

Be Careful Who You Follow: The Impact of the Initial Set of Friends on COVID-19 Vaccine Tweets

Izabela Krysińska¹, Tomi Wójtowicz¹, Agata Olejniuk¹,
Mikołaj Morzy¹, and Jan Piasecki²

¹Poznan University of Technology, Poland

²Department of Philosophy and Bioethics , Faculty of Health
Sciences , Jagiellonian University Medical College, Poland

izabela.krysińska@doctorate.put.poznan.pl

tomi.wojtowicz@doctorate.put.poznan.pl

agata.olejniuk@put.poznan.pl

mikolaj.morzy@put.poznan.pl

jan.piasecki@uj.edu.pl

October 3, 2021

Abstract

Although Twitter is regarded as one of the most potent sources of vaccine-related disinformation, relatively little is known about how Twitter constructs timelines for individual users. In this work, we examine the composition of the Twitter timeline conditioned on the initial selection of friends. We illustrate our method by analyzing how the initial selection of friends impacts the number of pro- and anti-vaccination information present in the timeline. Our experiment clearly shows the disproportionate prevalence of anti-vaccination content seeping into Twitter timelines even for accounts initialized with explicitly pro-vaccination friends. We also discuss ethical considerations of using automated bots for research purposes in a social network environment.

1 Introduction

Over the recent years, Twitter has become one of the leading channels of disinformation dissemination. The success of the global fight against the SARS-CoV-2 pandemic depends on the efficacy of vaccination campaigns, yet anti-vaccination propaganda is constantly being spread on Twitter. This spread can be amplified by internal Twitter recommendation algorithms constructing timelines, and the details of these algorithms remain hidden from the public.

The contents visible to users, i.e., the Twitter timeline, are generated by a sophisticated algorithm that tries to maximize users' engagement with the service. The details of the algorithm responsible for creating the timeline are not publicly available; thus, the observation remains the only experimental method useful for the public. In this work, we are using the Twitter API to read snapshots of the timeline. We must stress that the composition of the timeline visible via Twitter API is not necessarily identical to the composition of the timeline visible to regular users of the service. According to the Twitter API documentation, "*The endpoint returns a collection of the most recent Tweets and Retweets posted by the authenticating user and the users they follow.*" Our method retrieves the entire timeline accessible to a bot at a given point in time. This does not reflect the pattern in which regular Twitter users access the service. In the remainder of the paper, we make the simplifying assumption that the timeline represents the stream of tweets a regular human user would see. Still, one should remember that there may be considerable differences between the view exposed via Twitter API and the view exposed via the regular service interface.

Due to tweets' brevity and domain-specific language (hashtags, URLs, acronyms, neologisms), it is relatively hard to distinguish between human-generated and bot-generated content. Unfortunately, the contents of tweets and their origin are not the only things contributing to disinformation dissemination. Behavioral events, like re-tweeting, following, un-following (both genuinely originating from humans and resulting from the algorithmic procedure built into bots) can be easily mistaken for clues of social acceptance and support for fringe ideas and anti-scientific stances.

In this paper, we introduce a new method of analyzing the social impact of Twitter via experimental observation using passive bots. We create multiple bots that periodically check their timelines and download all tweets contained in those timelines. The bots execute in the same time period, and their only difference is the composition of the initial seed set of followed accounts (i.e., friends). Although we use the live Twitter ecosystem, our method is almost non-intrusive. The only potential impact of our bots comes from social signals

produced by the fact that the bot starts following an account.

The experimental scheme presented in this paper can be easily extended to cover any contentious and divisive subject discussed on Twitter. As the use case proving the viability and usefulness of our approach, we choose to examine the number of pro- and anti-vaccination information presented to a user conditioned on the initial set of friends. We create five bots that significantly differ in the initial seed set of followed accounts, and we allow our bots to access Twitter’s timeline a few times a day during the period of one week. We carefully examine the profile of tweets observed during this period and draw conclusions about the composition of timelines.

Despite recent interest in trustworthy and ethical artificial intelligence, ethical considerations related to research methodologies are often overlooked in computer science. To encourage other researchers to take ethics seriously, we extend our paper with an entire section dedicated to ethical considerations of using automated bots to conduct experiments in a real social network environment, with human users being a part of the experiment.

2 Related Work

From the very start of the SARS-CoV-2 pandemic, Twitter has been scrutinized as one of the primary engines for spreading disinformation about the pandemic, vaccines, and public health^{8,17,18}. Our work draws inspiration from³, where the authors employ a similar technique to gather observational data from Twitter. The main difference is that the authors focused on the social network analysis, especially on how social interactions such as mentioning and re-tweeting differ in groups of anti-vaxxers and pro-vaxxers rather than analyzing personal feeds in these groups.

