Understanding the Bystander Effect on Toxic Twitter Conversations

[Hansoul Jung, Ana Aleksandric, Mohit Singhal, Anne Groggel, Shirin Nilizadeh]

Department of Computer Science and Engineering. University of Texas at Arlington, Arlington, Texas, 76019

Abstract

Did you know that every single minute, an average of 350,000 new tweets are posted on Twitter making up more than 500 million Tweets a day? As users embrace social media, hate speeches and cyberbullying become a growing concern. By analyzing more than randomly selected 187k tweets, we outline bystanders and the tone of initial responses to a toxic reply as explanatory factors which affect whether others feel uninhibited to post their own abusive replies.

Introduction

In this paper, we assess a random sample of Twitter conversations to examine how group dynamics can be a social determinant of online behavior. Our finding suggest that the bystander effect is present in the social media environments of Twitter. The current study proposes and tests a series of hypotheses about the effects of group dynamics on the toxicity of Twitter conversations.

Table 1: Formulated Hypotheses

Hypothesis type	Hypothesis description
Hypothesis 1	# of users participating in a conversation before observing the first toxic reply is negatively associated with the number of users who post non-toxic replies after the first toxic reply.
Hypothesis 2	If a user posts a non-toxic reply immediately after the toxic reply, then more users post non-toxic replies
Hypothesis 3	If a user posts a non-toxic reply immediately after the toxic reply, then the toxicity of the conversation after this reply is more likely to be non-toxic.

Study Framework

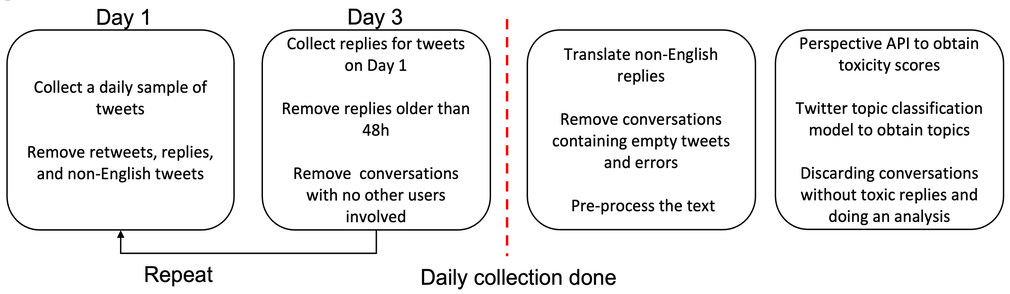

Figure 1: Study Framework

Methods

Table 2: Regression models used to test each hypothesis

Hypothesis type	Model type
Hypothesis 1	Poisson regression model
Hypothesis 2	Poisson regression model
Hypothesis 3	Linear regression model

Table 3: Variables assigned

Type	Variables
Independent Variables	**users_before_toxic**: the number of unique users engaged in the conversation before the first toxic comment occurred **reply_toxicity**: toxicity score of the 1st comment posted that is a direct reply to the 1st toxic comment in conversation
Dependent Variables	**unique_users_non_toxic**: total # of unique users who posted non-toxic comments after the 1st toxic reply in the conversation thread **remaining_toxicity**: ratio of all toxic replies that occurred in the conversation thread after the first toxic reply and the total number of replies in the tread
Control Variables	Account and conversation characteristics

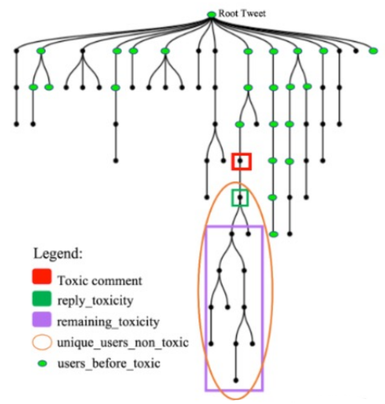

Figure 2: Example of a conversation tree

Results

Final data set:
- 9,107 conversations
- 187,658 tweets from 118,609 unique users

Dataset 1 (2021): 5,115 conversations
Dataset 2 (2023): 3,992 conversations

Datasets are significantly different allowing us to examine our hypotheses applying consistently

H1: With the Poisson regression model, we found that **more user participation before the first toxic comment leads to a lower number of users replying in a non-toxic way.**

H2: With the Poisson regression model, we found that **toxic reply begets more toxicity**.

