REPORT 11

Detection of Bots Focused on Political Polarization on Social Networks: An Approach Through Speculative Design

[ANONYMIZED AUTHOR]
[ANONYMIZED INSTITUTION / PROGRAM]
[ANONYMIZED ADDRESS]
[ANONYMIZED CONTACT]

Abstract

This paper explores the application of Speculative Design Methodology for detecting polarizing bots on social media. Using an approach grounded in workshops and article analysis, the study was structured into three notebooks: ”Where are we?”, ”Where are we going?” and ”Where do we want to go?”. Notebook 1 involved analyzing articles and using tools such as Innovation Map and the Cone of the Future to map the current scenario and identify emerging trends. Notebook 2 projected a future reconfiguration of the scenario, considering the evolution of technologies and their implications. Finally, Notebook 3 proposed an innovative solution, the ”Anti-Virus,” an advanced AI designed to infiltrate and neutralize botnets, addressing the shortcomings of current detection methods. The paper concludes with a critical analysis of limitations and suggestions for future developments, emphasizing the need for integration with social media APIs and legal reviews.

1. Introduction

Digital Culture is a powerful movement regarding technological progress, currently corroborated by the intense consumption of digital products and media. According to [Bortolazzo 2020], this evolutionary tendency of humanity toward the digitization of information results in two important aspects related to how information is handled and obtained nowadays. The first aspect is decentralization in the transmission of knowledge. Today, a wide variety of content can be accessed from anywhere and at any time, in various formats. This creates a phenomenon in which the technological apparatus itself is capable of producing new knowledge. The second aspect highlighted by the author is the miniaturization and mobility of mobile devices. The transformation of these devices into lightweight and portable tools significantly altered the ways in which we acquire, consume, and access knowledge. In this way, knowledge, previously restricted to fixed media such as printed books and newspapers, adapted to the context of contemporary mobility.

Following this line, one of the most relevant and accessible forms of information transmission in Digital Culture is Online Social Networks (frequently abbreviated as OSN, from the English term “Online Social Networks”), which constitute an extremely influential type of social media. [Gatkal et al. 2021] even states that “social media are currently omnipresent in our daily lives”, a fact that, according to the researchers, facilitates communication with massive audiences, and makes this type of virtual environment attractive to companies, politicians, and news channels.

In the context of OSNs, which are the category of social media whose main purpose is to create and strengthen networks of personal or professional connections, such ease, despite presenting great potential for beneficial purposes, contains its risks. Taking Twitter as an example, which has a vast amount of data on diverse subjects, many users assume the network as a reliable means of obtaining and disseminating information, through which these users form their opinions consciously or unconsciously from news and content shared by third parties.

The presence of “Official” users, such as political representatives, famous personas, and very well-known organizations strengthens this trust in the medium, as well as the possibility of exchanging information with any other person around the globe. However, this practice can be extremely harmful. According to [Cai et al. 2017], illegal users frequently use social bots to manipulate public opinions, spread rumors, and produce false classifications or ratings on social networks. [Khattak et al. 2014] expands the possible concerns by pointing out that bots may have purposes such as information gathering, cyber fraud, malware dissemination, cyber warfare, unsolicited marketing (spam), and disruption of network services. Therefore, malicious bots can bring negative effects to public and individual security. They may even influence users’ views regarding their political options in elections of their respective countries, as investigated by [Gatkal et al. 2021], [Ceron-Guzmán and León-Guzmán 2016].

It becomes fundamental, therefore, to create and invest in ways to protect the OSN public from this type of manipulation. In this way, several researchers have already proposed means of identifying bots and botnets on social media, as well as methods of suppressing rumors and fake news that undermine their influence. The major challenge in these cases is to find optimized ways to analyze the large volume of data present in networks, because its exponential size and its constant state of updating are characteristics that hinder the process. Furthermore, social bots become more similar to real users each day, making the complexity level to define their behavior patterns higher.

In the area of Information Systems, this type of proposal usually fits into areas such as Cybersecurity and Network Science. Community identification, a tactic that some researchers use to identify bots and their respective families, as proposed by [Gatkal et al. 2021], is a recurring topic in the Networks area, while several taxonomic terms used to refer to botnet structures and their detection tend to be present in cybersecurity literature, as well as ways of dealing with their components.