Applying social network analysis methods to Twitter conversations around the pandemic is a popular research direction^{7,12,15}. Although very interesting, we feel that these works miss an important point. They treat Twitter timelines as the source of data, and they focus on the behaviors displayed by users in response to the data. Our approach stresses the fact that Twitter itself is the source of bias and an important trigger for users’ behaviors.

Our research tries to address the problem of rising hesitancy towards vaccination in the face of a global pandemic. We draw inspiration from works such as¹⁹ where the authors analyze many tweets and identify major themes and topics discussed in anti-vaxxer Twitterverse. Our findings seem to confirm the hypothesis that merely exposing users to opposing views may lead to further polarization. Bail *et al.* present an experiment measuring the impact of the exposure to opposite views on liberal/conservative Twitter

users¹. We find a very similar effect concerning vaccines, but the asymmetry is much stronger than in the case of political views.

Robert Gorwa and Douglas Guilbeault created a functional and historical typology of bots⁶. This is not an ideal typology since at least three types are hardly distinguishable from one another. Gorwa and Guilbeault differentiate between:

- crawlers and scrapers whose primary function is to index the content of the Web,
- chatbots that are a form of computer interface allowing human users to interact with a machine,
- spambots, a malicious kind of bots that created unwanted, rubbish messages,
- social bots, also known as sybils, are pieces of software producing and spreading malicious content on social networks. Social bots interact with human users on social media, can mimic them, and be used to promote certain content; social bots can be used for good and bad goals, e.g., activists can use them to promote political participation (good goal) or to spread political misinformation (bad goal),
- sockpuppets and trolls, which are automated social media profiles with fabricated identities that impersonate real users and interact with other human users; they differ from social bots because they allow different degrees of human supervision,
- cyborgs and hybrid accounts are another type of human and bot cooperation, which again can be used to promote valuable content on social media, or to produce, promote or spread malicious content

Peter M. Kraft *et al.*⁹ propose a typology of bot's intervention that is based on certain aspects of bot's characteristics. Intervention on actions consists of a bot or bots behaving in a certain way, for instance, up-voting (share or like) certain content to test users' reactions to the "up-voted" content. Intervention on attributes requires changing some attributes of a bot's profile. For example, the gender of a bot retweeting certain content may be indicated as either female or male to test the reaction of human users. Finally, the intervention on algorithms requires differentiation between different bot's action patterns (for instance, modifying the frequency of retweeting the same contents to measure human engagement).

3 Methods

3.1 The architecture of Twitter bots

We build several bots that share the same operational procedure. The only thing that differentiates the bots is the composition of the initial seed set F_0 of Twitter accounts that each bot follows. Let us introduce the notation used throughout the paper.

Given a subject S and a tweet t , we use the notation $t \in S$ if the tweet concerns the subject S , and $t \notin S$ if the tweet t does not concern the subject S . Furthermore, we denote the fact that the tweet t expresses a positive attitude towards the subject S by $t \in S_+$. Neutral and negative attitudes are denoted by $t \in S_0$ and $t \in S_-$, respectively.

Our framework requires the training of two classifiers. A binary classifier $C_S(t)$ decides whether the tweet t concerns the subject S or not. The ternary classifier $C_A(t, S)$ decides the emotional attitude towards the subject S expressed by the tweet t .

The procedure followed by each bot is presented in Figure 1:

- The bot is created with the seed set F_0 of friends
- The bot "wakes up" and accesses n tweets from its current timeline T
- For each accessed tweet $t \in T$ the classifier $C_S(t)$ is applied to check if $t \in S$
 - if $t \in S$ and t type $k_t \in \{RETWEET, QUOTE, LIKE\}$, then author a of the retweeted, quoted or liked tweet is added to F (the bot starts following the author a)
- The bot "falls asleep" for the time period τ and becomes inactive
- After the predefined number of days λ , the bot halts. We apply the classifier $C_A(t, S)$ to all the tweets $t \in S$ which the bot has ever accessed to measure the percentage of pro-, neutral, or anti-vaccination tweets.

