Bad Robot: A Preliminary Exploration of the Prevalence of Automated Software Programmes and Social Bots in the COVID-19 #antivaxx Discourse on Twitter

Antonia Egli
Dublin City University Business School
Dublin City University
Dublin, Ireland
Email: antonia.egli@dcu.ie

Theo Lynn
Dublin City University Business School
Dublin City University
Dublin, Ireland
Email: theo.lynn@dcu.ie

Pierangelo Rosati
Dublin City University Business School
Dublin City University
Dublin, Ireland
Email: pierangelo.rosati@dcu.ie

Gary Sinclair
Dublin City University Business School
Dublin City University
Dublin, Ireland
Email: gary.sinclair@dcu.ie

***Abstract*—Health information is regularly sourced from social media platforms. However, health-related mis- and disinformation, particularly regarding vaccinations, has become increasingly prevalent on social networks since the spread of COVID-19. Automated attempts to manipulate or deceive the public by spreading false information on social media have adverse effects within the online vaccination discourse, for example by potentially converting vaccine hesitant individuals into vaccination deniers. 8,949 English-language tweets featuring the #antivaxx (i.e., anti-vaccination) hashtag generated by 7,721 discrete users in December 2020 were collected, a period when COVID-19 vaccines were first released in the United States. These were examined to determine (a) the prevalence of automated software and social bots in the #antivaxx discourse on Twitter during the focal period, (b) the prevalence of social bot use by active and visible users, and (c) the effectiveness of social network platforms to moderate misinformation. While there is evidence of use of automated software and social bots in the #antivaxx discourse during the period, such software is used by less than 1.5% of users and accounts for between 3.6% and 5.5% of the overall discourse. We also find that active users are more likely to be classified as bots than visible users. Furthermore, Twitter would seem to be effective in identifying and suspending highly active accounts associated with distributing potentially harmful information relating to vaccination.**

***Index Terms*—Twitter; Social Bots; Social Marketing; Social Media; Public Health Communications; Vaccination; #Antivaxx; #Antivaccination; Anti-vaccination.**

I. Introduction

Using social media to gather health information poses quality issues - despite its clear benefits in promoting public health information online, responding to health crises, and tracking disease outbreaks [1]–[4]. These issues arise specifically in regard to the accuracy of health information shared on social media, particularly due to a lack of regulation surrounding non-expert health sources and outdated or incomplete content [1] [5]. Research also suggests evidence of manipulative and deceptive practices [6] [7] which interfere with public health communication by creating a false sense of uniformity and validity, thus endangering public consensus and legitimising questionable or downright false information [7]–[9].

A lack of regulatory structure forces shifts in quality control responsibilities: what used to be monitored by content producers now lies in the hands of consumers' online information hygiene habits [10]. Psychological distance in virtual networks has further lowered norms of appropriate behaviour and increased the likelihood of malpractice [11]. These include tactics such as mimicking grassroots campaigns from higher-authority entities (*astroturfing*) or generating a high volume of content, replete with related hashtags and keywords, to de-emphasise or obscure some other type of activity or content (*smoke screening*) by humans or automated software programmes, i.e., *social bots* [6].

Social bots operate social media accounts and mimic human users with the aim of influencing specific online discussions. While their objective may be benign, there is substantial evidence of malicious use, for example with the aim of spreading rumors, spam, false information, slander, or noise [8]. Research suggests that social bots retweet more frequently than humans, while generating fewer replies, retweets, and mentions from human users [8]. Social bots may be used individually or in *social botnets*, which typically include large groups of bots under a single coordinator, or (*botmaster*), who coordinates their interactions to generate tweets independently of each other or within a retweet chain [13] [14]. In the context

 ISBN: 978-1-61208-869-3

of Twitter, Dickerson et al. [12] categorise such malicious bots into the following:

1) *Spambots*, which spread spam content only.
2) *Paybots*, which make money by tweeting content from accredited sources, but add links leading to sites that pay for traffic.
3) *Influence bots*, which try to sway conversations on Twitter in a specific direction.

In combination with users' increasing tendencies to turn to social media for health information, such malpractice impacts the validity of content shared and the prevalence of rumors or panic spread within a specific health-related online discourse [15]. Extant research suggests that the anti-vaccination movement has employed social bots in the past to influence narratives and decision making processes with respect to vaccination [16]. Repeated exposure to such content has been shown to result in increased hesitancy towards vaccinations, as well as a preference for finding information online rather than from accredited health care organisations [17]–[20].

