ILLINOIS INSTITUTE OF TECHNOLOGY
impact

Nayar Prize II, Phase I Quarterly Progress Report (*Q4*)

October 2017

Project: **A Cyberbullying Early-Warning and Response System**
Team: **Aron Culotta, Libby Hemphill**
Students: **Joshua Guberman, Ping Liu**

Progress Summary of Nayar Prize II, Phase I

Highlights:

The key accomplishments achieved during Phase I of this Nayar Prize II project are as follows:

(1.) Conducted dozens of interviews with parents, school administrators, and law enforcement to understand where cyberbullying takes place, what types of cyberbullying are the most important to detect, who is the most appropriate person to alert, and how the output of a forecasting system should be communicated to relevant stakeholders.

(2.) Collected more than 15 million comments from more than 400,000 Instagram posts, of which more than 30,000 comments have been manually annotated for the presence of cyberbullying content to train and validate the forecasting model.

(3.) Designed and implemented several competing machine learning models to forecast the presence and intensity of cyberbullying in Instagram posts. The best method is able to predict the appearance of a hostile comment on a post ten or more hours in the future with 77% accuracy and can furthermore distinguish between high and low levels of future hostility with 87% accuracy.

(4.) Designed and implemented a web-based user interface that allows parents to monitor the Instagram accounts of their children's online community and receive alerts of imminent cyberbullying threats. The prototype is accessible at http://nayar.casmlab.org with the username "libbyh@gmail.com" and password "password".

(5.) Submitted one NSF proposal and one computer science research article based on this work, with a second NSF proposal under development.

These accomplishments cover all of the milestones proposed for Year 1, as well as some of the milestones proposed for Year 2. In this document, we will first provide more details on the main accomplishments of Year 1, followed by an outline of the research plan for Year 2.

Objectives:

Bullying is a widespread public health issue with grave short- and long-term consequences, including physical violence, depression, and substance abuse (Hawker and Boulton 2000). As social connections have migrated online, so has bullying—a third of teenage internet users have been victims of *cyberbullying* (Duggan 2014), the 21st century version of bullying through electronic means such as online social networks. The anonymity, wide distribution, and 24/7 nature of cyberbullying pose novel threats to wellbeing and present new challenges to prevention and response. In addition to online verbal harassment, cyberbullying in many cases leads to more severe repercussions, including stalking, invasion of privacy, and destruction of property. Government efforts have recently begun engaging parents and school administrators through education and training programs ("Home" 2012), yet cyberbullying remains prevalent and will likely continue to rise as online social networking becomes a central part of life.

Addressing this new digitally mediated form of bullying requires new technology. As capabilities such as natural language processing, machine learning, and social network analysis have matured, it has now become feasible to monitor online social interactions to forecast future events. The broad, long-term objective of this project is to develop software tools to forecast imminent cyberbullying threats and vulnerabilities in online social networks. Such forecasts can then be used by individuals and communities to take timely actions to avoid escalation and reduce further harm. The specific objectives are as follows:

1. Implement a cross platform application that predicts imminent cyberbullying threats in real time by monitoring a user's online social interactions.
2. Optimize the accuracy and efficacy of the approach by modeling linguistic and social histories of users' interactions using machine learning.
3. Conduct in-depth user studies and experimental validation to ensure the validity of the approach and to determine the appropriate intervention protocols.

Phase One Summary:

Engagement with Stakeholders

We have purposefully approached this problem from a holistic sociotechnical perspective—while creating the appropriate predictive analytics technology is important, understanding who will use the system and how are just as critical to success. The social complexity stems in part from the diverse perspectives of stakeholders (students, parents, school administrators, law enforcement) as well as from the many sensitive topics that arise, including user privacy, freedom of speech, and discrimination. We have thus spent significant effort conducting interviews with parents, school officials, and law enforcement to ensure the maximum impact of the final system. Through these interviews, we answered several important questions to guide our approach.

Our approach:

