9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

How do Toxic Comments Induce Fearful Responses in Anti-vaccine Videos on YouTube?

Keywords: Emotional contagion, Fear, Toxicity, Vaccine hesitancy, YouTube

Extended Abstract

In the face of a pandemic, it is crucial to mitigate vaccine hesitancy to prevent the spread of contagious diseases. Although the efficacy of vaccines has been repeatedly confirmed, vaccine hesitancy continues to persist. Anti-vaccine groups often employ toxic messages as part of their communication strategy on social media [8], which may exacerbate anti-vaccine movements. It is also known that messages that are emotionally and negatively charged have higher inclinations toward vaccine aversion [1]. However, empirical evidence linking toxic messages with vaccine hesitancy remains largely absent.

This study aims to examine the influence of toxic comments about anti-vaccine videos on YouTube—one of the most popular information sources regarding a pandemic and vaccines. The comment section on YouTube videos is a venue not only for feedback for video creators but also for communication and information sharing among viewers, providing new experiences for video audiences. Despite its importance, the comment section has been suffering from uncivil messages, particularly on anti-vaccine videos [7]. Also, in the context of anti-vaccine, fear is one of the most prominent emotions because many narratives behind vaccine hesitancy and anti-vaccination rely on fear of vaccines. Quantifying the influence of toxic comments on other users would provide valuable insights for platformers to effectively moderate anti-vaccine video comments and maintain a safe and informative environment.

To this aim, we quantitatively evaluate the relationship between toxicity and fear in comments to anti-vaccine videos on YouTube. We use a dataset of 484 anti-vaccine videos and corresponding 414,436 comments collected in published studies [9]. To quantify the level of fear and toxicity in every single comment, we use state-of-the-art machine learning models: Perspective API [5] for toxicity and a RoBERTa-based model [4] for fear, respectively.

First, we focused on comment-by-comment dynamics, finding that highly-fearful and highly-toxic comments are only part of the comments. Fig. 1a shows that the fear and toxicity scores of the top 20th percentile comments are 0.03 for fear and 0.29 for toxicity on a 0-1 scale, respectively, indicating the skewness of the distribution. An example of temporal dynamics of fear and toxicity in a comment section is shown in Fig. 1b. We found no discernible temporal trend in both fear and toxicity, with intermittent spikes of highly-toxic and highly-fearful comments. Therefore, instead of focusing on individual spikes, we measure the average fear and toxicity while varying the number of comments for subsequent observations.

Second, we analyzed the relationship between fear and toxicity of aggregated comments per video. A linear regression analysis was conducted with the mean toxicity of the comment section as the independent variable and the mean fear as the dependent variable (Fig. 2a). The analysis revealed that the variable of toxicity (comment), as indicated at the bottom of the figure, showed a significant coefficient, even when controlling for other variables such as topics, engagement metrics, published date, and emotion/tone of title, description, and transcripts. This suggests that toxicity and fear of comments aggregated at the video level are strongly associated. We also found the topics of the virus itself and children's disease were significantly

associated with fear of comments, which aligns with existing research that suggests that these topics are prevalent causes of fear for anti-vaccine groups [6]. With regard to the video emotion/tone, we found that fear of title/description/transcript is positively associated with fear of comments, which is consistent with previous research that suggests that the emotional content of videos is contagious to viewers' emotions [3].

Lastly, we examined contagion dynamics between toxicity and fear in aggregated comments. For this, we divided the comments into the early part and the later part, and then conducted a linear regression analysis that estimates the fear in later comments using the toxicity in early comments (Fig. 3). The result shows that the toxicity of early comments is slightly associated with the fear of subsequent comments. Most importantly, the toxicity of early comments that received a high number of likes is disproportionately linked to the fear of later comments on videos, with an average increase of 1.3 times. Note that there is no simple positive correlation between toxicity and fear at the individual comment level (Fig. 2b). Therefore, the contagion from toxicity to fear in the aggregated comments is not trivial. Furthermore, we tested their relationship in the inverse setting that estimates the later toxicity as dependent using early fear as the independent variable (Fig. 4). The coefficients for fear of early comments are largely positive and significant, meaning that the influence of the toxicity and fear of comments is *bidirectional*. We observed the contagion between the same features, i.e., from toxicity to toxicity and from fear to fear, in both model experiments.