In this work, through the Speculative Design method, we will present the current scenario of the theme in question, some possibilities of futures and trends associated with this theme, and a proposal to mitigate the possible harms brought by political polarization botnets in the future.

Thus, this article is divided as follows: in Section 1, we have this Introduction. Section 2 will bring the Foundation of terms relevant for better understanding of the text in the remaining sections. Section 3 presents the Methodology used to conduct this study. Section 4 brings a Mapping of the current state of the theme; in Section 5, we have the Speculation of Possible Futures; and in Sections 6, 7, and 8 we will have, respectively, the Projection of the Desirable Future, the Conclusion, and the References.

2. Foundation

In this section, the adopted definitions of some specific terms present in the body of the article will be presented, so that there is better understanding of the text. All described terms were extracted and defined based on the literature present in the references of this work.

Online Social Network/OSN: It is a group of people, or more precisely, their digital representations, that are connected through relationships derived from data about their activities, shared contacts or direct links collected in internet-based systems. [Gatkal et al. 2021]

Bot or Social Bot: They are software programs designed to perform automated tasks on social networks, often operating independently, without the need for continuous human intervention. [Barhate et al. 2020]

Botnet: A network of compromised machines, or accounts, (bots) that receive and respond to commands from a command and control (C&C) server. [Khattak et al. 2014]

C&C: Server that serves as a rendezvous mechanism between the Bots and the Botmaster. [Khattak et al. 2014]

Botmaster: Human being who controls the Botnet. [Khattak et al. 2014]

Bot Binary: It is a malicious executable program that infects a machine (or an account in this case) and turns it into a bot within a Botnet. [Khattak et al. 2014]

A simplified scheme of the structure of a Centralized Botnet is presented in Figure 1, to better elucidate the terms described so far:

Figure 1. Simplified Botnet Scheme

Botnet Detection: Detection of all components of a botnet, including the botmaster, C&C server(s), C&C means and (all or a subset of) bots. [Khattak et al. 2014]

Bot Detection: Detection of machines, or accounts, infected by botnets. [Khattak et al. 2014]

Bot Family Detection: A bot detection class focused on finding sets of bots belonging to the same Botnet. [Khattak et al. 2014]

Community: A group of nodes that are more densely interconnected among themselves than with the rest of the network. [Gatkal et al. 2021]

3. Methodology

The Methodology used in this case was entirely based on Speculative Design concepts. From the definition of the theme and a Speculative Design workshop taught at **[ANONYMIZED FOR BLIND REVIEW]**, articles about the subject were read to generate a more grounded contextualization of the theme. Then 3 notebooks were answered following the following lines of reasoning: "Where are we?", "Where are we going?" and "Where do we want to go?". For record purposes, these notebooks are based on situating the chosen subject in current and future scenarios.

Notebook 1 was answered from the prior analysis of scientific articles on the theme, as well as the elaboration of a visual scheme of the current scenario to situate the present context. The Innovation Map tool was also used to identify trends in this research area and generate speculations about possible technological developments that can both improve and hinder the identification of polarizing bots. Finally, the Cone of the Future was used to elucidate possible situations inherent to the theme.

In Notebook 2, the trends and possibilities pointed out in Notebook 1 were used to speculate a possible scenario reconfiguration in an interval of 15 years in the future, if there are no interventions in current trends. At this point, the Cone of the Future proved extremely useful to help catalyze the main ideas developed up to this moment of the work.

Finally, in Notebook 3 the possible negative implications of the proposed scenarios were analyzed and a new form of solution for the problem in question was suggested. In this specific case of Detecting Polarizing Bots, although the initial theme was limited to detection, the suggested IT solution went a bit beyond, also being capable of containing the operation of malicious social bots.

The conclusion of the document was written based on the analysis of the union of all notebooks cited above. However, there was also the addition of possible failures in the speculations made, technical difficulties and ethical issues that may affect the development of the suggested solution, as well as possible biases and what could be done as future work from this.

