Between Overload and Indifference: Detection of Fake Accounts and Social Bots by Community Managers

Svenja Boberg(✉), Lena Frischlich, Tim Schatto-Eckrodt, Florian Wintterlin, and Thorsten Quandt

University of Muenster, 48143 Muenster, Germany
svenja.boberg@uni-muenster.de

Abstract. In addition to the increased opportunities for citizens to participate in society, participative online journalistic platforms offer opportunities for the dissemination of online propaganda through fake accounts and social bots. Community managers are expected to separate real expressions of opinion from manipulated statements through fake accounts and social bots. However, little is known about the criteria by which managers make the distinction between "real" and "fake" users. The present study addresses this gap with a series of expert interviews. The results show that community managers have widespread experience with fake accounts, but they have difficulty assessing the degree of automation. The criteria by which an account is classified as "fake" can be described along a micro-meso-macro structure, whereby recourse to indicators at the macro level is barely widespread, but is instead partly stereotyped, where impression-forming processes at the micro and meso levels predominate. We discuss the results with a view to possible long-term consequences for collective participation.

Keywords: Online journalism · Community management · Moderation · Fake accounts · Social bots

1 Introduction

The emergence of participatory journalism has fundamentally changed communication between citizens, public actors, and the mass media. The quasi-permanent stream of news on the Internet is now accompanied by a multitude of participatory offerings and often by direct feedback from readers [1]. Some articles are commented on, shared, or criticized within minutes.

Along with citizens who can expand their opportunities to participate in society through participatory offerings, strategic actors are using participatory formats to place hidden propaganda through the use of fake identities. These "pseudo users" can either be operated manually in the form of fake accounts or set up (partly) automatically as social bots.

© Springer Nature Switzerland AG 2020
C. Grimme et al. (Eds.): MISDOOM 2019, LNCS 12021, pp. 16–24, 2020.
https://doi.org/10.1007/978-3-030-39627-5_2

Community managers are expected to guard these "open gates" [1] of online newspapers and carefully separate the authentic opinions of citizens from manipulative statements. Yet this responsibility carries with it the danger of either censoring public expression or allowing propagandists to abuse the reach and credibility of their own media house.

So far there are few studies that deal with the question of what criteria journalists use to distinguish between "real" and "fake" users. It is clear that journalists feel responsible for what happens in their participative channels [2]. Users classified as "fake" will most likely be excluded from the discussion. However, the criteria on which these decisions are based are hardly known.

The present study addresses this gap. With the help of expert interviews (N = 25) with selected community managers and digital editors of German national and regional online newspapers, we examined their experiences with fake accounts and social bots as well as their criteria used to classify users as "fake".

2 Identifying Features of Social Bots and Fake Accounts

The term "social bot" has lately gained a lot of media attention. It refers to a "superordinate concept which summarizes different types of (semi-) automatic agents. These agents are designed to fulfill a specific purpose by means of one- or many-sided communication in online media" [3]. A special form are political bots, which are used to spread masses of political or even propagandistic messages. Bots pretend to be ordinary citizens in order to take advantage of the supposed trust that other users in social networks have emplaced in them. However, the level of automation is difficult to assess; thus the differentiation between social bots and fake profiles, which also pretend to be normal social media users, is almost impossible. Fake profiles are often operated manually either by highly engaged online users or even paid actors. For example, hate comments are observed to be disseminated by coordinated groups that set up a series of accounts in order to spread a certain agenda [4].

Regarding the impact of social bots, the research results are somewhat mixed. While Bastos and Mercea [5] report on a Twitter botnet during the Brexit referendum that helped to spread hyper-partisan pro-Brexit messages, Neudert, Kollanyi, and Howard [6] found moderate levels of automation in Germany. Bots are also often associated with spreading spam. Badri et al. [7] show that Twitter is only able to detect the original propagators of spam, whereas retweeted networks are not blocked.

Generally, the activity of an account serves as a key criterion to detect bots. Woolley and Howard [8] classify accounts as bots if they post more than 50 tweets a day. It can be argued that frequency as the only criterion is not sufficient, since many regular accounts post as much, or programmers give their bot networks more realistic activity patterns. The botometer project [9] takes other metrics into account, such as interaction patterns of profiles, sentiments, or the reaction rates of accounts. All of these scientific approaches have one thing in common: they rely on big data analysis to detect underlying patterns—tools and procedures that normal users and forum moderators don't necessarily have access to.