The contribution of this study is twofold. Our findings have important implications for comment moderation on online platforms. Specifically, we show evidence that toxic comments on anti-vaccine movies can have a fear-inducing effect, suggesting that such comments should be given higher priority for content moderation. Additionally, the fact that the toxicity of *highly-liked* comments is more significantly associated with fear of subsequent comments highlights the need for platforms to re-evaluate their comment sorting algorithms. It is also worth noting that we found heterogeneous contagion between toxicity and fear in online communications. In *emotional contagion*, it is often assumed that exposure to a particular emotion is considered to lead to a cascade of the same emotion (i.e., from anger to anger) [3, 2]. The heterogeneous contagion found here, however, may be beyond the assumption and we may need a new explanation of contagion.

[1] C. Betsch, C. Ulshöfer, F. Renkewitz, and T. Betsch. The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making*, 31(5):742–753, 2011.

[2] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions, 2017.

[3] E. Ferrara and Z. Yang. Measuring emotional contagion in social media. *PloS one*, 10(11):e0142390, 2015.

[4] J. Hartmann. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`, 2022.

[5] Jigsaw. Perspective api. `https://perspectiveapi.com/`. (Accessed on 01/29/2023).

[6] A. Kata. Anti-vaccine activists, web 2.0, and the postmodern paradigm–an overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*, 30(25):3778–3789, 2012.

[7] M. S. Locatelli, J. Caetano, W. Meira Jr, and V. Almeida. Characterizing vaccination movements on youtube in the united states and brazil, 2022.

[8] K. Miyazaki, T. Uchiba, K. Tanaka, and K. Sasahara. Aggressive behaviour of anti-vaxxers and their toxic replies in english and japanese. *Humanities and social sciences communications*, 9(1):1–8, 2022.

[9] K. Papadamou, S. Zannettou, J. Blackburn, E. De Cristofaro, G. Stringhini, and M. Sirivianos. "it is just a flu": Assessing the effect of watch history on youtube's pseudoscientific video recommendations, 2022.

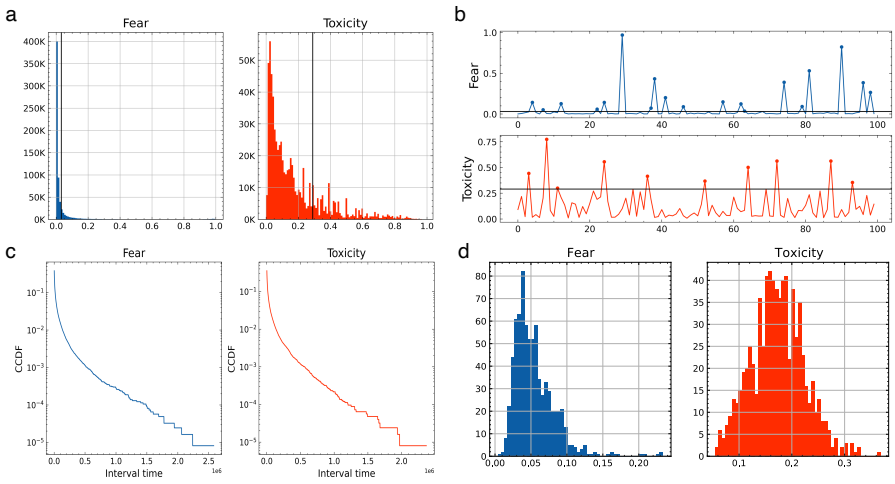

Figure 1: **Prevalence and dynamics of fear and toxicity of comments.** **a**: Histograms of the scores of fear and toxicity for all comments, and the top 20th percentile thresholds are shown as vertical lines. **b**: Time series of fear and toxicity of 100 comments for a certain video. The horizontal lines are the 20 percentile thresholds shown in a. **c**: Log-log plot of CCDFs indicating the probability of the interval time of all highly-fearful and highly-toxic comments. **d**: Histogram of the mean of fear and the mean toxicity of comments for videos.