4. Mapping the Current State of the Theme

4.1. Current Context

Figure 2. Current Scenario

In this section, we will explain how the current situation of the theme stands, presenting not only a macro scenario, but also the State-of-the-Art of polarizing bot detection methodologies up to the present work.

Starting with the scenario, in Figure 2, we have 3 groups of components separated by colors: The blue group corresponds to components belonging to the botnet; the purple corresponds to the environment where interactions occur, as well as components that connect both to real users and to automated users; and finally, pink is the group where human components are found, the users and those responsible for social media platforms.

In summary, from the interest of a stakeholder, a botnet is created or adapted with the specific objective of manipulating public opinion. In this context, the Botmaster strives to reach such objective using their bot families, which they command through the Command and Control Server (C&C). To avoid being discovered through conventional tracing techniques, the Botmaster also resorts to Stepping-Stone techniques.

Once the botnet is coordinated, automated users infiltrate social media, disseminating information or amplifying the reach of existing publications, which alters human users’ perception about the information displayed on the network.

To prevent mass alienation and other harms caused by bots, methods for detecting these accounts are developed. The most common methods identify individual bots or, in more advanced cases, bot families through comparison of posting patterns and divergent behavioral characteristics between human and automated users.

However, some researchers highlight that, despite the countless research efforts developed in the area, the official adoption by social media of detection methods is still relatively low. This suggests resistance on the part of development teams and other professionals responsible for managing these networks in implementing new detection technologies. As a result, human moderators, inefficient [Cresci et al. 2017] due to the massive volume of data and the evolution of the mimicry technique, are frequently used to analyze posts and interactions, while a growing number of malicious automated accounts continues to emerge in this virtual environment. The variety of social bot types also continues to increase over time, since each social bot is a product of human imagination [Arin and Kutlu 2023].

Furthermore, another aggravating fact about this theme is that in few cases the developed systems focus on detecting the human responsible for the botnet. This translates into impunity for botmasters and brings the possibility of restructuring for botnets already identified and contained previously. However, thanks to concealment strategies of these components, this would be a considerably more difficult approach, being normally explored in sectors such as cybersecurity.

Considering this context, the detection of political polarization bots on social networks, especially on Twitter, has benefited significantly from deep learning-based methods, as shown in recent studies [Cai et al. 2017], [Arin and Kutlu 2023], [Belokurov et al. 2021]. One of the most notable architectures identified in the study in question, developed by [Arin and Kutlu 2023], consists of three long short-term memory (LSTM) models and a fully connected layer, used to capture diverse data from a Twitter account. In this model, a user’s tweets are analyzed by an LSTM to capture the semantics of tweets and other implicit characteristics. Subsequently, the outputs of this model and the tweet metadata feed a second LSTM, which captures sharing patterns and semantic similarity. A third LSTM captures the semantics of the account description, and the outputs of the LSTMs, together with account metadata, are used for the final prediction.

Despite this, constant improvements need to be considered regardless of how efficient a model seems currently, because social bots continue evolving at a fast pace. In this sense, investigation of the state of the art in social bot detection reveals significant challenges in terms of scalability and precision. The most common methods, in general, include neural network techniques, supervised and unsupervised machine learning, and community detection. And recent studies such as [Zhang and Wu 2020] highlight the importance of analyzing behavioral characteristics of bots, which increasingly resemble human users, instead of focusing only on improving classifiers. Behavioral indicators such as posting patterns, interactions, and content type are examples of crucial characteristics for such detection.

Research also indicates that sentiment analysis in tweets and temporal posting frequency are effective parameters for this context [Gatkal et al. 2021] [Cai et al. 2017]. Fragmented methods, such as the "Bot-Score" by [Barhate et al. 2020], which evaluates multiple characteristics, have shown promise. Techniques such as LDA, CNN, LSTM and transformer-based language models (such as BERT) are widely mentioned in the found research. In addition, community detection using algorithms such as Louvain and K-means helps reduce the amount of analyzed data.

Another additional relevant approach is the technique based on the concept of users' "digital DNA", proposed by [Cresci et al. 2017]. This method transforms each user's timeline into sequences similar to DNA chains, mapping tweet types to nucleobases (retweets as C, replies as T, and other tweets as A). Similarity between these "DNA chains" is then calculated using the longest common subsequence (LCS) metric, and users are grouped based on their similarity scores. This approach enables the identification of coordinated behaviors among bots, which can be important clues for detecting entire bot families.