3 Guarding the Gates Against Intruders: The Journalists' Need to Defend Their Platforms

Gatekeeping is one of the most studied areas of communication research. Gatekeeping deals with the question of how editorial decisions are made and how topics, events, and interpretative patterns are arranged [10]. The emergence of user-generated content has not only changed journalistic decision-making processes but also the position of traditional media in the information flow. Now citizens are able to add their views to participatory platforms curated by journalists and thus open up the public communication processes [11]. So the traditional role of journalists as gatekeepers has changed to "gatewatching" [12]. Despite the promise of increased user participation [13], participatory formats also allow for irrelevant or even uncivil content to be posted, such as attacks against other persons [14] or social groups [15].

The reason why news media still enable user comments is rooted in the journalistic role of the "press advocating for the public [and] serving as its voice in a mass-mediated society" [16]. In that regard, comments are seen as an additional tool to create a deliberative public sphere. The prevalence of veiled or even manipulative actors might damage the relationship between readers and media brands by putting off users who want to engage in a constructive discussion, as well as making journalists question the benefit of having comment sections.

As a consequence, community managers operate in a field of tension between their perceived moral obligation to permit fruitful discussions and keep out manipulative content. They have to balance the risk of letting undesirable comments slip through and repelling users who would prefer a focused discussion, *or* restricting the forum too much and thereby being accused of censorship. As a result, journalists need to develop strategies to recognize false actors in order to preserve the comments sections for their target readership. Yet little is known about how journalists perceive fake accounts and social bots, which detection criteria they use, und how they evaluate the problem.

Therefore, we state the following research questions:

1. How do gatekeepers detect fake accounts and social bots?
2. How do gatekeepers define fake accounts and social bots?
3. Do gatekeepers perceive fake accounts and social bots as a problem?

4 Method

We conducted a series of guided interviews (N = 25) that addressed community managers' detection strategies and experiences with fake accounts and social bots at German newspapers. In the following, the selection of participants and qualitative analysis are briefly described.

4.1 Participant Selection and Sample

Participant Selection. We selected our interview partners via a purposeful multi-level procedure. (For a detailed description see Frischlich, Boberg, and Quandt [17]. We considered only professional journalists [18] working at mainstream newspapers with their

own websites, that have attracted more than 100,000 unique visitors in the first quarter of 2016. In order to create a sample that most accurately represents the different regions, reaches, and editorial lines in the German newspaper landscape, a pre-study was conducted. Newspapers were rated regarding their editorial leaning, ascribed influence, and perceived trustworthiness. On that basis, the online magazines were grouped in clusters ranging from nationwide conservative and liberal, to regional newspapers and low-trust yellow journals. To represent this variability, we interviewed 50% of the newspapers within each cluster, thus ensuring that different types of media organizations were represented in our sample.

Sample. Within each selected newspaper, we approached the person responsible for social media management—that is, the digital/social media editor or community manager. The social media staff was defined as curating user comments on the newspapers' profiles on Facebook, Twitter, Instagram and WhatsApp and/or moderating the comments sections that are hosted by the online magazine itself. A total of N = 25 (10 females) interviews were conducted.

Data Collection. All interviews were carried out between January and March 2017. The interviews had an average length of 42 min (range 31–70 min). Interviews were transcribed and pseudo-anonymized. The interviews followed a pilot-tested, semi-structured guideline. Two experienced interviewers asked interviewees about (a) experiences with fake accounts, (b) definition of social bots, (c) the prevalence of social bots, (d) detection strategies, and (e) the interviewees' evaluation of fake accounts and social bots as a problem.

Data Analysis. The interview transcripts were analyzed using qualitative content analysis, following Mayring [19]. This analysis combines deductively determined pre-set categories and inductively developed categories that emerged during the initial coding of a subsample. A subsample of eight interviews was coded to develop the inductive categories and check for reliability via MaxQDA12. The coders agreed on 83–89% of the assigned codes. Disagreements were solved via discussion. After coding the whole sample following the developed category system, we used the coded interviews to identify underlying types among the comment moderators.