Contemporaneous with the COVID-19 pandemic, the Director General of the World Health Organization (WHO) has labelled an emerging phenomenon of widespread false information surrounding the 2019 coronavirus as the "infodemic." At the time of writing, the WHO has identified over 30 discrete topics that are the subject of misinformation within the COVID-19 discourse [21]. In this short paper, we examine the prevalence and focus of automated software programmes in the COVID-19 anti-vaccination discourse. We specifically ask to which degree automated software programmes are used and examine those users who are highly active and highly visible within the December 2020 COVID-19 #antivaxx (i.e., anti-vaccination) conversation on Twitter.

The remainder of this short paper is organised as follows. Section II briefly describes the data gathered and methodology applied in three distinct analyses to examine levels of influence, types of software, and prevalence of moderation attempts within the data set. This is followed by an overview of preliminary findings in Section III, categorised by insights into generator software identified, user activity and visibility levels, and an examination of Twitter's platform moderation efforts. Finally, Section IV summarises the value of these findings and outlines elements of future research to be conducted on the subject.

II. Data & Methodology

Twitter is used widely for health surveillance and research [25], and is a popular channel within the anti-vaxx movement [9] [22]. The first release of COVID-19 vaccines took place at the beginning of December 2020. Using Twitter's enterprise application programming interface (API) platform, GNIP, we prepared a data set of 8,849 English language tweets generated in December 2020 featuring the #antivaxx hashtag. *Table 1* presents an overview of the data set.

We performed three discrete analyses in order to assess the prevalence of automated software programmes. First, we identify the most active and visible users in the data set as a proxy for influence [23]. In this case, activity is measured as the sum of tweets, retweets, and replies posted by a user, while visibility is measured as the number of retweets and replies received by a user [24]. Second, to explore the sophistication of technologies used in the discourse, we examine the type of software used to generate tweets. We use the generator metadata available from GNIP to identify the software utility that was used to post the Tweet. This metadata includes the name and a link for the source application that generated the tweet. The general public typically use official Twitter clients or other social networking platforms for cross-posting (e.g., Instagram, Facebook, etc.), while commercial actors are more likely to use social automation and other marketing automation software. The generator metadata can also provide evidence of bot applications. Third, a machine learning algorithm designed specifically for detecting social bots on Twitter, the Indiana University Network Science Institute (IUNI) Botometer, is used to identify the use of social bots by the data set's most active and most visible users. The IUNI Botometer leverages a thousand features from a Twitter account and its activity (such as astroturfing, spamming, potential self-declaration as a bot, or the account's number of 'fake followers') in order to evaluate the similarity of that account to the known features of social bots. In doing so, the IUNI Botometer reports a social bot detection accuracy in excess of 95% [8] [26].

TABLE I
Descriptive Analysis

Metric	Count	Percentage
No. of Distinct Users	7721	-
Total Tweets	8949	100%
No. of Original Tweets	2301	25.71%
No. of Replies	336	3.74%
No. of Retweets	6312	70.53%
No. of Tweets with URLs	1680	18.77%
Avg. Tweets per User	0.3	-

III. Findings

A. Generator Software

Our analysis of the generator metadata confirms our expectations in that the main software used by users in the #antivaxx discourse are official Twitter client software (97%); between 1% and 2% of end users used identifiable bots or automated software to generate tweets. However, these accounts generated between 3.5% and 5% of tweets. Approximately 45 generators (54% of all generators) were self-identified as bots or exhibited bot behaviour.

It is important to note that generator software is not a highly accurate predictor of black hat techniques. For example, some automated social botnets may use official Twitter clients, while others may be human operated, often *en masse* in click farms. The human social botnet analog, referred to as 'meat puppetry,' typically involves paid networks of real Twitter users operating under the direction of a single user who sells the network's reach for a price [13]. Such networks are extremely difficult to

 ISBN: 978-1-61208-869-3

TABLE II
GENERATOR OVERVIEW

Generator	No. of Tweets	No. of Users
Twitter Client	94.68%	97.71%
Bot	3.36%	0.86%
Third Party Twitter Client	1.21%	0.89%
Social Network	0.20%	0.18%
Other	0.54%	0.36%
Total	100.00%	100.00%

identify as the network comprises real users. These networks can be used to generate spam tweets independently of each other or as a single retweeting tree or retweet chain [14].

B. Activity & Visibility

Table 2 offers a deeper breakdown of original tweets (i.e., non-replies and non-retweets), showing the number of those posted by the data set's most active and most visible users. The percentages of original tweets published by the data set's most active and most visible users remain comparably low, indicating a widely spread online conversation with only few highly influential actors functioning as information sources.