3.2 Data Annotation

Our system uses two different classifiers. Classifier $C_S(t)$ is used to decide whether a given tweet is in any way related to the topic of COVID-19 vaccines. The primary aim of our experiment was to measure how does selective exposure affects the polarization of views about a single topic — vaccines. Thus, we have reduced the impact of non-vaccine-related content and taught

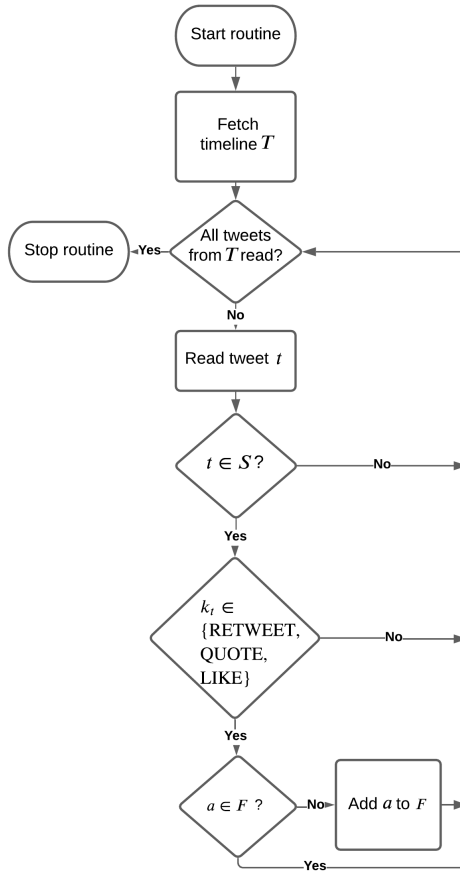


Figure 1: Twitter bot routine

the bot how to identify tweets related to vaccines. All other tweets were discarded from further analyses. The topic classifier trained to identify vaccine-related tweets was trained as a binary classifier with two exclusive classes: whether a tweet is related to vaccines or not.

To train the topic classifier $C_S(t)$, we have created a dataset containing positive examples (tweets related to COVID-19) and negative examples (other tweets). For training the classifier, we have used open-source datasets. The first dataset consisting of tweets related to COVID-19 vaccines is a collection of CoVaxxy⁴ tweets from the period 14-17.01.2021. The second dataset containing tweets unrelated to vaccines consists of tweets from different sources: celebrities dataset², gender classification dataset¹¹, Russian trolls dataset^{14,16}, and events dataset²⁰. We have selected a sample of 200 000 tweets with the distribution of 10% positive class (tweets related to COVID-

19) and 90% negative class (tweets not related to COVID-19). The classifier has been trained using `TextCategorizer` from the SpaCy ¹ text processing library using a convolutional neural network. The classifier reached almost perfect classification with $AUC = 0.995$. Still, one should remember that the classification task was straightforward: practically all vaccine-related tweets contained tokens from a small vocabulary (covid, vaccine, jab, vax, cov19, az, pfizer) and were tagged with particular hashtags (#vaccine, #covidiot, #stayhome, #covid19), so the classifier had no problem in memorizing these features.

The second classifier $C_A(t, S)$ was trained to discover the attitude expressed in the tweet towards the subject of COVID-19 vaccines. The training of this classifier has proven to be much more challenging. First, we have manually annotated over 2500 tweets using the following annotation protocol. We have assumed the existence of three classes:

- **PRO**: positive, the tweet unequivocally suggests support for getting vaccinated against COVID-19
- **NEUTRAL**: the tweet is mostly informative, does not show emotions vs. presented information, contains strong positive or negative emotions but concerning politics (vaccine distribution, vaccine passports, etc.)
- **AGAINST**: the tweet is clearly against vaccination and contains warnings, conspiracy theories, etc.

Tweet annotation has been conducted using Prodigy². The annotators were provided with the following instructions:

- Do not spend too much time on a tweet and try to make a quick decision, the slight discrepancy in labeling (especially if you are deciding between **PRO** and **NEUTRAL**) will not affect the classifier significantly.
- Assign tweets that seem to originate from news sites as **NEUTRAL** and use **PRO** for tweets that express unequivocal support for getting the vaccine.
- There are many tweets on vaccination and politics. They should fall into the **NEUTRAL** class unless they contain a clear call to action: go get vaccinated!

¹<https://spacy.io>

²<https://prodi.gy>

- Use only the contents of the tweet to label it, do not open the links if the content of a tweet is not enough for labeling (e.g., “Hmm, interesting, <https://t.co/ki345o2i345>”), skip such tweets instead of giving it a label.
- Use the option to skip a tweet only when there is nothing in the tweet except for an URL or a few meaningless words, otherwise do not hesitate to put the tweet in the **NEUTRAL** class.

Below we present selected examples of tweets and their respective labels.

