Crowdsourced Ethics with Personalized Story Matching

Henry Lieberman
Karthik Dinakar
Birago Jones
MIT Media Lab
20 Ames St.
Cambridge, MA 02139 USA
{lieber, kdinakar, birago}
@media.mit.edu

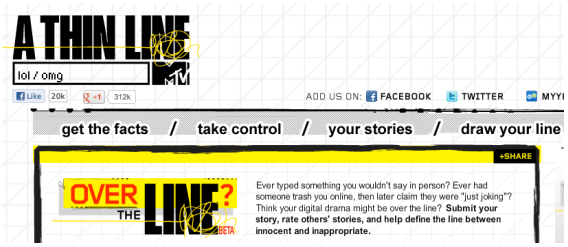

Figure 1. Intro to MTV's "Over the Line"

CHI'13, April 27 – May 2, 2013, Paris, France.
ACM 978-1-XXXX-XXXX-X/XX/XX.

Abstract

Cyberbullying is a widespread and growing social problem, threatening the viability of social networks for youth. We believe that one of the best ways to combat this problem is to use these incidents as "teaching moments", encouraging teens to reflect on their behavior and choices. Sites that offer community discussions around the ethical aspects of social situations can help teens feel less alone in their plight, and provide useful advice and emotional support. But the success of these "crowdsourced ethics" sites depends critically on the user feeling like discussions are relevant to their own personal experience.

We have augmented the crowdsourced ethics site "Over The Line", offered by MTV Networks, with a personalized story matcher that classifies stories according to dynamically discovered high-level themes like "sending nude pictures online" or "feeling pressure in relationships". The matcher uses a mixed-initiative LDA machine learning technique [2], and a commonsense knowledge base specialized to the bullying problem. The site is currently public, and attracts an audience of thousands of users daily.

Author Keywords
Cyberbullying; Crowdsourcing; Commonsense reasoning; Personalization; Machine Learning;

ACM Classification Keywords
H.5.3 Collaborative Computing

The Cyberbullying problem

Cyberbullying, harassment of youth by their peers on online networks, is a growing social problem [1], [9]. Studies show that almost half of young people experience it at some point [8], and extreme cases can lead to tragedies such as suicides.

Cyberbullying presents the same kind of threat to participation by youth on social networks, that spam did to e-mail in the early days of the Internet. If we allow it to get too prevalent, it will destroy the viability of social networks as a communication medium for youth. Just as spam filters provided a technical solution that, while they didn't eliminate it, kept the problem under control, we need a new generation of technical solutions adapted to this particular problem.

Reflective Interfaces

We would like to avoid heavy–handed approaches, such as censoring a post or banning the perpetrator. Our general approach to interface design is called *Reflective Interfaces [5]*. Reflective Interfaces are designed to encourage the user to think about why they made the choices they did; what the consequences are for themselves and others, and their options for future action. This is in contrast to conventional interfaces, which focus on just the immediate content of the interface choice or the communication taking place.

Reflective Interface design borrows its framework from principles espoused by Donald Schön [11]. Schön talked about the need for successive stages of action, then reflection, which in turn can have influence on future action.

Crowdsourced ethics: MTV's "A Thin Line" and "Over the Line" sites

While many sites offer advice, and many schools have anti-bullying programs, their effectiveness is limited. Advice tends to be vague ("Tell a trusted adult", "Don't let it get to you", etc.), unconnected to a child's personal situation, and delivered far from the times and places when bullying actually occurs.

However, in our opinion, one of the best current sites on this topic is MTV's "A Thin Line". It adopts a unique *crowdsourced ethics* approach that encourages dubious situations to be evaluated by the young people themselves.

The crowdsourced ethics approach has a number of advantages for helping youth cope with potentially bullying situations. Even the "Thin Line" title recognizes that it shouldn't try to preach a single correct answer. It encourages the user to reflect on how their particular experience relates to the experience of others.