TABLE III
ORIGINAL TWEET ANALYSIS

Metric	Count	Percentage
No. of OTs by Most Active Users	387	16.82%
No. of OTs by Most Visible Users	100	4.35%

As discussed, we analysed the most active and visible users using Botometer. Firstly, 22% of the Top 100 active users and 17% of the Top 100 visible users were either suspended or no longer available via Twitter. This typically, although not exclusively, means that the user has deleted their account or their account has contravened Twitter guidelines. Of the remaining Top 100 accounts, 21% of the most active accounts and 6% of the most visible accounts were rated as exhibiting social bot behaviour. This was largely driven by (i) self-declaration that the account was a bot, (ii) high volumes of retweeting (*echo chamber* behaviour), and (iii) a higher number of fake followers than average users. Presenting bot behaviour does not necessarily mean that the account is a bot or has malicious or malign intent. Many of the most visible and active users in our data set, while exhibiting bot-like behaviour e.g., retweeting in high volumes, are either benign, e.g., media outlets or automated news feeds, or represent high-volume Twitter users who may be passionate about countering anti-vaccination messaging through their own form of smoke screening. These include doctors and other health advocates. While outside the scope of this particular paper, the lack of anti-vaccination proponents may suggest a shy anti-vaccination supporter hypothesis and is worthy of future research.

It is important to note that the use of automated software and/or social bots should not be taken to mean that the user or their messages are anti-vaccination. Manual analysis of the most visible and active users suggests that in both cases the majority of accounts identified as behaving like bots were pro-vaccination. Of those rated with a medium to high probability of bot behaviour across both the most active (n=21) and most visible (n=5) users, only one of the accounts was actually anti-vaccination. A greater proportion of the accounts designated unavailable were true anti-vaccination supporters - seven of the 22 most active unavailable accounts were anti-vaccination and only two of the most visible unavailable accounts were anti-vaccination.

TABLE IV
BOT SCORE OVERVIEW

Bot Score	Active Users	Visible Users
Very High	9	1
High	12	5
Medium	2	6
Low	19	21
Very Low	36	51
Suspended/No Longer Accessible	22	17

C. Platform Moderation

Given the higher proportion of true anti-vaccination accounts in the unavailable accounts, and the relatively low number of anti-vaccination promoters in the most visible and most active accounts, one could reasonably posit that Twitter as a platform has been effective at moderating potentially harmful anti-vaccination messaging.

IV. CONCLUSIONS

Vaccine hesitancy is a significant contributor to avoidable deaths and disease burden worldwide. The availability of COVID-19 vaccinations presents an opportunity to control a highly transmissible disease that has resulted in a significant number of deaths as well as an unprecedented health, social, and economic burden on society worldwide. False information on social media can result in individual consumers becoming vaccine hesitant or, *in extremis*, vaccination deniers, and thereby reduce the effectiveness of COVID-19 vaccination programmes.

Few studies consider the computer as a social actor in the context of the anti-vaxx movement and the use of black hat techniques, which may influence the associated discourse. This short summary gives preliminary insights into the prevalence of automated software programmes in the online COVID-19 #antivaxx discourse on Twitter. It forms part of a wider research project (i) analysing over 24.5 million tweets generated on COVID-19 vaccination during December 2020, and (ii) comparing the wider anti-vaccination discourse on Twitter pre-COVID-19 (2018) and during COVID-19 (December 2020 onwards).

Countering the anti-vaccination movement is a significant multi-stakeholder challenge that requires active interventions by public health agencies, policy makers and professionals, pharmaceutical companies, and social media platforms themselves. Greater understanding of the different mechanisms being used by anti-vaccination promoters can help

 ISBN: 978-1-61208-869-3

pro-vaccination stakeholders mitigate the adverse effects of the anti-vaxx movement and restore faith in vaccines and vaccination programmes.

REFERENCES

[1] S. A. Moorhead et al., "A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication," Journal of Medical Internet Research, vol. 15/4, e85, 2013.

[2] C. Chew and G. Eysenbach "Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1120 outbreak," PloS One, vol. 5/11, e14118, 2010.

[3] Y. Zhao and J. Zhang, "Consumer health information seeking in social media: a literature review," Health Information & Libraries Journal, vol. 34/4, pp.268-283, 2017.

[4] P. Rutsaert, Z. Pieniak, A. Regan, A. McConnon, and W. Verbeke, "Consumer interest in receiving information through social media about the risks of pesticide residues," Food Control, vol. 34, pp. 386-392, 2013.

[5] I. C. Sinapuelas and F. N. Ho, "Information exchange in social networks for health care," Journal of Consumer Marketing, vol. 36/5, pp. 692–702, 2019.

[6] M. Kovic, A. Rauchfleisch, M. Sele, and C. Caspar, "Digital astroturfing in politics: Definition, typology, and countermeasures," Studies in Communication Sciences, vol. 18/1, pp.69-85, 2018.