- **PRO** tweets

- If you think that NHS staff should receive full COVID-19 vaccine courses (2 doses) immediately, please RT this 🙏 NHS staff are not currently allowed 2 doses. Not even staff who are shielding with medical conditions, who want to return to their frontline jobs 😞
- 48 hours later, and I’m happy to report that my mom has experienced zero side effects. <https://t.co/S0emF9iKLV>
- Got that vaccine today 💪💉

- **NEUTRAL** tweets

- The problem we have is that it won’t be the GP that will give the vaccine to the care home residents. It has to go to the care home or something.
- Actually, resigning abruptly as HHS secretary in the middle of a surging pandemic and a vaccination crisis doesn’t seem all that responsible. At this point, Azar should be working his ass off until the last minute.
- Trump administration accused of deception in pledging release of vaccine stockpile <https://t.co/MD8stxY110> via @Yahoo
- Inside Track: Behind instantaneous clearance of Bharat Biotech’s corona vaccine <https://t.co/KtYRqQGuk5>

- **AGAINST** tweets

- If there are 23 in the tiny population of Norway, how many are there in the US that are not being publicly revealed? Just askin’.
- Now they are forcing vaccines on we the people. We will no longer be in charge of our own bodies. <https://t.co/Ag0zyweFLS>

- So the AstraZeneca vaccine was trialed in Brazil, South Africa & the U.K. & now we have virus mutations from these areas? Is this not being analyzed by the media? Anyone? <https://t.co/TMKLKJ2BF3>
- 1/ @cdcgov had released its updated batch of reported Covid vaccine side effect data through Jan. 8, when roughly 6.6 million doses had been administered. And it is... not good.

Our goal was to have an obvious distinction between **AGAINST** and **PRO** classes, where the former expresses explicit support for vaccination, and the latter tries to discourage the reader from getting vaccinated. All tweets that had negative emotional valence (e.g., tweets criticizing a government of incompetent handling of vaccine distribution) were classified as **NEUTRAL**, even if one could deduce from the tweet that an author supports the vaccination. Thus, in the following sections, the **PRO** class contains only the tweets which explicitly encourage people to vaccinate, and all other tweets, even if implicitly supporting vaccination efforts, were classified as **NEUTRAL**.

We have asked 8 annotators to annotate the same set of 100 tweets using the guidelines proposed in the annotation protocol to verify the annotation protocol. We have measured the inter-rater agreement using the Fleiss' κ coefficient⁵. The results were as follows:

- when measuring the agreement with four possible classes (**PRO**, **NEUTRAL**, **AGAINST**, **NONE**, where the last class represents tweets that were rejected from annotation), the agreement is $\kappa = 0.3940$
- when measuring the agreement after removing tweets that were rejected, the agreement is $\kappa = 0.3560$
- when measuring the agreement if rejected tweets are classified as **NEUTRAL**, the agreement is $\kappa = 0.3753$
- when measuring the agreement for only two classes (using **PRO**, **NEUTRAL** and **NONE** as one class, and **AGAINST** as another class), the agreement is $\kappa = 0.5419$

According to a popular interpretation of Fleiss' κ ¹⁰, the annotators are in fair agreement in the first three scenarios and moderate agreement in the last scenario. These results suggest that the annotators are struggling to distinguish between **PRO** and **NEUTRAL** classes, and sometimes they have divergent opinions on whether the tweet should be rejected from training.

Still, they are coherent when labeling **AGAINST** tweets. Since the main goal of our experiment is to validate how much disparaging information (explicitly advocating against the vaccines) is presented to a Twitter user, we conclude that the agreement between annotators is sufficient to produce a high-quality training dataset for the classifier.

The final classifier has been trained on 2000 manually annotated tweets related to COVID-19 vaccines. The training dataset contained 366 examples of pro-vaccine tweets, 637 examples of anti-vaccine tweets, and 1119 examples of tweets that were neutral. As in the case of the first classifier, we have used the `TextClassifier` from SpaCy library, training the model with dropout $d = 0.4$ for 100 iterations using compounding batch sizes. The baseline AUROC was 0.470, and the training managed to improve the overall AUROC to 0.729. The AUROCs for **PRO**, **NEUTRAL**, and **AGAINST** classes were 0.706, 0.679, and 0.802, respectively. Even though the training set was imbalanced, both classes of interest (**PRO** and **AGAINST**) are recognized with precision and recall sufficient for the main goal of the study.

3.3 Seed Accounts for Twitter bots

This section describes our methodology for selecting accounts that will be used to create the initial set of bot’s friends. The point of departure is creating two lists of polarized seed accounts. The first list consists of vaccine advocates, and the second list consists of vaccine skeptics. We define vaccine advocates as a group of people who produce content that urges people to get vaccinated, describe the positive effects of vaccines, debunk conspiracy theories on COVID-19 vaccines, or present a negative attitude to vaccine hesitancy. On the other hand, vaccine skeptics create content that warns others about getting the vaccine, share information about exaggerated side effects following COVID-19 vaccines, and disseminate conspiracy theories.

