RESEARCH

User's Reaction Patterns in Online Social Network Communities

Azza Bouleimen[1,2*], Nicolò Pagan[2], Stefano Cresci[3], Aleksandra Urman[2], Gianluca Nogara[1] and Silvia Giordano[1]

*Correspondence: azza.bouleimen@supsi.ch
[1]Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland (DTI, SUPSI), Lugano, CH
Full list of author information is available at the end of the article

Introduction

Misinformation, hate speech, toxicity, trolling, and malicious bots are examples of undesired behavior on Online Social Networks (OSNs) with the potential for serious implications in real life for individuals and societies. These harmful impacts range from escalating physical violence in protests [1], threatening democratic elections' integrity [2], to leading to genocides [3]. To address these harms, OSNs platforms introduced some moderation strategies that aim at mitigating these problems. Most of these interventions follow a one-size-fits-all approach, as policies are equally applied to all users [4]. However, these interventions sometimes exacerbate the phenomena instead of limiting them [5, 6]. A possible explanation could consist in the fact that users present diversified reactions to moderation policies [7], in particular because users are typically grouped in communities within OSNs. In this context, some studies highlight the need for personalized moderation interventions [4]. However, to do so, we need to better understand user's susceptibility defined as the factors that drive them toward particular reactions. In other words, we aim to get insights on what makes users more or less likely to engage in undesired behavior. In this work, we aim at studying the reaction of users in network communities as a first step toward understanding user's susceptibilities on the individual level. In turn, this represents a preliminary step towards designing personalized moderation strategies.

Results

Dataset. We base our study on the VaccinItaly dataset [8]. It is a collection of tweets related to the Covid-19 discussion in Italy ranging from Dec, 20th, 2020 to Oct, 22nd 2021. The topic has been very controversial all around the world. Consequently, the choice of this dataset is suitable for studying the susceptibilities of users in contexts that prompt adversarial reactions. The dataset consists of ∼12 million tweets in Italian, half of which are retweets. It involves 551,816 unique users, where 86% of them have less than 20 tweets in the dataset. We selected a subset of users that are involved enough to reflect the core discussion on the Covid-19 vaccines in Italy. To do so, we adapted and applied the definition of a *core user* from [6] to our dataset. This reduces the number of users to 9,278 (1.7% of users) who are responsible for nearly half of the tweets.

The purpose of the study being to observe the behavior of users within the network structure they are in, we built the retweet network of users. It is a directed weighted network where nodes represent the users and edges represent retweets. The weights of the edges represent the number of retweets from one user to another.

Community detection. We applied the Louvain community detection algorithm [9] on the network with a resolution parameter of 0.7. We obtained two main communities that gather 87% of the nodes in the network. We qualitatively analyzed the tweets of the nodes with the high authority scores in these two communities [10]. In one community, the nodes with high authority scores tweet content in favor of the vaccines while, in the other community, the nodes with high authority scores are against the adoption of vaccines and the government's measures to contain the spread of the virus. The same observations are found when analyzing the most retweeted tweets or the tweets of the most central users in these two communities. Hence, we assume that one community is dominated by a **Pro vaccination** discourse (Provax community: 3,980 nodes) and the other is dominated by an **Anti vaccination** discourse (Novax community: 3,831 nodes).

Since our work aims at measuring the differences in the reactions of users belonging to different communities, we first need to understand whether the communities were stable over time, or whether they evolved. In the latter case, differences between the two communities could be simply due to the flow of users between them. To do so, we ran different instances of the community detection algorithm on different sub-periods of time and evaluated the user's flows across communities. Our analysis showed that the composition of the two communities remains overall stable over time, hence we can use the community partitioning based on the whole dataset.

Toxicity in communities. In this abstract, due to space restrictions, we limit the analysis to measuring the user toxicity. Nevertheless, an analysis of negativity was also done and similar observations were obtained. To measure the toxicity of the text of the tweet, we used the Detoxify library [11]. It is a state-of-the-art method for computing toxicity [12, 13]. Detoxify has a multilingual model for non-English texts. For Italian it reaches an AUC of 89.18% [11]. The model returns a score ranging from 0 (low toxicity) to 1 (high toxicity).