- **Which online social networks are the most important to monitor?** We learned that SnapChat, Instagram, and Twitter (in that order) are the top three sites of importance to the stakeholders, both because of their popularity and the prevalence of bullying. Because of this, we updated our Y1 plans to focus first on Instagram, reserving Twitter for Y2. As SnapChat data is closed and private, we do not plan to consider it for this project.
- **What type of cyberbullying is the most important to detect?** The academic literature has many competing definitions of cyberbullying, so we asked the stakeholders to help us construct an operational definition that is most meaningful to them. We quickly learned that both the academic and mainstream press have mischaracterized the kinds of cyberbullying that have the greatest impacts on the day-to-day lives of high school students. Whereas anonymous attacks from unknown persons online have gained a lot of attention, they are of low priority to stakeholders in part because they are exceedingly rare. Instead, the stakeholders were unanimous in describing the sorts of harassment that most concerned them: situations in which students who know each other offline escalate a social conflict through online communications, often culminating in a physical altercation. This insight informed our data collection methodology to focus specifically on users that likely know each other offline.
- **Who should be the users of the system?** While we initially planned that school officials would be the natural stakeholders to notify of imminent cyberbullying threats, after interviews we have discovered a number of drawbacks to this approach. First, schools have limited resources to respond to such alerts; second, they are wary of assuming new liabilities; and third, alerting the schools frames the system as a potentially punitive approach, raising concerns of privacy and profiling. Instead, we have identified parents as much more suitable users of our forecasting system. Parents clearly have an incentive to respond to cyberbullying threats to their children, and by designing the system to be “opt-in,” we can alleviate many privacy concerns.
- **What form should the cyberbullying alert take?** Once the system has predicted an imminent cyberbullying threat, we must determine how best to communicate this to the parent. In interviews, parents expressed desire for an email or text message alerting them of potential incidents. Parents were emphatic in their desire for only limited information about the incident specifics—they are sensitive to allowing their children space to interact outside of their supervision. Instead, what they most want is a prompt from our system so that they know when to check in with their children about potential problems. We find this to be a much more practical and impactful type of intervention—rather than involving

school officials with the potential for a punitive response or signaling potential bullies out for shaming or attention, we enable parents to discuss ongoing issues with their children, letting them collaborate on a response that is appropriate to the situation. Indeed, research suggests that one of the biggest dangers of cyberbullying is that the victims often “suffer in silence,” because they are ashamed, fear retaliation, or feel that no one will believe them. Thus, we view our system as a “conversation starter” between parents and their children, providing a low-friction method to open a dialogue at just the right moments to prevent escalation of online harassment. The results from these interviews have guided all of the technical components of the system. Below, we describe our efforts in Year 1 to build and evaluate a working prototype of a cyberbullying forecasting system.

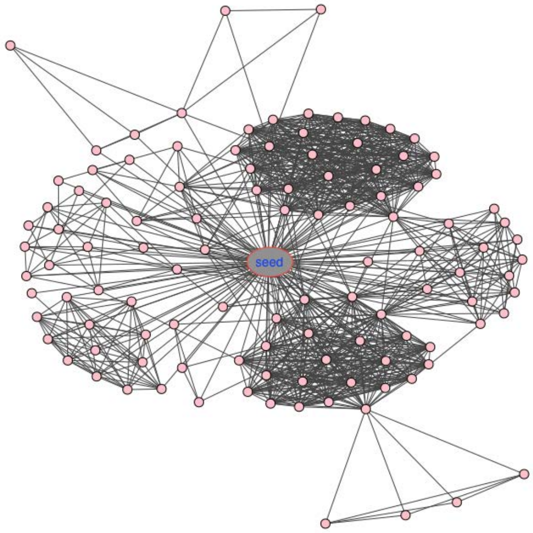
Forecasting Tasks

Based on our discussions with stakeholders, we defined hostile comments as those containing harassing, threatening, or offensive language directed toward a specific individual or group. We have initially focused our data collection on Instagram. On this site, a user creates a post by uploading an image. Each post has a thread of associated comments. We find that the conversation attached to a post often takes place over long periods of time (over two months, on average). Our goal is to develop a machine learning method that can analyze the linguistic and social information in the initial comments of a post to forecast future hostility. We formulate two specific forecasting tasks:

1. **Hostility presence forecasting:** Given the initial sequence of *non-hostile* comments in a conversation, predict whether some future comment will be hostile.
2. **Hostility intensity forecasting**: Given the first *hostile* comment in a post, predict whether the post will receive more than *N* hostile comments in the future. By decomposing the problem into these tasks, we provide flexibility in how the system will be used. For example, some users may want to be alerted whenever there is a high risk of future hostilities. Other users may instead only want to be alerted when a rapid escalation of hostilities is predicted.

Data Collection and Annotation:

Community Identification. Because we are targeting interactions between users who know each other offline, we developed an algorithm that can identify, for a given user, a set of users that are likely to be known to this user. To do so, we construct a graph of users, where a weighted edge between users indicates how often they have participated in the same conversation. Starting with a seed user, we expand the user set by including directly connected users, as well as second and third degree users (e.g., “three degrees of separation”).