We queried the Twitter search engine with manually curated hashtags such as `#coronavaccine`, `#getvaccinated`, `#mRNA`, `#PfizerGang`, `#VaccineNoThankYou`, `#vaccinesWork`, `#BillGatesVaccine`, `#VaccinesKill`, etc. to fetch tweets related to COVID-19 vaccines. Then we have searched for tweets with conspicuous emotional load, both negative and positive. As a result, we have identified 24 Twitter accounts with a positive attitude towards COVID-19 vaccines (vaccine advocates) and 50 Twitter accounts with a negative attitude towards COVID-19 vaccines (vaccine skeptics).

Then, we have manually curated 5 configurations (one configuration per bot) of the initial seed set of friends. The parameters of bots are presented in Table 1. As can be seen, the bots differ in the ratio of anti-vax to pro-vax, starting with a bot that has befriended only anti-vaxxers (Anthony) and

ending with a bot that has befriended only pro-vaxxers (Jessie). Unfortunately, we could not control variables such as number of followers, number of comments, or tweeting frequency across configurations and Twitter user groups (advocates versus skeptics). The selection of initial friends for each bot configuration was random, and we have not prioritized friends with many followers or friends producing more tweets.

Table 1: Twitter bots’ initial configuration

Bot name	Anti-vax ratio	#anti-vaxxers	#pro-vaxxers
Anthony	1.0	10	0
Alice	0.9	9	1
Julia	0.5	5	5
Alfred	0.1	1	9
Jessie	0.0	0	10

4 Results

Our bots were active for the period of 5 days (from May 18, 11:00 a.m. to May 22, 5:00 p.m.) and were following the procedure outlined in Section 3.1. The period τ of inactivity of the bot has been set to one hour. Table 2 presents the results of the entire experiment³. The results below do not include tweets for which COVID-19 relevance was computed below 0.50. The attitude classifier has been re-trained on filtered data following the suggestions from¹³. Columns marked with "avg" represent the average score per tweet for each class.

Table 2: Results of the experiment: number of tweets, number of friends, and the composition of individual bots’ timelines

Bot name	# tweets	# covid tweets	# friends	PRO	NEUTRAL	AGAINST	avg PRO	avg NEUTRAL	avg AGAINST
Anthony	1967	1033	772	2%	63%	35%	12%	49%	39%
Alice	392	244	40	1%	58%	41%	11%	47%	42%
Julia	2118	1084	803	2%	63%	35%	13%	49%	38%
Alfred	1594	620	794	5%	81%	14%	19%	55%	26%
Jessie	682	296	240	4%	90%	6%	18%	61%	22%

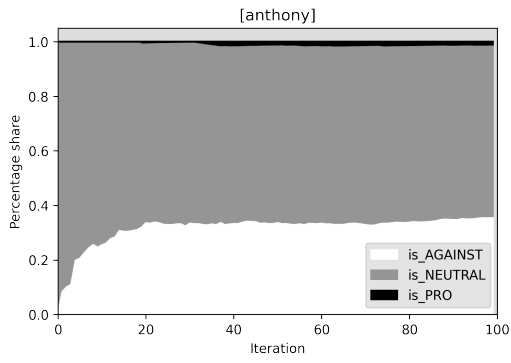
³The small number of tweets received by Alice is due to the fact that Twitter has rejected this bot after a few days of operation from adding new friends, we could not identify the reason behind this suspension.

The results presented in Table 2 depict the final view of the timeline, i.e., the analysis of the entire set of tweets "visible" to the bot during the entire experiment. Equally interesting is the temporal view of the change in the composition of the bot's timeline. Below we present the composition of all bots' timelines in each iteration. This view allows us to observe the gradual change of the profile of information presented to each bot. We find that any significant changes to the composition of the timeline happen only during first iterations, when bots start to extend the set of followed accounts, but (quite surprisingly) the compositions of timelines very quickly reach the stable state, irrespective of the initial composition of the timeline.