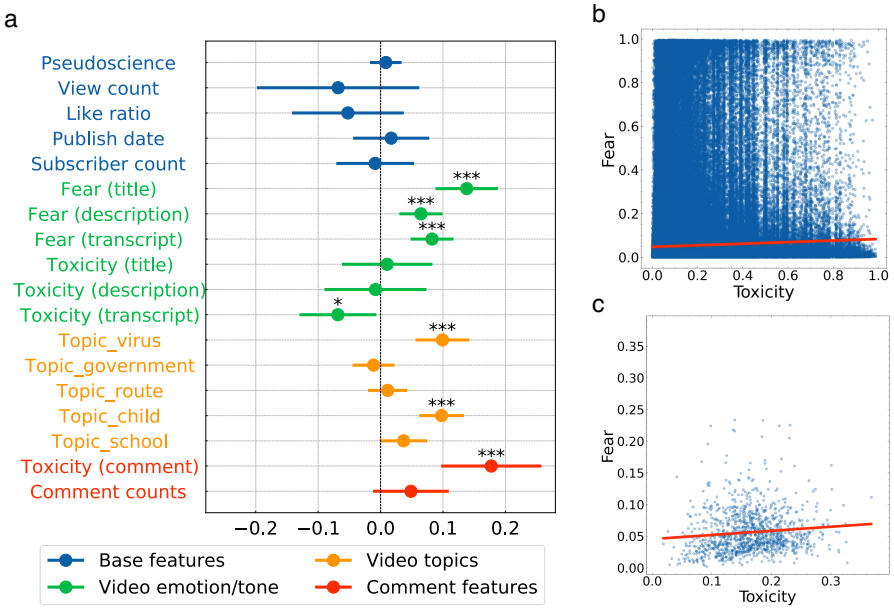

Figure 2: **Results for video-level regression.** **a**: The coefficient of variables with 95% CIs. The stars indicate the p values of the t-test: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$. **b**: Scatterplot of comment levels between toxicity and fear. Data points indicate comments; red lines indicate regression lines. The Pearson correlation coefficient was 0.04 ($p = 0.00$), and the coefficient of the single regression analysis was 0.03. **c**: Scatter plots at the video level between mean toxicity and mean fear. Data points indicate videos, and the red line shows the regression line. The Pearson correlation coefficient was 0.10 ($p = 0.00$), and the coefficient of the single regression analysis was 0.06.

