Toxic comments reduce activity of volunteer editors on Wikipedia

Keywords: Wikipedia, toxicity, harassment, Wikipedia editors, text analysis

Wikipedia is arguably one of the most successful collaborative projects in history. It is the largest encyclopedia ever created and it is used by millions of users worldwide on everyday basis as the first source of information when encountering a new topic, for fact-checking and in-depth research [5]. The content of Wikipedia is created and curated by thousands of volunteer-editors known as Wikipedians. English Wikipedia, the largest language edition, has 130,652 active editors.

As Wikipedia relies solely on volunteer effort, it might be particularly vulnerable to toxic speech. Affected by toxic comments, Wikipedians might reduce their contributions or abandon the project altogether. Despite Wikipedia's "No personal attacks" policy [7], toxic speech and harassment have been frequently observed on the platform [1, 4, 6, 8]. The effect of such behaviors on editors' contributions is, however, not well understood. The largest study to date relies on a voluntary opt-in survey of the 3,845 Wikipedians conducted in 2015 [6]. It reports that 20% of users witnessing harassment have stopped contributing for a while, 17% considered not contributing anymore and 5% stopped contributing at all.

In this paper, we analyzed all 57 million comments made on user talk pages of editors on the six most active language editions of Wikipedia (English, German, French, Spanish, Italian, Russian) to understand the impact of toxicity on editors' contributions. User talk pages are a place for editors to communicate with each other either on more personal topics or to extend their discussion from an article's talk page. The majority of toxic comments are left on user talk pages [3]. The comments we study were extracted from revision histories of talk pages and, thus, include even those toxic comments that were later archived or deleted by the page owner.

To identify toxic comments we used a model from the Perspective API [2] – the state-of-the-art toxicity detection algorithm. It is a BERT-based model trained on comments from a variety of online sources, including Wikipedia. When presenting results, we define a comment as toxic if the model score is larger than 0.8 and as non-toxic if the model score is lower than 0.2. However, we show that our findings are robust with respect to selecting different toxicity thresholds.

To estimate the effect of a toxic comment, we compute the proportion of users who were active on day X before or after receiving a toxic comment (Figure 1). We find that, on average, editors are more active near the time when they receive a toxic comment. This is a rather unsurprising observation since toxic comments are often made as a reaction to an edit made by a user and, thus, users are expected to be active around the time of a toxic comment. We also find that the average activity across all users who have received a toxic comment is lower during all 100 days after the event compared to the corresponding days before (dashed and solid red lines in Figure 1b). To rule out the possibility that this is due to a general drop in activity over time or a drop in activity after any comment, we select a control group of users who have received a non-toxic comment, and whose average activity in the 100 days before the comment is the same as the average activity of users who received a toxic comment. In contrast to a toxic comment, a non-toxic comment does not lead to a significant decrease in activity (dashed and solid blue lines in Figure 1b). Similar results hold for all six language editions that we have examined (Figure 1c-g).

Note that given that thousands of users have received at least one toxic comment, even a moderate loss per user could result in many human-years of lost productivity for Wikipedia. Our estimates range from 5 human-years for Russian Wikipedia to 265 human-years for the English edition. The reason for the lasting effect of toxicity is that some users are discouraged by a toxic comment and choose to leave the project altogether. To further investigate this effect we compare the probability of leaving Wikipedia after receiving a toxic comment with the probability of leaving Wikipedia after receiving a non-toxic comment.

We observe that the probability of leaving Wikipedia after N contributions declines with N following a power law. While the probability of leaving the project after the first contribution is high ($P_1 = 47\%$ for English Wikipedia), the risk of leaving Wikipedia drops to 0.7% for users who have made 100 contributions (Figure 2). At the same time, the risk of an editor leaving after a toxic comment is consistently higher for all editions and regardless of the contribution number.

Toxicity is especially dangerous to Wikipedia which relies on the contributions of volunteer editors to create articles with broad, global exposure. To ensure the objectivity and comprehensiveness of articles, it is critical to establish and maintain an open and inclusive collaborative environment. However, this is not always the case on talk pages, where clashing editor opinions can lead to toxic and harmful comments. As we have shown in this paper, this comes at a cost and has negative effects on editors. It could also lead to the idea that harassment is a possible tool for silencing users.

Wikipedia plays a crucial role in the global information infrastructure. It is often considered as a neutral and comprehensive source of knowledge and for that reason its articles are used by many services as a first choice information source, e.g. Wikipedia articles are featured in the results of major search engines. While many are aware of the darker corners of the Internet, it is often assumed that Wikipedia is free from toxicity. Our findings demonstrate that the toxicity is not only present but its impact is substantial.

References

[1] Danielle J Corple. *Beyond the Gender Gap: Understanding Women's Participation in Wikipedia*. PhD thesis, Purdue University, 2016.

[2] Perspective API. Technical documentation, 2022.

[3] Iris Qu, Nithum Thain, and Yiqing Hua. Wikidetox visualization. In *Wiki Workshop*, 2019.

[4] Michael Raish. Identifying and classifying harassment in arabic wikipedia: A "netnography", 2019.

[5] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why we read wikipedia. In *Proceedings of the 26th international conference on world wide web*, pages 1591–1600, 2017.

[6] Support and Safety Team. Harassment survey 2015, 2015.

[7] Wikipedia. Personal attacks, 2022.