We present in Fig. 1 the daily average toxicity of the text written by the users belonging to the Provax and Novax communities. Fig. 1 shows that, from the beginning of the data collection until mid-June, the toxicity level of the Provax community is lower than the one of the Novax community. Interestingly, this trend is inverted starting from mid-June where we notice that Provax users become on average more toxic than Novax users. Overall, the toxicity of both communities increases throughout time as shown by the Mann-Kendall test for trends. However, the Provax toxicity rate increases more than twice as fast as the Novax one.

Toxicity around specific events. To deepen the analysis, we look into the reaction of the Provax and Novax communities around specific events related to the Covid-19 pandemic in Italy. The selected events are presented in Tab. 1. In Fig. 2, we present box plots of the toxicity of the tweets posted by the users of the two communities for the three days following a specific event. Both Provax and Novax have a similar reaction, in terms of toxicity, to the start of the vaccination campaign on Dec 27th. In fact, there is no significant difference between the toxicity of the two communities. For Mar. 15th and Apr. 22nd, the toxicity of the Novax community is significantly higher. This relates to the trends of toxicity observed in Fig. 1. For

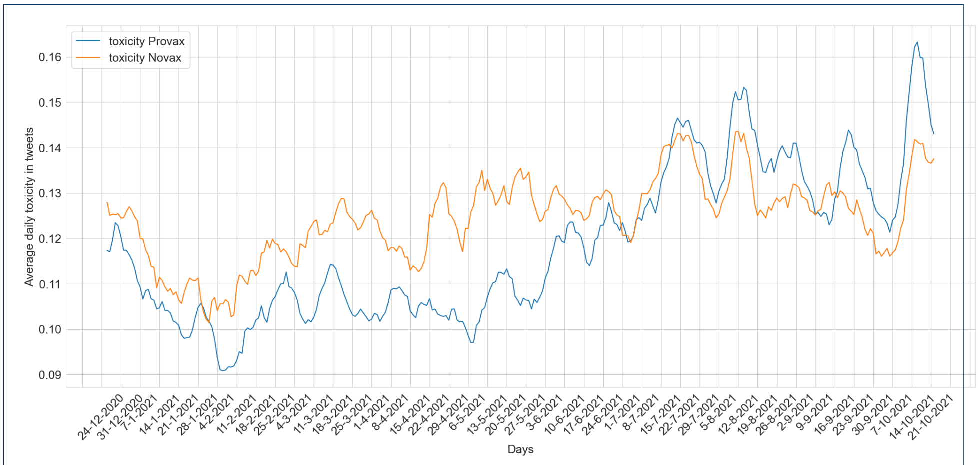

Figure 1 Daily toxicity average in written text for Provax and Novax communities. A moving average of a 7-day window was applied to the plot.

Table 1 List of event related to the Covid-19 pandemic in Italy. Events highlighted in bold correspond to pics in the number of tweets posted by the user on that day.

Date	Event
Dec. 27th, 2020	Start of the vaccination campaign in Italy.
Mar. 15th, 2021	The Italian government announces lockdown during eater vocation.
Apr. 22nd, 2021	Introducing the use of the Green Pass.
Jun. 11th, 2021	Death of a young woman after receiving Astrazeneca shot.
Jul. 23rd, 2021	Announcing mandatory Green Pass starting from the 6th of August 2021.
Aug. 6th, 2021	Mandatory Green Pass required to access several public spaces.

Jun. 11th, the difference between the two communities is not significant anymore. This happens at the time where, in Fig. 1, we recognize an increase in the toxicity level of the Provax community reaching the level of the Novax one. For Jul. 22nd and Aug. 6th, it is the Provax that is this time significantly more toxic than the Novax community. Fig. 2 illustrates that users belonging to different communities develop different reaction patterns depending on which events they are confronted with. It also supports the inversion in the temporal trend of the toxicity within both communities observed in Fig. 1. This result shows that some events might trigger among groups of users a reaction that can shape the behavior of a whole community, while the impact can be non-existent for other groups of users. This supports the need for intervention strategies targeted at the group and at the individual level.