5 Results

In the following, the characteristics that journalists use to identify fake accounts and social bots are presented *(RQ1)*. These detection strategies are largely dependent on how much prior knowledge and experience exists with such veiled actors *(RQ2)*. We also shed light on the journalistic evaluation of social bots and fake accounts as a problem or even a threat *(RQ3)*. A total of seven types of evaluators can be identified that differ in terms of their experience, their detection strategies, and their problem perceptions.

5.1 Journalistic Detection Strategies of Fake Accounts and Social Bots

Regarding *RQ1,* forum moderators rely exclusively on their personal experience and tend to review the comments manually. Only two newsrooms in the sample have experimented with machine learning algorithms to identify undesirable content, but these methods were not yet found to be satisfying. If the moderators notice something unusual, they look primarily at the individual comment or the corresponding profile. Few consider the context of the comment, such as interactions with other suspicious content or actors.

At the micro-level of the individual comment, journalists focus primarily on topics and familiar argumentation patterns (n = 16). This can also include certain buzzwords such as "thank you, Merkel" which is often used ironically to express the harm the German chancellor allegedly has done.

> "The comment as such can be identified. Of course, this is also vague and a bit based on experience. The wording." (IV 10)

The language of the comment, such as spelling, syntax, or orthographic mistakes, is also used as a criterion, especially when Russian profiles are not set up in correct German.

After looking at the comment itself, forum moderators get a general overview of the profile, with regard to thematic focus or the amount of available information (n = 13). Posting behavior is often obvious here, especially if the profiles are monothematically oriented and similar posts appear in large numbers (n = 21).

> "They all have a certain topic, which drives them. They also interpret this in every current topic. […] That's something very idealistic." (IV 13)

Also, community managers take a look at the person behind the profile. They get suspicious if the profile has no picture or a picture that looks like a stock photo (n = 12), if the creation date of the profile is very recent (n = 6), if the relationship between followers and followees is unbalanced (n = 8), or if the profile is a member of suspicious or shady groups that the community managers have encountered before (n = 1).

> "Then you see a weird comment without a profile picture and go to the profile and there's little information or just three friends." (IV 21)

Even though most of them only look at the profile and comment itself, some also include contextual features on the macro level. These are, for example, the so-called flooding with comments.

> "There used to be one, two, kinds of hacker attacks, where we were spied […] from a […] account, where hundreds of comments came within a few minutes, which paralyzed our system for a short time." (IV 6)

Lack of interaction with other users (n = 4) and recurring profiles (n = 4) can be seen as a further indication.

> "They'll be banned and then they'll come back […], they'll be old acquaintances. You can tell by the way they express themselves, by what they call themselves. So they're not so smart that they would somehow give themselves a new name now, instead of Anton B he's called Anton C." (IV 15)

5.2 How Do Journalists Define and Evaluate Fake Accounts and Social Bots?

With respect to *RQ2*, gatekeepers do not differentiate between human-like fake accounts or automated social bots when they reflect on suspicious user profiles. When asked directly about social bots, community managers have different ideas about what they are dealing with. While some have no deep understanding at all, other journalists argue more technically, while others associate the term with a buzzword that stands for the ongoing public debate and scaremongering about the danger posed by social bots. Regarding their prevalence, all respondents have had prior experience with fake accounts, but there is a great uncertainty as to whether they are automated accounts. Here, journalists rely primarily on their feelings, but admit that automation cannot be determined without specific tools.

> "But even there, it's very difficult to determine and understand whether they're actually bots or agreed-upon people who've organized themselves somehow." (IV 14)

All in all, the respondents reported that fake accounts infiltrate public discussions, especially on political issues. On Facebook in particular, fake accounts were described as a constant phenomenon, whereas social bots were primarily attributed to Twitter. But manipulation attempts were also observed on their own forums.

In addressing fake accounts and social bots as a problem *(RQ3)*, community managers have different perceptions. Some of the respondents are not aware of the problem, or have not really thought about it yet, or are sure that the fear of social bots is exaggerated.