[8] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

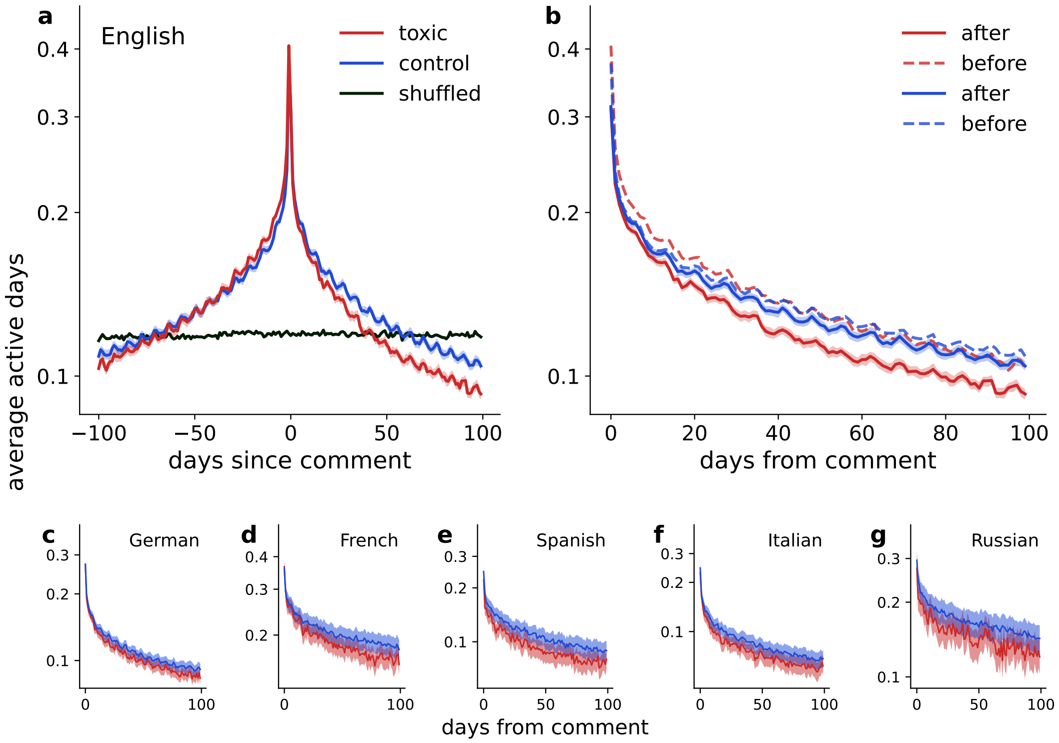

Figure 1: **After receiving a toxic comment, editors become less active.** On average, users are more active near the time when they receive a toxic comment (peak at zero for the red line in **a**). Average activity across all users who have received a toxic comment is lower in all 100 days after the event compared to the corresponding days before (dashed and solid red lines in **b**). This cannot be explained by a baseline drop in activity after non-toxic comment (dashed and solid blue lines in **b**). Similar results hold not only for the English edition but also for other five editions (**c-g**).

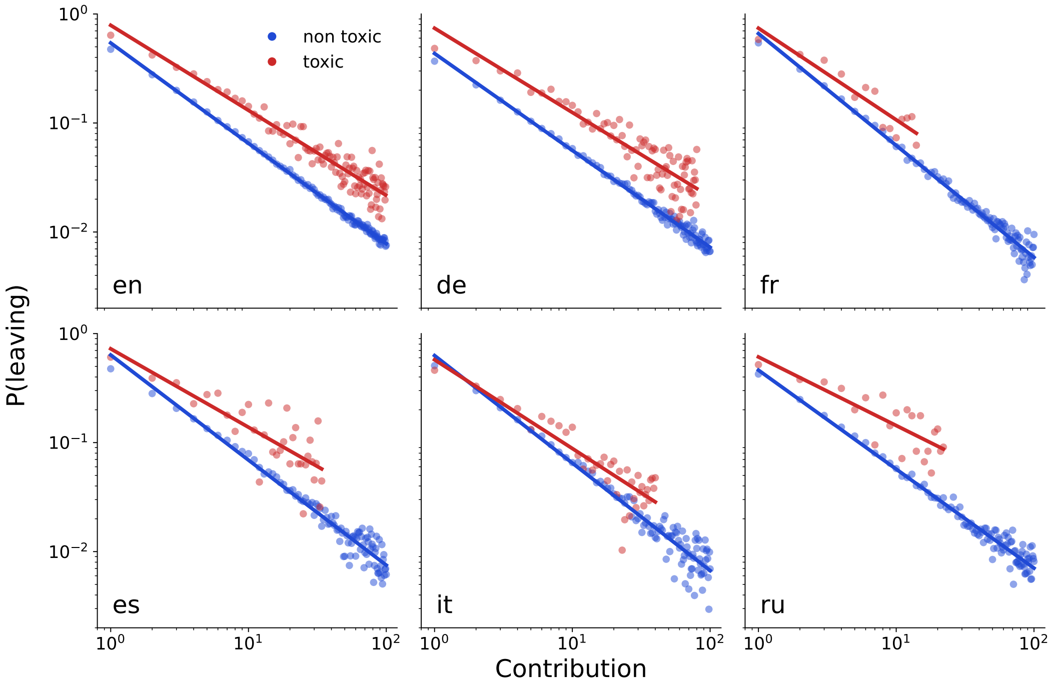

Figure 2: **The probability of leaving Wikipedia after receiving a toxic comment is substantially higher than after a non-toxic comment.** For all six editions the probability of leaving declines with the number of contribution approximately following the power law. At the same time, this probability is substantially higher after receiving a toxic comment that might be expected otherwise. Dots are probability estimates and solid lines are best linear fit on a log-log scale.