Conversational Resilience: How Online Discussions Recover from Destructive Interruptions

Faculty supervisor: Libby Hemphill

Anti-social, destructive behaviors online (e.g., online harassment, cyberbullying, hostility, trolling) are increasingly problematic. Attempts to derail productive conversations online are frequent and pervasive, but they are not always effective. This project aims to understand what makes some online conversations resilient to anti-social behaviors.

This project uses natural language processing and discourse analysis approaches to study conversation trends in online social networks. Our goal is to identify the linguistic, social, and technical features of conversations that are able to recover from attempts like trolling or harassment to continue productively. We can use this knowledge to inform design of interventions and interfaces for successfully responding to destructive behaviors.

Using existing models and measures of hostility, insults, and conversation stoppers, work in the summer of 2019 will include one or more of the following:

- Discover cultural norms around hostility in a particular space
- Measure when labels (e.g., hate speech) apply and/or drift among contexts
- Compare model’s predictions about whether particular posts are problematic
- Automatically learn phrases that de-escalate hostilities
- Uncover the social structure of resilience

Interested students should have strong Python skills and at least some experience working with social media data and/or natural language processing. Experience using the Reddit and Perspective APIs is especially helpful.

I will meet weekly with the student. I expect students to work well independently but will happily help with all aspects of the project.