Perhaps its greatest effect is to help the user realize that they are not alone, that their plight is shared by others, and that they can reach out to others for help. The most extreme negative reactions to bullying situations such as depression and even suicide, tend to occur only when a person feels like they're totally isolated in their predicament, and that there is no way out.

The top level of the "Thin Line" site (Figure 1) acts as a portal, offering education on bullying topics, help lines, sign up for anti–bullying campaigns, and a blog. The "Your Stories" link leads to the "Over the Line" page, which is concerned with personal stories contributed by users. Users can simply read stories contributed by others, and they can also rate them "over", "on", or "under" the line of acceptability. Users can also see the results of how others rated a particular story.

http://www.athinline.org/overtheline/

Figure 2. A story rated by the community

Personalization for crowdsourced ethics by understanding high-level themes

But the stories appear on the site with no particular topic organization. You can order them by recency, popularity, rating, or a (hand-selected) category "outrageous". But what's likely to make the stories most effective is to be able to find the ones that are most personally relevant for you.

Personalization for an application such as this is challenging. We can't rely on many traditional methods, such as keyword spotting, such as is typically used in spam filters, conventional topic modeling, and other information retrieval techniques. Just getting a superficial match to the particulars of the story is generally not helpful (even though a story mentions that it took place at a basketball game, other stories about basketball are probably irrelevant).

Our goal is, instead, to identify high-level themes of the story that are likely to be relevant to the topics and shared with other stories. This isn't the place to go into details of the story matching algorithm; for that see the references [3], [4].

Briefly, though, we employ a mixed-initiative variant of the machine learning technique Latent Dirichlet Allocation (LDA) [2]. LDA models a document as being produced from a set of underlying topics by a probabilistic process, and infers backwards from the actual words, to a distribution of topics. The topics can be discovered by this algorithm, and need not be specified in advance.

The mixed-initiative aspect is that, we ask people to describe each theme from the word cluster produced by LDA, then verify that it agrees reasonably well with a human interpretation of the story. This only needs to be done for each topic once a corpus is analyzed, not each time we interactively analyze a story. Figure 3 shows some of about 30 topics from a corpus of about 10,000 stories.

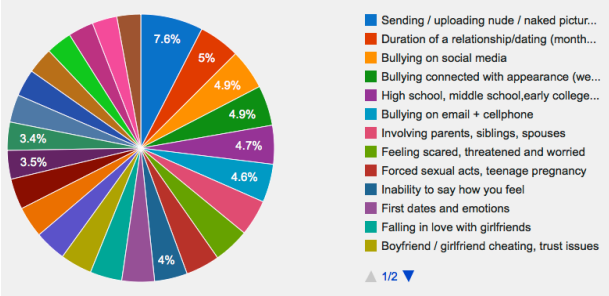

Figure 3. Theme distribution in the story corpus. 30 major themes were identified from 10,000 stories.

We also use a commonsense knowledge base [12] that contains over 1 million simple assertions about everyday life. For example, it might contain the assertion, "french fries make you fat", which helps us relate a story about eating too many french fries to the topic of insulting someone's appearance, a common bullying topic. We also augment the commonsense knowledge base with a more specialized knowledge base, BullySpace, containing specific knowledge about bullying topics, like racial and ethnic slurs.

In addition, we are also exploring finer grained methods of understanding the stories. While our current techniques relate to the topic of the entire story, we are also exploring modeling specific actors and events in the story, in order to recognize stereotypical story patterns like Schank's "scripts" [6], or Winston's Genesis system [6]. A companion paper [6] reports on some preliminary work.

Interaction with "Over the Line"

Figure 4 shows the screen with the system's response for finding stories similar to a given story. At top, the original story input by the user. Directly below it, a similar story retrieved by the system. The user can page through all matching stories, in order of relevance. The system also asks the user to rate the match (as well as rating the individual stories).

The "Resources" bar at the bottom is also personalized, using the computed story theme to filter the list of resources, taken from the other categories of information on the site. For example, if the inferred theme is, "pressure for having sex", the resource Dating Abuse Hot Line will be offered.