H3: With the linear regression model, we found that **toxic reply after first toxic reply leads to uncivil behavior.**

Table 4: Results of the regression analysis

	Dependent variable:		
	unique_users_non_toxic		remaining_toxicity
	H1: Poisson	*H2: Poisson*	*H3 :OLS*
users_before_toxic	−0.020*** (0.004)	NA	NA
reply_toxicity	NA	−1.615*** (0.103)	0.086*** (0.015)
num_followers	−0.0 (0.0)	−0.0** (0.0)	0.0 (0.0)
num_friends	0.0 (0.0)	0.0 (0.0)	−0.0 (0.0)
num_tweets	0.0* (0.0)	0.0 (0.0)	−0.0 (0.0)
listed_counts	0.00001 (0.00002)	0.0001** (0.00003)	−0.0 (0.00001)
verifiedTrue	−0.471*** (0.09)	0.363*** (0.09)	−0.01 (0.02)
account_age	−0.010* (0.004)	−0.001 (0.004)	0.001 (0.001)
has_URLTrue	−0.045 (0.03)	−0.007 (0.03)	0.018** (0.007)
description_length	−0.001* (0.0003)	0.0004 (0.0003)	−0.0001 (0.0001)
has_locationTrue	0.107** (0.04)	−0.031 (0.04)	0.008 (0.008)
width	0.001** (0.0004)	0.003*** (0.0005)	−0.0004* (0.0002)
depth	0.021*** (0.001)	0.003 (0.002)	−0.004*** (0.0004)
Observations	9,171	3,770	1,959
R^2	NA	NA	0.064
Note:			*p<0.05; **p<0.01; ***p<0.001

Conclusion

This study examines the impact of group dynamics and the bystander effect on Twitter conversations, focusing on the disinhibition of toxic replies.

In this research we…
- Find evidence of a bystander effect with increased conversational participants being associated with fewer Twitter users standing up to a toxic reply.
- Highlight the importance of initial responses to a toxic tweet within a conversation.
- Show that posting a toxic reply immediately after a toxic comment predicts that the Twitter conversation will become increasingly toxic

Process and Insights

With guidance from my research advisors, I was able to take insight into data science and machine learning, applying these technical skills to analyze social and psychological behaviors within Twitter conversations. I acquired proficiency in utilizing pandas in Python to extract data from CSV files. Additionally, I looked at R codes for the first time, where I gained the ability to interpret and execute code lines aimed at testing various regression models to validate our hypotheses. This experience not only broadened my understanding of collaboration and research methodologies but also significantly enhanced my expertise in my major field of study: Computer Science

References

[1] Google Perspective API. 2021. https://www.perspectiveapi.com/
[2] Performance Overview. 2023. https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US&tabset-20254=3
[3] Twitter. 2022. Twitter API. https://developer.twitter.com/en/docs/twitter-api
[4] Kimberley R Allison and Kay Bussey. 2016. Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. Children and Youth Services Review 65 (2016), 183–194.
[5] Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Daehun Nyang, and David Mohaisen. 2021. Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in youtube. In Companion Proceedings of the Web Conference 2021. 508–515.
[6] Ashley A Anderson, Sara K Yeo, Dominique Brossard, Dietram A Scheufele, and Michael A Xenos. 2016. Toxic talk: How online incivility can undermine perceptions of media. International Journal of Public Opinion Research 30, 1 (2016),156–168
[7] Peter Fischer, Joachim I Krueger, Tobias Greitemeyer, Claudia Vogrincic, Andreas Kastenmüller, Dieter Frey, Moritz Heene, Magdalena Wicher, and Martina Kainbacher. 2011. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. Psychological bulletin137, 4 (2011), 517.
[8] Stephanie D Freis and Regan AR Gurung. 2013. A Facebook analysis of helping behavior in online bullying. Psychology of popular media culture 2, 1 (2013), 11.

UNIVERSITY OF TEXAS ARLINGTON

Contact: Hansoul Jung, hxj2886@mavs.uta.edu
UTA Department of Undergraduate Research
www.uta.edu/research/opportunities/undergraduate-research