Conclusion

Through the case study of the VaccinItaly dataset, we studied the reaction of users to specific events on OSNs considering their position in the network. We found that these events impact differently the OSN users. They can significantly alter the behavior of a community as a whole and invert the dynamics of behavior within the whole network. Our work highlights the presence of an understudied phenomenon which is the user's susceptibility to undesired behavior. It stresses out the importance of understanding the reasons behind the changes in users' reactions and fine-tuning the research to the individual's level. Possible paths forward include investigating social contagion effects, the interplay between the reactions in the two communities, and the existence of a relation between the structure of the network,

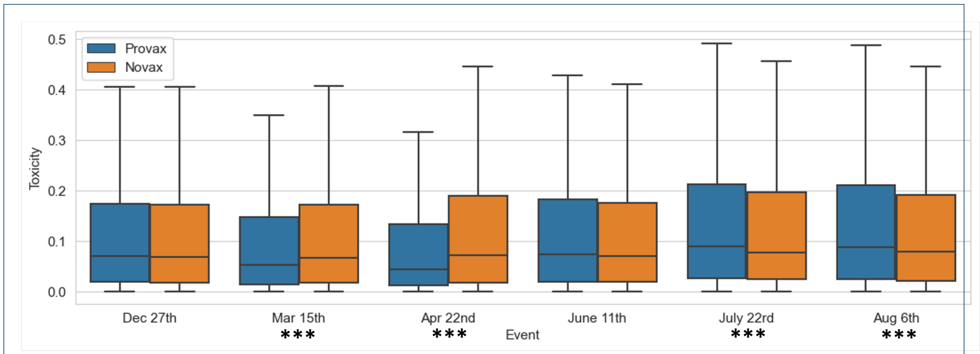

Figure 2 Box plots of the toxicity of tweets published by users of the Provax and Novax communities three days starting from the event date. The stars under the event date on the x-axis represent the significance of the difference between the two communities.

the position of a user within the graph, and their reaction to a particular event. We hope, through our contribution, to pave the way towards building better OSNs' intervention strategies centered on the user.

Author details

[1]Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland (DTI, SUPSI), Lugano, CH. [2]Institute of Informatics, University of Zurich (IfI, UZH), Zürich, CH. [3]Institute for Informatics and Telematics, National Research Council (IIT, CNR), Pisa, Italy.

References

1. Gallacher, J.D., Heerdink, M.W., Hewstone, M.: Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. Social Media+ Society **7**(1), 2056305120984445 (2021)
2. Luceri, L., Deb, A., Giordano, S., Ferrara, E.: Evolution of bot and human behavior during elections. First Monday (2019)
3. Yue, N.: The" weaponization" of facebook in myanmar: A case for corporate criminal liability. Hastings LJ **71**, 813 (2019)
4. Cresci, S., Trujillo, A., Fagni, T.: Personalized interventions for online moderation. In: Proceedings of the 33rd ACM Conference on Hypertext and Social Media, pp. 248–251 (2022)
5. Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., West, R.: Do platform migrations compromise content moderation? Evidence from r/The_Donald and r/Incels. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2), 1–24 (2021)
6. Trujillo, A., Cresci, S.: Make Reddit Great Again: Assessing community effects of moderation interventions on r/The_Donald. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–28 (2022)
7. Trujillo, A., Cresci, S.: One of Many: Assessing User-level Effects of Moderation Interventions on r/The_Donald. In: Proceedings of the 15th ACM Web Science Conference 2023, pp. 55–64 (2023)
8. Pierri, F., Pavanetto, S., Brambilla, M., Ceri, S.: Vaccinitaly: monitoring italian conversations around vaccines on twitter (2021)
9. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (2008). doi:10.1088/1742-5468/2008/10/P10008
10. Kleinberg, J.M., *et al.*: Authoritative sources in a hyperlinked environment. In: SODA, vol. 98, pp. 668–677 (1998)
11. Hanu, L., Unitary team: Detoxify. Github. https://github.com/unitaryai/detoxify (2020)
12. Rossetti, M., Zaman, T.: Bots, disinformation, and the first impeachment of us president donald trump. Plos one **18**(5), 0283971 (2023)
13. Maleki, M., Arani, M., Buchholz, E., Mead, E., Agarwal, N.: Applying an epidemiological model to evaluate the propagation of misinformation and legitimate covid-19-related information on twitter. In: Social, Cultural, and Behavioral Modeling: 14th International Conference, SBP-BRiMS 2021, Virtual Event, July 6–9, 2021, Proceedings, pp. 23–34 (2021). Springer