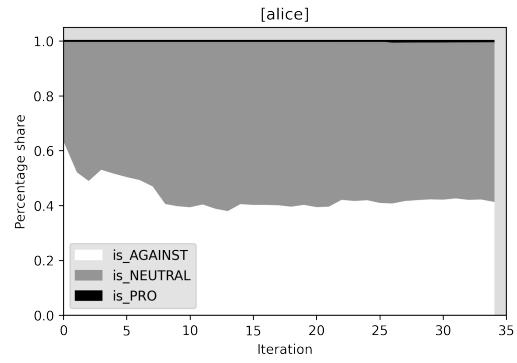
5 Discussion

Firstly, we want to stress that the framework of experimental observation of Twitter concerning divisive and contentious topics presented in this paper is not limited to the subject of vaccines. It can be readily adapted to every topic under discussion in the Twitterverse. The only action required for such adaptation is the training of two classifiers: the topic classifier responsible for recognizing that a tweet is relevant to the topic and the attitude classifier responsible for discovering the attitude of the tweet to the topic. Soon, we are planning to examine in detail the impact Twitter has on the public discussion on COVID-19 related topics such as wearing masks, lockdowns, and vaccination passports. We also plan to extend our inquiries into other topics which define current public discourse, such as racism, climate catastrophe, gender equality, or acceptance of sexual minorities.

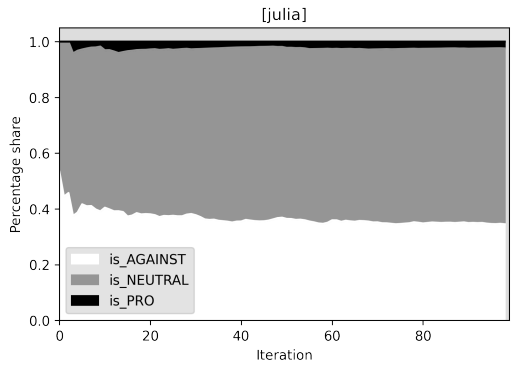
The results of the experiment are depicted in Table 2. For each bot, we report the number of tweets in the bot's timeline, the number of tweets related to COVID-19, and the number of friends gathered during the experiment. We also present aggregated composition of bot's timeline in terms of tweets that are **PRO**, **NEUTRAL**, and **AGAINST** the vaccines. To provide more nuanced insight into the timelines, we also report on the averaged responses of the $C_A(t, S)$ classifier responsible for evaluating the attitude expressed towards vaccines by a tweet. When classifying a tweet t , the response of the classifier is the distribution over three classes. For instance, the classifier might predict that there is 70% that the tweet is pro-vaccine, 18% that the tweet is anti-vaccine, and 12% that the tweet is neutral. We choose the label with the highest probability to be the final label of the tweet. However, this approach may not be representative when classifier responses are fuzzy. The average class assignment of tweets in bot's timeline are presented in columns **avg PRO**, **avg NEUTRAL**, and **avg AGAINST**.



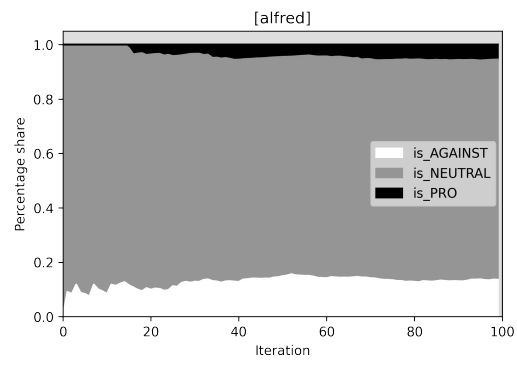
(a) Anthony's timeline



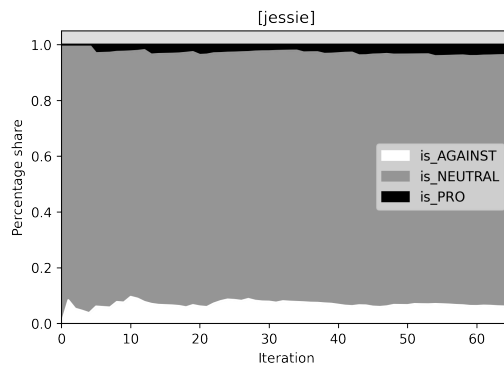
(b) Alice's timeline



(c) Julia's timeline



(d) Alfred's timeline



(e) Jessie's timeline

Figure 2: The timeline compositions.

The results of our experiment are quite discouraging. We can see that even for bots that were seeded with exclusively (or almost exclusively) pro-vaccination friends (Jessie and Alfred), the percentage of tweets expressing the unwavering support for vaccination against COVID-19 is tiny, not exceeding 5%. It quickly drops to 1%–2% for bots, allowing for a more "balanced" seed of friends. If a bot starts with strongly anti-vaccination friends (as is the case with Anthony and Alice), the Twitter timeline will aggressively strengthen and affirm the anti-vaccination stance by presenting a large amount of anti-vaccination material. Unfortunately, this anti-vaccination material manages to seep into timelines of bots that start with pro-vaccination friends, and the amount of anti-vaccination propaganda is substantial. Even if a bot does not have a single anti-vaxxer friend, she will still see 6% of anti-vaccination tweets, and by allowing just 10% of anti-vaccination friends, the number of anti-vaccination tweets grows to 14%. We see the lack of symmetry between bots seeded with pro-vaccination friends (Alfred and Jessie) and bots seeded with anti-vaccination friends (Anthony and Alice). Trying to "balance the opinions" leads to a disaster. Julia (who starts with half pro-vaxxer friends and half anti-vaxxer friends) sees only 2% of pro-vaccination tweets compared to 35% anti-vaccination tweets.