Therefore, the SOTA in political polarization bot detection is a combination of efficient and current methods that vary according to the researcher's objective. To identify bots that promote political polarization, deep learning methods to categorize the theme of the data and language models for contextual analysis are ideal. Community detection can optimize the process and help identify botnet components. However, challenges persist in the constant updating of bots' behavioral characteristics and in handling the large volume of data on social networks. Incorporation of techniques such as "digital DNA" offers an innovative perspective for detection, addressing analysis of coordinated behaviors among suspicious accounts, but so far no found technique is free from the possibility of improvement.

4.2. Trends and Signals

Considering the entire context presented in Subtopic 4.1, the main trends associated with this environment were also identified, and how the technologies that will develop will affect this scenario.

The main identified trend, and one of the closest to being realized, is the increase of similarity between bots and human users, to the point that the difference between them is almost unrecognizable. In the context of political polarization, botmasters will be able to further automate bots' social engineering strategies using machine learning or liquid neural networks to create false personas indistinguishable from real ones. To combat this, researchers will have to develop automated security mechanisms based on artificial intelligence that are as efficient as, or more than, malicious chatbots. Automated social honeypots may also be used to interact with automated accounts, resulting in a scenario where two artificial intelligences try to convince each other that they are real people, generating a variation of the Turing Test without human intervention. These technologies will have a significant impact on manipulation and protection of online political discussions.

This trend is supported by several researchers in the area, who indicate the constant need to update the behavioral characteristics analyzed by detection systems, as well as by the Innovation Map in the topics "Honeypot-Based Social Engineering Defense" and "Liquid Machine Learning" (prediction of a deep learning model that learns while executing a task, adapting to new data in real time, providing speed and adaptability for machine learning). Furthermore, continuous technological evolution is making these tools more accessible, with a clear trend that technology becomes increasingly cheaper and efficient. This means that both botmasters and researchers interested in the area will have access to more advanced and low-cost technologies, which can intensify technological competition between both sides.

Another trend pointed out by the Innovation Map is that of artificial superintelligence, a technology that would surpass human intellectual capacity. Although this trend is more distant, if it materializes, intelligent systems that encourage political polarization could make parts of the botnet dispensable. With superintelligence, the need for human intervention to manage and coordinate the botnet could be drastically reduced. Bots controlled by superintelligence would be capable of learning and adapting in real time much more effectively than traditional bots, reducing the need for continuous updates and adjustments. Superintelligence could also develop self-sufficient bots, making detection by traditional botnet tracing techniques more difficult. The Botmaster itself would have a reduced role in this aspect, making attribution of responsibility to entities that use this form of mass manipulation even more complex. On the other hand, this technology would also be extremely useful in containing this problem, leading to the possibility of an internal war between AIs, where the determining factor for the success of one of the sides would be the human agent.

5. Speculation of Possible Futures

5.1. Where Are We Going?

If no intervention is carried out, the trends raised in 4.2 suggest a substantial reconfiguration of the political polarization scenario on social media. The increase in similarity between bots and human users will make it increasingly difficult to distinguish authentic and automated accounts, allowing botmasters to further automate social engineering strategies. With the use of machine learning and liquid neural networks, bots will be able to create false personas that blend perfectly among real users, amplifying manipulation of public opinion and making disinformation even more prevalent.

The absence of significant interventions will allow researchers and bot developers to continue evolving their technologies in a faster and more efficient way. As these tools become cheaper and more accessible, the number of bots used to manipulate online political discussions may increase exponentially (as currently happens with betting bots on Instagram, for example, although these are bots of another type), making social networks a digital battleground where truth and authenticity are constantly challenged.

Furthermore, the arrival of artificial superintelligence could make parts of current botnets dispensable. Bots controlled by superintelligence will be self-sufficient, adaptable in real time and extremely difficult to detect, reducing the need for human intervention to manage and coordinate these networks. This would not only increase efficiency of disinformation campaigns, but also complicate attribution of responsibility, since these operations will be conducted by advanced autonomous entities. Also making the action of legal agents more complex.

