Let's Gang Up on Cyberbullying

Henry Lieberman, Karthik Dinakar, and Birago Jones, *MIT Media Lab*

The novel design of social network software can help prevent and manage the growing problem of cyberbullying.

Cyberbullying has emerged as a major problem in recent years, afflicting both children and young adults. Tragic stories in the news about suicides of bullied teens have drawn public attention to the issue, and statistics indicate that its prevalence is growing. A 2006 survey commissioned by the National Crime Prevention Council showed that more than 43 percent of US teens were subjected to cyberbullying at some point in the previous year (www.ncpc.org/cyberbullying), while a 2008 survey by UCLA researchers reported that nearly three-quarters of teens had been bullied online at least once in the past 12 months (www.safeinyourspace.org/2008juvonengross.pdf).

Social networks provide many benefits for youth, like helping to start and maintain friendships and providing a personally meaningful context for practicing reading and writing (D. Boyd, *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life,* MIT Press, 2007). But if too many kids are bullied too often on social networks, kids will feel scared to join them and parents won't permit their children to participate.

In the Internet's early days, many people were surprised and discouraged by receiving spam in their e-mail. If no measures had been taken to combat the rapidly growing volume of spam, the Internet as we know it today wouldn't exist. Fortunately, a technical solution—spam filters—managed to get the problem under control, even if it hasn't totally eliminated spam. Are technical solutions available to likewise manage cyberbullying?

The MIT Media Lab's Time Out project is investigating a range of innovations in social network software to help prevent cyberbullying and mitigate the consequences when it does occur. Our efforts fall into two broad categories: detecting possible cases of cyberbullying by using machine learning to better understand cyberbullying language; and intervention technologies for participants as well as network providers and moderators. *Reflective interfaces* encourage participants to carefully consider their behavior and choices, while visualization tools can help providers and moderators control the escalation of cyberbullying.

DETECTION

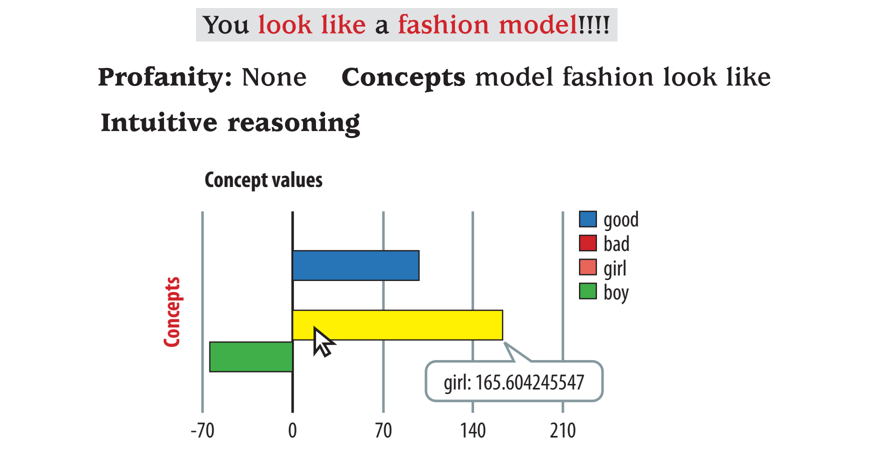
Detecting cyberbullying, which is personalized and contextual, is much more difficult than detecting spam, which is sent identically to large numbers of people. However, our analysis indicates that most cyberbullying occurs around a small number of topics: race and ethnicity, sexuality and sexual identity, physical appearance, intelligence, and social acceptance and rejection. If we can understand whether a message is about those topics, and whether its tone is positive or negative, we can identify many possible cyberbullying messages.

One class of bullying messages we have studied involves accusations of being gay or lesbian, with a negative intent. Often this takes the form of ascribing stereotypically female characteristics to a male or stereotypically male characteristics to a female. For example, a comment addressed to a male might be "You'd look great in lipstick and a dress."

This doesn't suggest people ought to feel bad if such things are said about them—we certainly don't endorse any stereotypes. But in practice, such statements are often used in cyberbullying and so are among several clues as to whether it might be occurring.