> "It's not like we're slapping our hands over our heads and say, 'Oh, God, how are we supposed to handle this?'" (IV 5)

Other journalists simply see themselves as not influential enough to be attractive to social bots and believe that such problems only affect the big media brands. Others, however, already see the handling of fake profiles and social bots as a problem, especially with regard to future elections:

> "I'm just afraid that this is an issue that will definitely occupy us. […] Or will occupy even more. Also now in the course of […] the Bundestag elections." (IV 10)

The results thus show that all journalists deal with the identification of veiled profiles on a daily basis, but differ greatly in the extent to which this is perceived as a problem. Based on the evaluations and experiences with fake accounts and social bots, seven types can be derived (see Fig. 1).

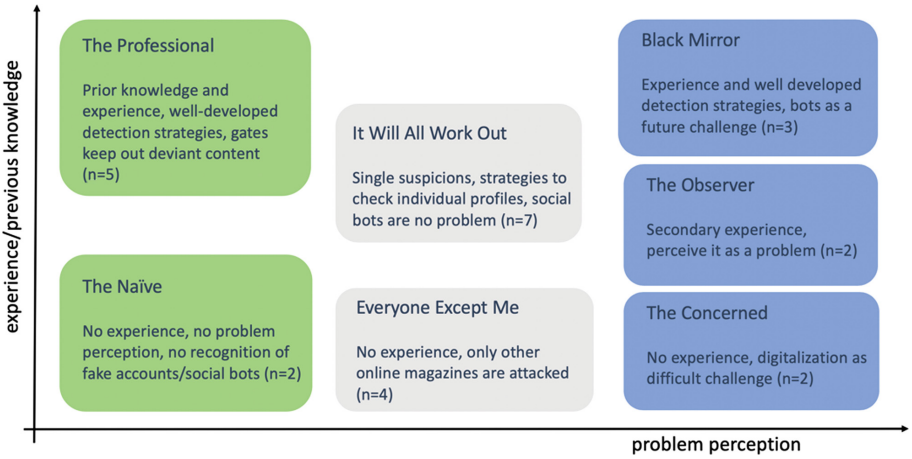

Fig. 1. Community managers' problems with perceiving fake accounts and social bots

When it comes to the detection of pseudo users, there are substantial differences in prior knowledge and competence. The two types that use the most differentiated strategies to recognize fake profiles and social bots can be contrasted by their perception of pseudo users as a problem. "The professional" have well-developed methods at their disposal, because they deal with pseudo users on a daily basis. They feel well equipped to deal with the problem. These journalists belong to large media brands that are coping with a large amount of comments and thus have institutionalized the moderation of comments to a great extent. The "black mirror" type has the same abilities, but at the same time accentuates the potential danger and the concern that the problem may be greater than is currently assumed. On the other side of the spectrum "the naïve" see no problem at all, mostly because they claim they do not have to deal with deceptive profiles apart from a few harmless fakes. They have a vague knowledge of recognition features, so it can be assumed that they also experience forms of pseudo users but simply do not recognize them. Closely related is "everyone except me" who also shows little knowledge and sees pseudo users as a problem for other magazines. The members of this type belong to smaller regional newspapers that perceive themselves as unimportant and not an attractive target for manipulation attempts. Between these extremes, the "it will all work out" type has encountered suspicious users and developed strategies to identify single profiles, but they do not fear social bot attacks and thus are confident in their ability to protect their comment sections. Lastly, "the observer" and "the concerned" both have little or no experience with pseudo users and express considerable distress. While "the observer" knows the characteristics of social bots from reports or second-hand experience and perceive them as a possible threat, "the concerned" rather refer to social bots as a buzzword and are generally skeptical about online phenomena.

6 Conclusion

The results show that experiences with fake profiles are consistent among the interviewed community managers. Without exception, all interviewees reported the prevalence of

pseudo users, even though most of them were uncertain about the degree of automation of these accounts. The basis of their judgments also varied greatly and was not bound to their own professional field, but was also fed by mass media coverage and the experiences of colleagues.

The criteria by which someone was classified as a "fake user" could be described along a micro-meso-macro structure, ranging from single comments to the overall context of a comment. The features of the comments (micro-level), the account and its digital networks (meso-level) as well as the overarching patterns (macro-level) became apparent. However, most respondents based their judgment exclusively on micro- and meso-level indicators (e.g. incorrect grammar, untrustworthy user names). Characteristics at the macro level, such as the interaction between accounts, were seldom used for impression building—although the interviewees attributed the latter with the best suitability for recognizing automated manipulation attempts. The clearly recognizable recourse to stereotypes also requires a critical reflection of the filtering processes in participative journalistic offerings.

