackburn, J. (2019). Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 218–226, New York, NY, USA. ACM. event-place: San Francisco, USA.

Zannettou, S., Sirivianos, M., Blackburn, J., and Kourtellis, N. (2018b). The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *arXiv:1804.03461 [cs]*. arXiv: 1804.03461.