Computers still can't fully understand English, but progress in natural-language processing means that sometimes we can partially understand some aspects of a text. Active areas of research include *topic detection*, a mainstay of search and database engines; and *affect analysis*,

0018-9162/11/$26.00 © 2011 IEEE



You look like a fashion model!!!!

Profanity: None **Concepts** model fashion look like

Intuitive reasoning

Figure 1. The term "fashion model" is commonly associated with females; thus, when directed at a straight male, the sentence "You look like a fashion model!" might be an example of cyberbullying.

determining whether a message conveys a positive or negative emotion.

Among the statistical tools we're using are machine learning classifiers, which are effective for many topic detection problems. We train these classifiers on a set of cyberbullying messages identified by humans and analyze the messages for statistical regularities (K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," *Proc. 2011 Social Mobile Web Workshop, Assoc. for the Advancement of Artificial Intelligence*, 2011, pp. 11-17).

Our "secret ingredient" in this process is a commonsense knowledge base with associated reasoning techniques. We've collected about one million sentences describing everyday life that provide the kind of background knowledge AI programs need to go beyond simple word matching and word counting. We use these to simulate the kind of vague, informal reasoning that people do, rather than reasoning with mathematical precision.

The knowledge base contains the kind of statements that help a detection system decide whether a sentence might be referring to stereotypical male or female concepts—for example, "lipstick is a kind of makeup" or "women wear dresses." As Figure 1 shows, the term "fashion model" is commonly associated with females; thus, when directed at a straight male, the sentence "You look like a fashion model!" might be an example of cyberbullying.

Similarly, commonsense knowledge about what people typically eat might indicate that the comment "You must have eaten six hamburgers for dinner tonight" was intended to insult someone for being overweight.

We would like to avoid directly accusing an individual of being a bully in any situation. Our goal isn't to achieve 100 percent certainty in detecting cyberbullying but to call out the possibility of its occurrence. If a pattern is repeated over time, seems to be escalating, or has a consistently negative tone, our confidence in estimation might increase.

INTERVENTION

There are many participants in the cyberbullying process: the perpetrator, the victim, friends, family, teachers, and so on. We can design interventions specifically for each role. As Figure 2 shows, when a possible cyberbulling message is detected, we could unobtrusively provide a link to educational material appropriate to the user's situation.

For potential cyberbullies, the material could encourage empathy for the victim and warn of possible damage to the bully's social reputation. The intervention could exhort victims to seek emotional support, learn how others have dealt with similar situations, give suggestions for appropriate responses (such as humor), and discourage them from retaliating. The material could induce friends to defend the victim rather than join in with the bully.

The key is to offer advice that is personalized, specific, and actionable. The material can take many forms including written stories, video, or interactive narratives.

Other measures could subtly change the social network interface to encourage reflection, or to slow the spread of a potentially insulting message that has been sent. Instead of just a simple "Send" message, for example, the button could be changed to remind the user of the consequences: "Send to 350 people in your network." Likewise, an "Are you sure?" confirmation could be added to potentially problematic messages. Or delivery could be delayed overnight to give the sender a chance to rescind the message in the morning before it's actually delivered.

Social network providers and moderators also have a role to play: they're obligated to provide a safe and welcoming environment for their participants, especially newcomers. To that end, we propose a kind of "air traffic control" interface called SpeedBump, shown in Figure 3, that helps a moderator visualize the community's connections, history, and topics.

SpeedBump's goal isn't to catch every instance of cyberbullying but to prevent incidents from escalating into major outbreaks. Social network providers inform us that such incidents tend to occur in clusters: typically they spread rapidly throughout a particular group, like students at a school, or are triggered