Related Work

Given the importance and ubiquity of this problem, it is surprising that there has been almost no prior work directly in the area of designing social network software to prevent and mitigate cyberbullying. There has been some work in educational games that simulate potentially bullying situations, e.g. [6], but these are fictional situations that do not directly relate to users' real-world lives.

Evaluation and Future Work

At the time of this writing, the site has just gone public, and we do not have sufficient data to report. Evaluation of the matching algorithm and some of the underlying techniques are reported in the references [3, 4]. By CHI press time, we will at least report some of the usage statistics collected by the public site. As seen in Figure 4, the site asks the user to rate the quality of the story match.

Figure 4. The story matching screen.

Other significant usage statistics are how often users invoke the story matcher, how many stories they examine with and without the matcher, whether reading matched stories increases the rate at which educational and help resource links are clicked, time spent on site, signup rates for the anti-bullying campaigns offered, etc. We are also conducting in-lab experiments where we can ask users directly whether they found the story matching personally valuable as opposed to just rating the match well.

We will continue to experiment with better ways of matching stories, interactive improvement of the knowledge base and machine learning techniques. We would also like to generate personalized advice dynamically by taking advice written for a general audience around one of the recognized story themes, and substituting in particulars of the user's story.

We are also exploring the potential for applications of these techniques to problems other than cyberbullying, such as support for autistic spectrum individuals.

Acknowledgements

We would like to thank Jason Rzepka of MTV for bringing our work to the public via their Over the Line site. We would like to thank Aneesh Chopra, former White House CTO, for sparking our collaboration at the 2011 White House Conference on Bullying Prevention.

We wish to acknowledge the financial support of the Simons Center for the Social Brain at MIT, and Formspring, as well as the over 70 corporate and organizational sponsors of the MIT Media Lab.

References

[1] Bazelon, Emily, *Sticks and Stones: Defeating the Culture of Bullying,* Random House, 2013.

[2] Blei, D., Ng, A., Jordan, M., Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, 993-1022, 2003.

[3] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R., Commonsense Reasoning for Detection, Prevention and Mitigation of Cyberbullying, ACM Transactions on Intelligent Interactive Systems (ACM TiiS), Volume 2, Number 3, 2012.

[4] Dinakar, K., Jones, B., Lieberman, H., Picard, R., Rosé, C. Thoman, M., Reichart, R., You Too?! Mixed-Initiative LDA Story Matching to Help Teens in Distress, International Conference on Weblogs and Social Media (ICWSM-12), Dublin, Ireland, June 2012.

[5] Jones, B., Reflective Interfaces: Assisting Teens with Stressful Situations Online, MS Thesis, MIT Media Lab, 2012.

[6] Macbeth, J., Adeyama, H., Lieberman, H., Fry, C., In-Depth Story Matching for Cyberbullying Prevention, submitted to CHI Works in Progress, 2013.

[7] Mancilla-Caceres, J.F., Pu, W., Amir, E., and D. Espelage. *Identifying Bullies with a Computer Game.* In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI-12), 2012.

[8] MTV Networks and Associated Press. (2011). MTV-AP Digital Abuse Survey. Retrieved from http://www.athinline.org/pdfs/2011-MTV-AP_Digital_Abuse_Study_Full.pdf

[9] Olweus D. Bullying at School: *What We Know and What We Can Do.* Oxford, England: Blackwell; 1993.

[10] Schank, R., Abelson, R., *Scripts, Plans, Goals and Understanding*, Lawrence Erlbaum, 1977.

[11] Schön, D. (1983) *The Reflective Practitioner, How Professionals Think In Action*, Basic Books.

[12] Speer, R., Havasi, C., Lieberman, H., AnalogySpace: Reducing the Dimensionality of Commonsense Knowledge, Conference of the Association for the Advancement of Artificial Intelligence (AAAI-08), Chicago, July 2008.

[13] Winston, P., The Strong Story Hypothesis and the Directed Perception Hypothesis, AAAI Fall Symposium on Advances in Cognitive Systems, 2011.