With the lack of interventions, online political polarization may intensify, leading to greater fragmentation of society. Political discussions on social networks will be dominated by sophisticated bots that influence and manipulate public opinion without being detected. Trust in social media platforms may decrease significantly, as users lose the ability to discern true information from false, but a large portion of the population will continue to be affected by its influence. With the intensity of political opinions inflamed, the population may reach extreme acts justifying themselves by undesired or controversial electoral results. The situation may worsen from speeches and attitudes of the political personas involved in the clash and eventually generate a civil war.

However, this same superintelligence technology that facilitates manipulation can also be used to combat these threats. If there are no interventions to develop and implement automated security mechanisms based on artificial intelligence, the battle between malicious bots and detection and defense systems will become an unbridled technological race, where the efficiency and adaptability of both sides will be tested until a new manipulation strategy gains more focus, or until governmental systems change to the point where polarization maneuvers no longer have effect on election final results.

To synthesize in a simplified form some presented predictions, we can use the scheme of Figure 3, shown below:

5.2. Cone of the Future

With the purpose of better elucidating how the possible directions of these trends and their implications were suggested, we used the Cone of the Future, presented in Figure 4.

Figure 3. Scheme of Causes and Consequences

6. Projection of the Desirable Future

6.1. IT Solution

The previously speculated future definitely would not be something desirable for most of the population. Thus, a solution to mitigate negative implications of political manipulation on social media could be the development of an Anti-Virus, an advanced artificial intelligence designed to infiltrate botnets through strategies similar to those used by bot binaries, detect their components and sabotage them. This anti-virus would act autonomously, integrating into the bot network and analyzing its communications and behavior patterns as a double agent.

Once infiltrated, the anti-virus would use the strategy of social bots against themselves. From the results obtained by its behavioral analyses, it would mimic the bots, acting as one of them until it can map the entire botnet by techniques derived from Digital DNA, and determine the specific role of each identified bot, such as dissemination of disinformation, amplification of messages, or data collection.

The objective would be to map the entire structure of the botnet, identifying each bot and its connections within the network. With this information, it could deactivate identified bots, interrupting their communication with the botmaster and reducing their operational capacity without alarming the people responsible for the botnet. In addition, the anti-virus would apply digital forensic analysis techniques to identify the botmaster, including IP tracing, analysis of communication patterns and data correlation, providing critical information to authorities and social media platforms. Thus, punishment and exposure of botmasters (and, if possible, their stakeholders) would allow legal actions to be taken against those responsible for manipulating public opinion, discouraging future attempts to use botnets.

Figure 4. Cone of the Future

To ensure long-term effectiveness, the anti-virus could also update itself continuously, adapting to new techniques and strategies of emerging botnets through liquid neural network methods. With this proactive approach, the solution would not only reduce the effectiveness of disinformation and political manipulation campaigns, but would also increase the security of social media platforms. Introducing this anti-virus would change the future scenario, creating a robust defense against automated threats and contributing to restoration of users’ trust in information shared online.

Another positive point would be the possible exposure of entities interested in using this type of tactic, which should affect the public’s perception of these political figures, promoting a more balanced and informed discussion environment.

6.2. Necessary External Actions

But for it to be possible to use such solution to deal with issues of polarizing bots, it would be necessary first to integrate the Anti-Virus into social media APIs, so that besides infiltrating, it can contain the effects of identified bots.

Next, the laws related to online information manipulation, mainly focused on political processes, need to be reviewed and adjusted so that botmasters and their financiers do not go unpunished after being identified. After all, dissemination of false information in this aspect is an anti-democratic act.

Finally, it would also be interesting that educational processes were disseminated to better inform the population how technology can be used negatively. And how information should be analyzed critically from data and facts corroborated by research and serious communication media.

7. Conclusion

7.1. Possible Limitations and Biases

Speculative Design and its various tools are immensely useful in future projections and improvements in product or service proposals. However, this work was carried out individually, which can mean a deficit of other views to complement the made speculations. In this sense, it is possible that some tools would have been better used in group and generated a more comprehensive result.