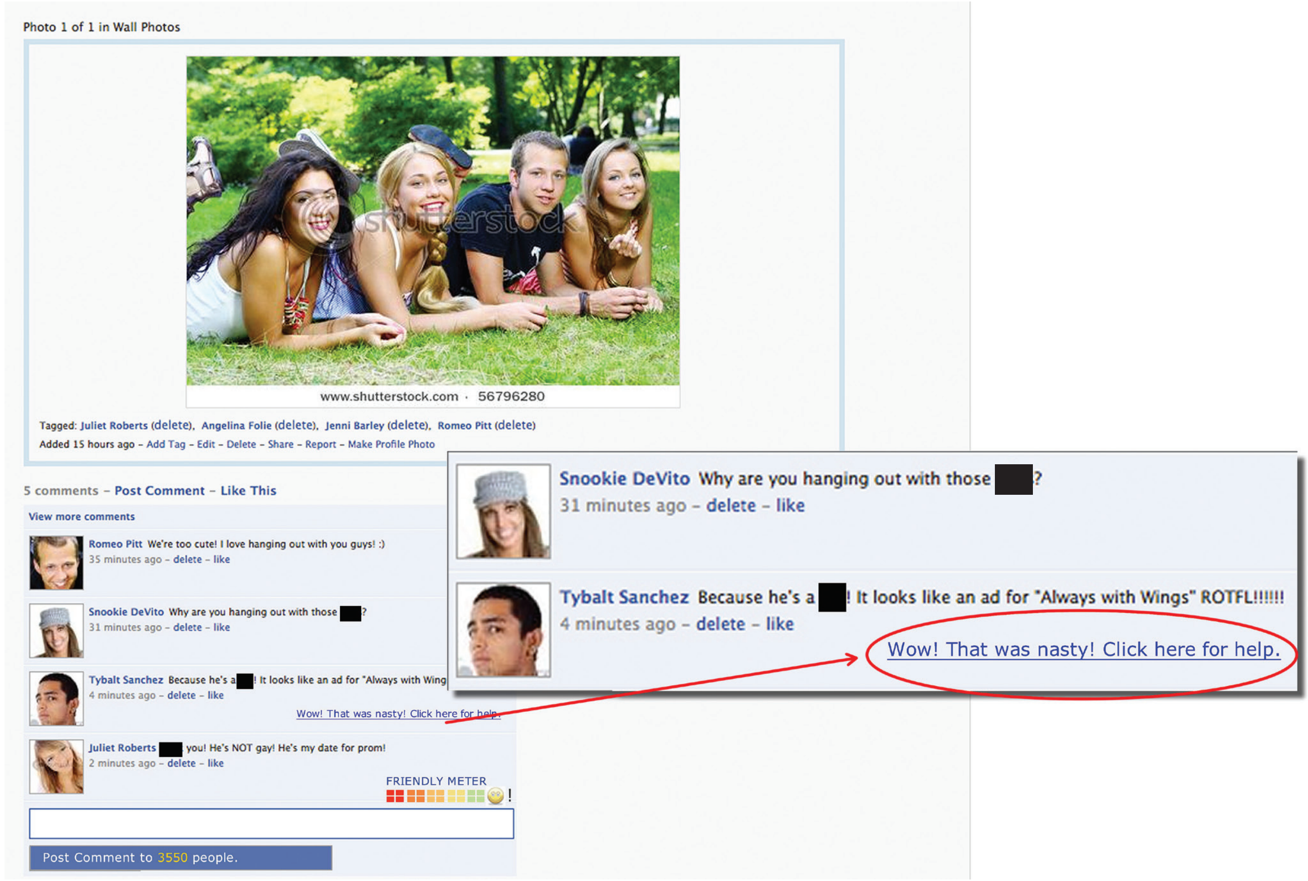


Photo 1 of 1 in Wall Photos

www.shutterstock.com · 56796280

Tagged: Juliet Roberts (delete), Angelina Folie (delete), Jenni Barley (delete), Romeo Pitt (delete)
Added 15 hours ago – Add Tag – Edit – Delete – Share – Report – Make Profile Photo

5 comments – Post Comment – Like This

View more comments

Romeo Pitt We're too cute! I love hanging out with you guys! :)
35 minutes ago – delete – like

Snookie DeVito Why are you hanging out with those [redacted]?
31 minutes ago – delete – like

Tybalt Sanchez Because he's a [redacted]! It looks like an ad for "Always with Wing
4 minutes ago – delete – like
Wow! That was nasty! Click here for help.