[7] T. Lynn, P. Rosati, G. L. Santos, and P. T. Endo, "Sorting the Healthy Diet Signal from the Social Media Expert Noise: Preliminary Evidence from the Healthy Diet Discourse on Twitter," International Journal of Environmental Research and Public Health, vol. 17/22, 8557, 2020.

[8] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," Commun ACM, vol. 59/7, pp. 96–104, 2016.

[9] D. A. Broniatowski et al., "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate," American Journal of Public Health, vol. 108/10, pp.1378-1384, 2018.

[10] X. Lin, P. R. Spence, and K. A. Lachlan, "Social media and credibility indicators: The effect of influence cues," Computers in human behavior, vol. 63, pp. 264–271, 2016.

[11] C. E. Naquin, T. R. Kurtzberg, and L. Y. Belkin, "The finer points of lying online: Email versus pen and paper," Journal of Applied Psychology, vol. 95/2, pp. 387-394, 2016.

[12] J. P. Dickerson, V. Kagan, and V. S. Subrahmanian, "Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?", Advances in Social Networks Analysis and Mining, IEEE/ACM International Conference, pp. 620-627, 2014.

[13] D. M. Cook, B. Waugh, M. Abdipanah, O. Hashemi, and S. A. Rahman, "Twitter deception and influence: Issues of identity, slacktivism, and puppetry," Journal of Information Warfare, vol. 13, pp. 58–71, 2013.

[14] S. Stieglitz, F. Brachten, B. Ross, and A. K. Jung, "Do social bots dream of electric sheep? A categorisation of 147 social media bot accounts," Proceedings of Australasian Conference on Information Systems, pp. 1-11, 2017.

[15] J. P. Allem and E. Ferrara, "Could social bots pose a threat to public health?", American Journal of Public Health, vol. 108/8, pp. 1005-1006, 2018.

[16] V. S. Subrahmanian et al., "The DARPA Twitter bot challenge," Computer, vol. 49/6, pp. 38-46, 2016.

[17] M. J. Smith and G. S. Marshall, "Navigating parental vaccine hesitancy," Pediatric annals, vol. 39/8, pp. 476-482, 2010.

[18] A. M. Jones et al., "Parents' source of vaccine information and impact on vaccine attitudes, beliefs, and nonmedical exemptions", Advances in Preventive Medicine, pp. 1-8, 2012.

[19] D. Jolley and K. M. Douglas, "The effects of anti-vaccine conspiracy theories on vaccination intentions," PLoS One, vol. 9/2, e89177, 2014.

[20] E. Dubé, M. Vivion, and N. E. MacDonald, "Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications," Expert Review of Vaccines. vol. 14/1, pp. 99–117, 2015.

[21] World Health Organisation, "Coronavirus disease (COVID-19) advice for the public: Mythbusters", 2020. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters [retrieved: June, 2021].

[22] J. Ruiz, J. D. Featherstone, and G. A. Barnett, "Identifying Vaccine Hesitant Communities on Twitter and their Geolocations: A Network Approach." Proceedings of the 54th Hawaii International Conference on System Sciences, p. 3964-3969, 2021.

[23] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy." Fourth international AAAI conference on weblogs and social media, pp. 1-8, 2010.

[24] B. K. Chae, "A complexity theory approach to IT-enabled services (IESs) and service innovation: Business analytics as an illustration of IES," Decision Support Systems, vol. 57, pp. 1–10, 2014.

[25] L. Sinnenberg et al., "Twitter as a tool for health research: a systematic review," American Journal of Public Health, vol. 107/1, e1-e8, 2017.

[26] C. A. Davis, O. Varol, O. E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," Proceedings of the 25th international conference companion on world wide web, pp. 273–274, 2016.

[27] T. Anderson and H. Kanuka, E-research: Methods, strategies, and issues. Boston, MA: Pearson Education, 2003.

[28] D. L. Olson and G. Lauhoff, "Descriptive data mining," In: Descriptive Data Mining (pp. 129-130). Springer, Singapore, 2019.

[29] L. Silver, C. Huang, and K. Taylor, "In emerging economies, smartphone and social media users have broader social networks." Pew Research Center, 2019.

[30] Royal Society of Public Health, "MOVING THE NEEDLE: Promoting vaccination uptake across the life course," 2018. Available at: https://www.rsph.org.uk/static/uploaded/3b82db00-a7ef-494c-85451e78ce18a779.pdf [retrieved: June, 2021].

[31] S. Blume, "Anti-vaccination movements and their interpretations," Social Science & Medicine, vol. 62/3, pp. 628-642, 2006.

 ISBN: 978-1-61208-869-3