The timelines of all bots are fairly similar. However, we can observe small departures in percentage shares of both the pro-vaccination and anti-vaccination tweets, with the majority of tweets belonging to the **NEUTRAL** class. The stability of the composition of timelines, which we observed after just a few iterations, may be interpreted as yet another effect of information bubbles created by Twitter. It seems that each bot is quickly reaching its information niche and does not experience significant shifts in the distribution of the polarity of information.

Interestingly, Anthony's timeline (starting with a seed of 100% anti-vaccination friends) was flooded with anti-vaccination tweets after the first 20 iterations. We do not observe a similar pattern in the case of Jessie, who started with a seed of 100% pro-vaccination friends. The observed disproportion between the shares of pro- and anti-vaccination tweets may suggest that anti-vaccination friends are much more active than pro-vaccination friends.

6 Ethical considerations

Bot activity in a social network can be considered as a research intervention. Such an intervention alters the "natural" informational environment of a social network available to users. The ultimate goal of the bot's research intervention (BRI) is to gain generalizable knowledge on either human be-

havior or the results of a social network algorithm that may ultimately affect human behavior. Therefore, the BRI does not necessarily meet the definition of research involving humans provided by the US Common Rule.

According to the US. Common Rule Research involving humans requires:

- obtaining information about an individual through intervention or interaction with individual and using, studying or analyzing that information(CFR46.101), or
- obtaining, using, studying, analyzing, or generating identifiable private information. (CFR46.101),

Intervention is understood as physical procedures interacting with a body of human subjects and alternations of subjects' environment (CFR46.101). In comparison, interaction is understood as communication or interpersonal contact between subject and researcher (CFR46.101). These definitions are important not only from a legal perspective but also from an ethics perspective. Research involving human beings may require additional ethical scrutiny.

According to the bot classification introduced by Gorwa & Guilbeault⁶, our bots best fit into the category of social bots. However, their goal is not to interact or to target other users' behavior. The bots are fully automated. They do impersonate human users, but they do not communicate with other users, and the only action they perform is to follow other users based on the pre-programmed algorithm. Using Krafft's *et al.* classification⁹, our bots test an intervention on algorithms: they investigate how initial configurations of Twitter friends impacts the exposure to the COVID-19 vaccine-related content.

The bots employed in the current project did not involve any human subjects. We also did not collect any data associated with individuals. It is reasonable to conclude that our bots did not involve human beings. Nevertheless, the bots were active in the human information environment, and we should carefully assess possible ethical perils of the research. The most obvious ethical challenge is the use of deception.

The experimental bots act under the disguise of normal human users. Therefore, there is an element of deception: other users, especially those followed by the experimental bots, might think that real human beings actually follow them. However, there is a rationale not to disclose the experimental bot's full identity. Namely, we assume that if the bot's identity would be revealed: the profile would be informed that this bot tests exposure to misinformation on COVID-19 vaccines. The experimental bots would be blocked

by those Twitter users who oppose vaccines. These practical concerns alone do not suffice to use deception. We deem that deception is justified by: the goal of the study, minimal intrusiveness of the bots, transparency, and leaving space for autonomous choice.

- *The goal of the study* is to describe Twitter bias towards misinformation about COVID-19 vaccines. The COVID-19 is a serious global health threat that can be mitigated only by public health interventions that require massive participation. Mass vaccination against COVID-19 is one of the most effective and economically promising solutions to stop the spread of the Sars-Cov-2 virus, which is responsible for the pandemic. Understanding how misinformation about COVID-19 vaccines is spreading in one of the globally most important social networks is paramount.
- *Minimal intrusiveness*: The bots do not produce any content. Their interference into a social network is limited to two factors: their visibility as separate “users” and their action of following other accounts. Twitter will notify users who are followed by the bots that they have a new follower. A human user may indeed infer the false information that there is someone interested in her tweets. We mitigate this risk by providing explicit information in bots’ profiles (exposing the fact that these are bots) and limiting the experiment to a relatively short time span. The interaction between human users and our bots is minimal and does not produce any unusual or additional risk.
- *Transparency and debrief*: The actions of our bots are described ahead of launching the project at the project website ⁴, and this information is disseminated by the project Twitter profile @webimmunization.
- *Space for autonomous choice and opt-out option*: The bots intervention still leaves space for users to make an autonomous decision if they want to be followed by the experimental bot. A human user who concludes that information contained in the bot’s profile is insufficient, too general or suspicious, has the freedom to block the bot.