Juliet Roberts [redacted] you! He's NOT gay! He's my date for prom!
2 minutes ago – delete – like

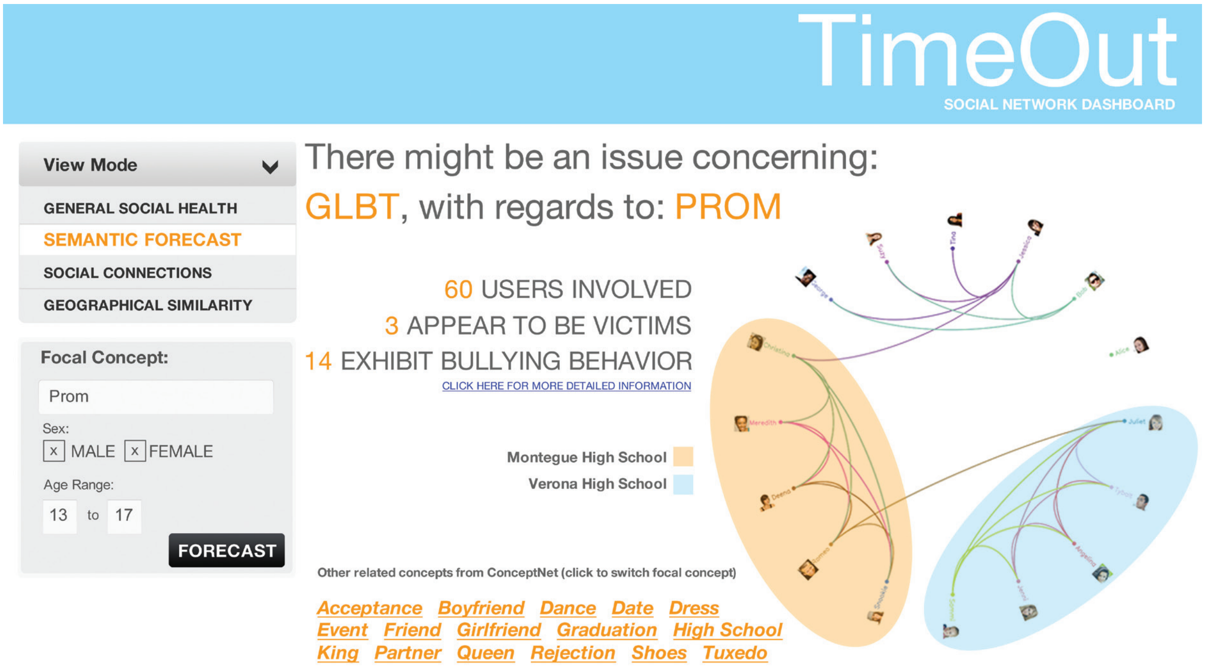
FRIENDLY METER
🟥🟥🟧🟧🟨🟨🟩🟩 !

Post Comment to 3550 people.

Snookie DeVito Why are you hanging out with those [redacted]?
31 minutes ago – delete – like

Tybalt Sanchez Because he's a [redacted]! It looks like an ad for "Always with Wings" ROTFL!!!!!!
4 minutes ago – delete – like
Wow! That was nasty! Click here for help.

Figure 2. When a possible cyberbullying message is detected, an automated intervention system could unobtrusively provide a link to educational material appropriate to the user's situation (offensive terms redacted by authors).



TimeOut
SOCIAL NETWORK DASHBOARD

View Mode
GENERAL SOCIAL HEALTH
SEMANTIC FORECAST
SOCIAL CONNECTIONS
GEOGRAPHICAL SIMILARITY

Focal Concept:
Prom
Sex:
☒ MALE ☒ FEMALE
Age Range:
13 to 17
FORECAST

There might be an issue concerning:
GLBT, with regards to: PROM

60 USERS INVOLVED
3 APPEAR TO BE VICTIMS
14 EXHIBIT BULLYING BEHAVIOR
CLICK HERE FOR MORE DETAILED INFORMATION

Montegue High School 🟧
Verona High School 🟦

Other related concepts from ConceptNet (click to switch focal concept)

Acceptance Boyfriend Dance Date Dress
Event Friend Girlfriend Graduation High School
King Partner Queen Rejection Shoes Tuxedo

Figure 3. SpeedBump helps a social network moderator visualize the community's connections, history, and topics.

by a particular event, like a high school prom.