The main bias that this work may bring is the view that botmasters and researchers are technologically matched. It is possible that due to greater financial support, depending on who finances the respective botnets, botmasters around the world have much more structural conditions to evolve, compared to researchers who, as in **[ANONYMIZED FOR BLIND REVIEW]**, depend on state funds in most cases. And therefore may take longer to have access to some technologies.

Another issue that may have been treated in a not-so-deep way in the present work is the scope of interdisciplinarity that guides the theme. Social bots are human reproductions in the virtual medium, therefore analyses of this type are strongly linked to sociological and cultural issues. Thus, for the development of any accurate solution, consulting specialists in the area focused on online behavior would be interesting. As well as conducting reviews in the used databases, so that in factors such as “sentiment analysis” the use of specific dialects does not cause false positives.

7.2. Contributions of the Work

The main contributions of this work are in the summary of the current condition of the polarizing bot detection scenario; in the suggestion of possible socially worrying futures, if the theme is not addressed; and in the suggestion of a solution considerably different from most proposals up to now.

7.3. Future Work

As future work, it would be interesting to investigate what the situation of research in this environment is in **[ANONYMIZED FOR BLIND REVIEW]** society, that is, whether there are **[ANONYMIZED FOR BLIND REVIEW]** datasets for bot detection, how much national researchers are interested in the area, and to better develop the justification of why it would make sense to apply new technologies to deal with this issue in the national scenario.

Developing an initial prototype of "Anti-Virus" would also be an attractive future approach, however, understanding the ethical issues surrounding the proposal would be priority in this case, to mitigate possible misuse, or even intrusive approaches in environments disconnected from the initial objective.

8. References *(kept verbatim from the original)*

Arin, E. and Kutlu, M. (2023). Deep learning based social bot detection on twitter. IEEE
Transactions on Information Forensics and Security, 18:1763–1772.

Barhate, S., Mangla, R., Panjwani, D., Gatkal, S., and Kazi, F. (2020). Twitter bot detection and their influence in hashtag manipulation. In 2020 IEEE 17th India Council
International Conference (INDICON), pages 1–7.

Belokurov, D. A., Shamakova, E. S., and Kolomoitcev, V. (2021). Using machine learning
techniques to identify bot accounts on a social network. In 2021 Wave Electronics and
its Application in Information and Telecommunication Systems (WECONF), pages 1–
5.

Bortolazzo, S. F. (2020). DAS CONEXÕES ENTRE CULTURA DIGITAL E EDUCAÇÃO: PENSANDO A CONDIÇÃO DIGITAL NA SOCIEDADE CONTEMPORÂNEA. ETD Educação Temática Digital, 22:369 – 388.

Bytheway, A. (2014). Investing in Information: The Information Management Body of Knowledge. Computer science. Springer International Publishing.

Cai, C., Li, L., and Zengi, D. (2017). Behavior enhanced deep bot detection in social media. In 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), pages 128–130.

Ceron-Guzmán, J. A. and León-Guzmán, E. (2016). A sentiment analysis system of spanish tweets and its application in colombia 2014 presidential election. In 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), pages 250–257.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In Proceedings of the 26th International Conference on World Wide Web Companion - WWW ’17 Companion, WWW ’17 Companion. ACM Press.

Delve (2024). Speculative design and a cone of possibilities. Accessed: 2024-07-15.

Gatkal, S., Panjwani, D., Barhate, S., Mangla, R., and Kazi, F. (2021). Community detection and impact of bots on sentiment polarity of twitter networks. In 2021 Asian Conference on Innovation in Technology (ASIANCON), pages 1–6.

Innovation Map (2024). Innovation map. Accessed: 2024-07-18.

Jahan, M. S. and Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546:126232.

Karataş, A. and Şahin, S. (2018). Application areas of community detection: A review. In 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), pages 65–70.

Khattak, S., Ramay, N. R., Khan, K. R., Syed, A. A., and Khayam, S. A. (2014). A taxonomy of botnet behavior, detection, and defense. IEEE Communications Surveys Tutorials, 16(2):898–924.

Zhang, C. and Wu, B. (2020). Social bot detection using ”features fusion”. In 2020 2nd International Conference on Information Technology and Computer Application (ITCA), pages 626–629.