7 Conclusion

In this paper, we present a method for evaluating the composition of the Twitter timeline conditioned on the initial selection of friends when engaging

⁴www.webimmunization.cm-uj.krakow.pl

in divisive and contentious topics of public discourse. As an example of such discourse, we analyze social media interactions related to the COVID-19 pandemic and the public sentiment towards vaccines and vaccination programs. We observe a strong asymmetry between the anti-vaccination and pro-vaccination sides of the discourse. While bots initialized as highly pro-vaccination received mostly neutral tweets in their timelines, bots initialized as highly anti-vaccination received firm support for their anti-vaccination stance. People who are convinced of the efficacy and usefulness of COVID-19 vaccines do not tend to enforce this stance among other pro-vaccine users. At the same time, anti-vaxxers seem to strengthen and fortify their message among other anti-vaxxers constantly. Of course, the procedure executed by our bots is a simplification of real human interaction with the Twitter service. Nevertheless, we feel that our results can be projected onto non-bot timelines and that these results provide an interesting insight into the effects of Twitter on the public discourse.

Acknowledgments

The research leading to these results has received funding from the EEA Financial Mechanism 2014-2021. Project registration number: 2019/35/J/HS6/03498. We would like to cordially thank the members of the #WebImmunization project for helping with data annotation.

References

- [1] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.
- [2] Raad Bin Tareaf. Tweets Dataset - Top 20 most followed users in Twitter social platform, 2017.
- [3] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 130–140, 2020.
- [4] Matthew R DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Christopher Torres-Lugo,

- Kai-Cheng Yang, Filippo Menczer, and John Bryden. Covaxxy: A collection of english-language twitter posts about covid-19 vaccines. *arXiv preprint arXiv:2101.07694*, 2021.
- [5] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- [6] Robert Gorwa and Douglas Guilbeault. Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, 12(2):225–248, 2020.
- [7] Man Hung, Evelyn Lauren, Eric S Hon, Wendy C Birmingham, Julie Xu, Sharon Su, Shirley D Hon, Jungweon Park, Peter Dang, and Martin S Lipsky. Social network analysis of covid-19 sentiments: application of artificial intelligence. *Journal of medical Internet research*, 22(8):e22590, 2020.
- [8] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraittem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3), 2020.
- [9] Peter M Krafft, Michael Macy, and Alex” Sandy” Pentland. Bots as virtual confederates: Design and ethics. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 183–190, 2017.
- [10] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- [11] Data For Everyone Library. Twitter user gender classification, 2015. data retrieved from Kaggle Datasets, <https://www.kaggle.com/crowdfLOWER/twitter-user-gender-classification>.
- [12] Elena Milani, Emma Weitkamp, and Peter Webb. The visual vaccine debate on twitter: A social network analysis. *Media and Communication*, 8(2):364–375, 2020.
- [13] Martin Müller and Marcel Salathé. Addressing machine learning concept drift reveals declining vaccine sentiment during the covid-19 pandemic. *arXiv preprint arXiv:2012.02197*, 2020.
- [14] NBC News. Russian troll tweets, 2018. data retrieved from Kaggle Datasets, <https://www.kaggle.com/vikasg/russian-troll-tweets>.

- [15] Paola Pascual-Ferrá, Neil Alperstein, and Daniel J Barnett. Social network analysis of covid-19 public discourse on twitter: Implications for risk communication. *Disaster medicine and public health preparedness*, pages 1–9, 2020.
- [16] Ben Popken. Twitter deleted 200,000 russian troll tweets. read them here. *NBC News*, 14:273, 2018.
- [17] Hans Rosenberg, Shahbaz Syed, and Salim Rezaie. The twitter pandemic: The critical role of twitter in the dissemination of medical information and misinformation during the covid-19 pandemic. *Canadian journal of emergency medicine*, 22(4):418–421, 2020.
- [18] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*, 2020.
- [19] Mike Thelwall, Kayvan Kousha, and Saheeda Thelwall. Covid-19 vaccine hesitancy on english-language twitter. *Profesional de la información (EPI)*, 30(2), 2021.
- [20] Arkaitz Zubiaga. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984, 2018.