At the March 2011 White House Conference on Bullying Prevention, President Obama said, "Today, bullying doesn't … end at the school bell—it can follow our children from the hallways to their cell phones to their computer screens. If there's one goal of this conference, it's to dispel the myth that bullying is just a harmless rite of passage or an inevitable part of growing up. It's not." (www.whitehouse.gov/blog/2011/03/10/president-obama-first-lady-white-house-conference-bullying-prevention)

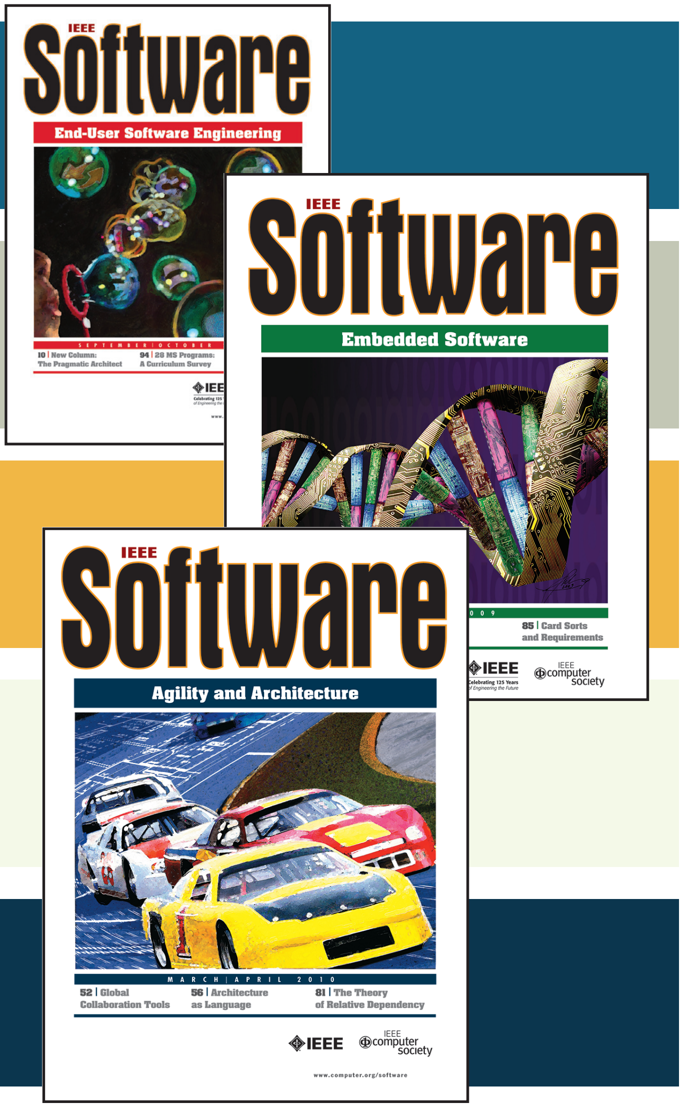
At its core, cyberbullying is really a people problem: no software can substitute for teaching kids how to have healthy personal relationships. But the novel design of social network software can help prevent and manage the problem.

__Henry Lieberman__ is a principal research scientist at the MIT Media Lab, where he heads the Software Agents Group. Contact him at lieber@media.mit.edu.

__Karthik Dinakar__ is a research assistant at the MIT Media Lab. Contact him at kdinakar@media.mit.edu.

__Birago Jones__ is a research assistant at the MIT Media Lab. Contact him at birago@media.mit.edu.

Editor: John Riedl, Department of Computer Science and Engineering, University of Minnesota; riedl@cs.umn.edu


IEEE
Software
End-User Software Engineering
SEPTEMBER | OCTOBER
10 | New Column: The Pragmatic Architect
94 | 28 MS Programs: A Curriculum Survey
IEEE
Software
Embedded Software
85 | Card Sorts and Requirements
IEEE
IEEE computer society
IEEE
Software
Agility and Architecture
MARCH | APRIL 2010
52 | Global Collaboration Tools
56 | Architecture as Language
81 | The Theory of Relative Dependency
IEEE
IEEE computer society
www.computer.org/software