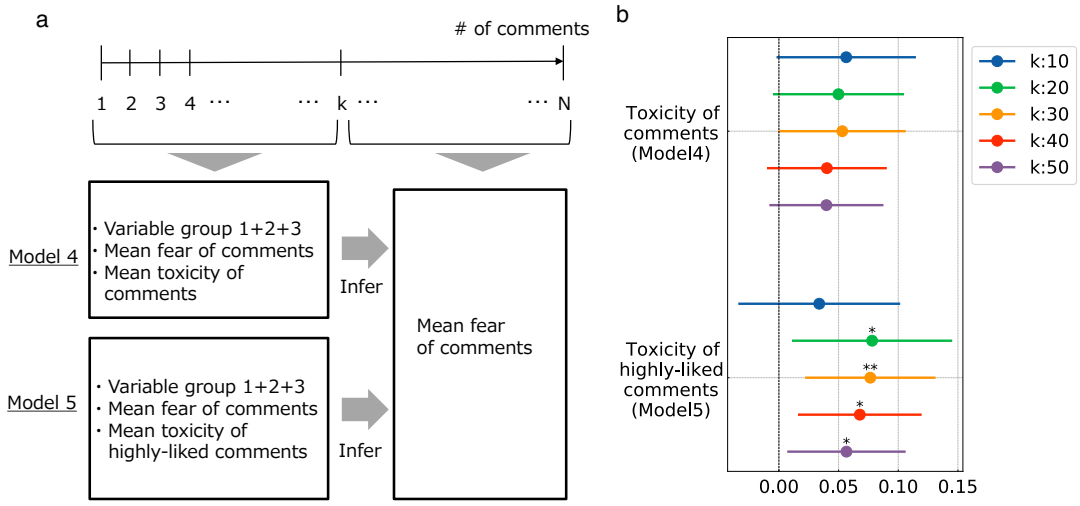

Figure 3: **Measuring the influence of toxicity of early comments on the fear of later comments. a**: Illustration of the problem set. N comments in chronological order for a given video are divided into early and later halves, separated by k. Then, the mean fear of comments in the comment range is predicted by the variables noted in Model 4 and Model 5, respectively, and the coefficients are obtained. **b**: Forest plots showing the coefficients of mean toxicity of comments and highly-liked comments, for $k = \{10, 20, 30, 40, 50\}$. Both are positive regardless of k, but the only mean toxicity of highly-liked comments is largely significant. The mean toxicity of highly-liked comments has a high coefficient compared to the mean toxicity of normal comments (1.3 times higher in the average of the value in the five windows).

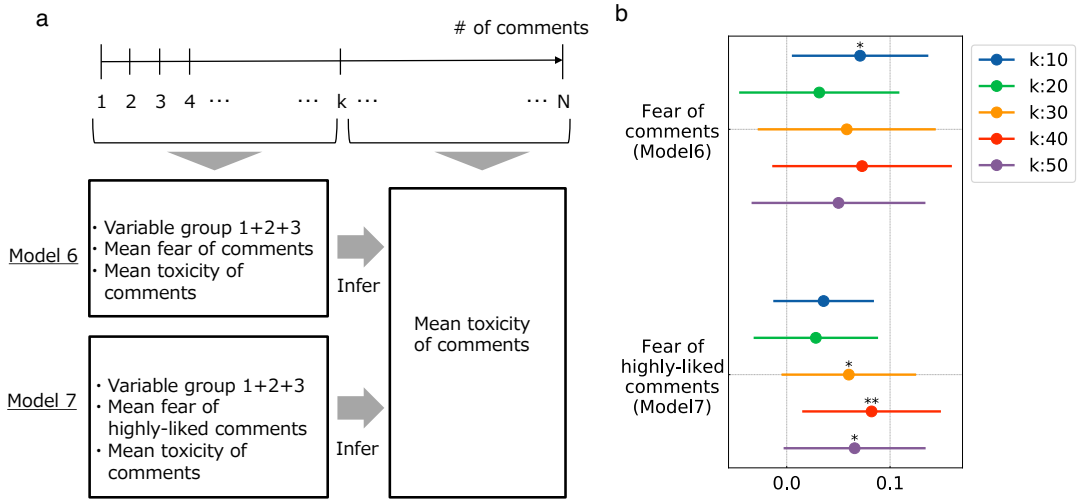

Figure 4: **Measuring the influence of the fear of early comments on the toxicity of later comments. a**: Illustration of the problem set. N comments in chronological order for a given video are divided into early and later halves, separated by k. Then, the mean fear of comments in the comment range is inferred by the variables noted in Model 6 and Model 7, respectively, and the coefficients are obtained. **b**: Forest plots showing the coefficients of the fear of comments and the fear of highly-liked comments, for $k = \{10, 20, 30, 40, 50\}$. Only the coefficients for fear of highly-liked comments are largely significant (3 in 5 cases).