Conversation-Level Resilience to Bad Actors in Reddit Communities

Charlotte Lambert and Eshwar Chandrasekharan
Department of Computer Science, University of Illinois at Urbana-Champaign

ILLINOIS

Introduction

Online conversations, just like offline ones, are susceptible to influence by bad actors.

We examine examine the notion of ***conversational resilience*** :

- What makes a conversation more or less capable of withstanding an adverse event?

Motivation

Questions:

1. Do adverse events always derail online conversations?
2. Where should moderators focus their limited attention?

Impact of this work:

1. Understand contributing factors to *resilience* in online conversations
2. Determine conversations that warrant more moderation resources
3. *(bonus)* Identify behaviors to encourage to promote resilience

Data + Methods

We focus on 10.5M comments from five mainstream subreddits (over 2.9M subscribers each).

Comments are organized into more than 5.8M comment threads (i.e., *conversations*).

Using 30K relevant conversations, we investigate the effects of different norm violations on the outcome of a conversation.

Motivating Examples

P N P ... N N **A** N A A

vs.

P N P ... N N **A** N P P

A : Adverse event (e.g., norm violation)

N : Neutral event

P : Prosocial event

Questions:

1. Why does one conversation not get derailed by a norm violation while another does?
2. How do we determine the differences and which should receive moderator focus?

Results

1. We find that norm violations that occur at the start of a conversation do not necessarily stop the flow of subsequent comments.

Independent Variable(s)	Dependent Variable: Number of Comments After *First Adverse Event*				
	r/books	r/games	r/nba	r/relationships	r/science
Depth of *First Adverse Event*	−	−	−	−	−

2. Additionally, highly-toxic violations are more likely to occur in comment threads containing other toxic comments, both before and after the violation itself. Possibly focus moderator attention on high toxicity violations.

Independent Variable(s)	Dependent Variable: Average Toxicity of Comments After *First Adverse Event*				
	r/books	r/games	r/nba	r/relationships	r/science
Toxicity of *First Adverse Event*	+	+	+	+	+
Avg. Toxicity Before *First Adverse Event*	+	+	+	+	+

3. Finally, more severe violations may not necessarily lead to higher rates of subsequent violations, while the comment author's history of toxic behavior might..

Future Work

1. How do community-level differences affect conversational resilience?
2. How can we capture the outcome of a conversation and classify it as prosocial or antisocial?
3. Does the presence of prosocial behavior before the adverse event lead to more resilience?

Questions

1. What aspects of online conversations do *you* think of when I say "resilience"?
2. What linguistic characteristics may be worth exploring to recognize prosocial behavior?