Overall, our study provided the first empirical insights into the experiences of journalists dealing with manipulation attempts by fake accounts and social bots in Germany. It contributes to the understanding of the criteria used to separate "real" from "fake" users. The results underline the need to address this issue, as the increase of manipulative attempts in comment sections could lead to a decrease of discussion quality, resulting in either biased online discourse or even the shutdown of participatory formats entirely.

References

1. Singer, J.B., et al.: Introduction. In: Singer, J.B., et al. (eds.) Participatory Journalism: Guarding Open Gates at Online Newspapers. Wiley Subscription Services, Inc., Sussex (2011)
2. Diakopoulos, N., Naaman, M.: Towards quality discourse in online news comments. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, pp. 133–142 (2011)
3. Grimme, C., Preuss, M., Adam, L., Trautmann, H.: Social bots: human-like by means of human control. Big Data **5**, 279–293 (2017)
4. Erjavec, K., Kovačič, M.P.: You don't understand, this is a new war!' Analysis of hate speech in news web sites' comments. Mass Commun. Soc. **15**(6), 899–920 (2012)
5. Bastos, M.T., Mercea, D.: The brexit botnet and user-generated hyperpartisan news. Soc. Sci. Comput. Rev. (2017). https://doi.org/10.1177/0894439317734157
6. Neudert, L.-M., Kollanyi, B., Howard, P.N.: Junk news and bots during the German parliamentary election: what are German voters sharing over Twitter? In: COMPROP Data Memo, vol. 7, September 2017
7. Badri Satya, P.R., Satya, B., Lee, K., Lee, D., Zhang, J.J.: Uncovering fake likers in online social networks. ACM Trans. Internet Technol. 2365–2370 (2016). https://doi.org/10.1145/2983323.2983695
8. Woolley, S.C., Howard, P.N.: Social media, revolution, and the rise of the political bot. In: Routledge Handbook of Media, Conflict, and Security, pp. 282–292. Routledge, New York (2016)
9. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: BotORNot: a system to evaluate social bots. In: WWW 2016 Companion, pp. 1–11 (2016)
10. Heinderyckx, F.: Gatekeeping Theory Redux. In: Vos, T.P., Heinderyckx, F. (eds.) Gatekeeping in Transition, pp. 253–268. Routledge, New York (2015)

11. Williams, B.A., DelliCarpini, M.X.: Unchained reaction: the collapse of media gatekeeping and the Clinton-Lewinsky scandal. Journalism **1**(1), 61–85 (2000)
12. Bruns, A.: Gatewatching. Collaborative Online News Production. Peter Lang, New York (2005)
13. Vos, T.P.: Revisiting gatekeeping theory during a time of transition. In: Vos, T.P., Heinderyckx, F. (eds.) Gatekeeping in Transition, pp. 3–24. Routledge, New York (2015)
14. Gagliardone, I., et al.: MECHACHAL: online debates and elections in Ethiopia - from hate speech to engagement in social media, Oxford (2016)
15. Engelin, M., De Silva, F.: Troll detection: a comparative study in detecting troll farms on Twitter using cluster analysis. KTH, Stockholm, Sweden, 11 May 2016
16. Braun, J., Gillespie, T.: Hosting the public discourse, hosting the public. J. Pract. **5**(4), 383–398 (2011)
17. Frischlich, L., Boberg, S., Quandt, T., Boberg, S., Quandt, T.: Comment sections as targets of dark participation? Journalists' evaluation and moderation of deviant user comments, vol. 9699 (2019)
18. Weischenberg, S., Malik, M., Scholl, A.: Journalismus in Deutschland 2005 Zentrale Befunde der aktuellen Repräsentativbefragung deutscher journalisten. Media Perspekt. **7**, 346–361 (2006)
19. Gläser, J., Laudel, G.: Experteninterviews und qualitative Inhaltsanalyse als Instrument rekonstruierender Untersuchungen, 4th edn. VS Verlag für Sozialwissenschaften/Springer, Wiesbaden (2010)