Because this quickly leads to a large set of users, many only indirectly related to the seed user, we then apply a clustering algorithm based on the density of edges connecting nodes in the graph. This identifies a community centered on the seed user. The figure above shows a sample community of about 100 users identified for one seed user. We can see that the user participates in a few distinct sub-communities. The algorithm identifies users that the seed user may have not yet communicated with, but is likely to do so in the future given the high “friends of friends” overlap. We then continuously crawl all the available posts and their associated comments for each user in the discovered community. This gives us a target set of users to track when forecasting future bullying activities.

Data Annotation. Training and evaluating machine learning models requires a large and representative dataset for which the ground truth labels are known. Using a seed set of users from local high schools, we used the community identification method above to expand the set of users; then we crawled public Instagram posts and the associated comments. In total, we collected over 15 million comments from over 400 thousand posts. We then manually annotated a representative sample of this data to indicate which comments contain hostility. We annotated over 30 thousand comments from over 1,100 posts using Amazon’s Mechanical Turk, a popular crowdsourcing platform. To ensure data quality, we required each annotator to undergo a thorough training program that instructed them about what should and should not be considered hostility. We also had at least two different annotators label each comment, adding a third annotator to break ties. Table 1 shows statistics of this data.

	Posts	Comments	Hostile Comments
Hostile Posts	591	21,608	4,083
Non-Hostile Posts	543	9,379	0
Total	1,134	30,987	4,083

Table 1: Statistics of the manually annotated training and validation data. “Hostile” posts are those with at least one hostile comment.

Investigating this data revealed several insights:

- The commenting feature is heavily used on Instagram. On average, each post received about 27 comments, though many received well over 100.
- Comments on a post take place over extended time periods. On average, the final comment on a post was written 80 days after the post was originally created. This suggests that, far from being a place just to post photos, Instagram posts serve as meeting grounds for users to conduct long-term conversations on a topic.
- Examining the chronology of when hostile comments appear in each post, we find that the volume and temporal frequency of hostile comments varies substantially between posts, which we investigate further in the next section.

Discovering Temporal Patterns of Hostility. To better understand the opportunity for a technological intervention in forecasting online hostilities, we performed a cluster analysis to identify groups of posts that exhibit similar temporal characteristics. To do so, we construct a time series for each post by counting the number of hostile comments posted within each hour from the time the post was created. Thus, each time series indicates how many hostile comments appear over the lifespan of a post. We then cluster the time series using an algorithm (K-Spectral Centroid) that is invariant to scaling and shifting, enabling us to identify posts that have similarly shaped time series of hostile post frequency. Table 2 below shows the results.

These results provide some insight into what sorts of interventions may have an impact and what their limitations may be. For example, posts from cluster 0 appear to be a low priority for intervention—the hostile comments are isolated incidents that occur too quickly for any preventative action. In contrast, in clusters 1 and 3 the time lag before hostile comments appear presents an opportunity to intervene before hostilities escalate—e.g., by contacting parents, blocking comments, taking down a post, or other means. This motivates our first task of forecasting a future hostile comment in a post. Additionally, while the first hostile comment appears similarly quickly in clusters 0 and 2, if we could distinguish between the two after the first hostile comment appears, we could prioritize which posts require intervention. This motivates our second task of forecasting the total volume of hostile comments a post will eventually receive.

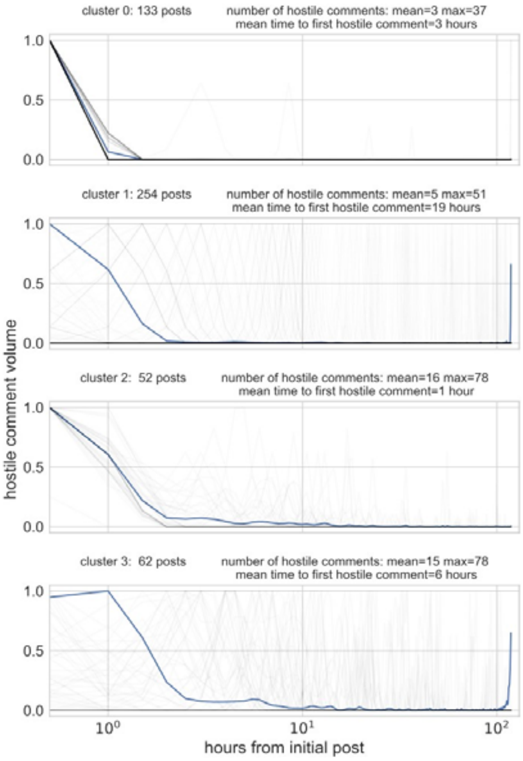
	Low hostility at start of conversation.
	Low hostility in middle of conversation.
	High hostility at start of conversation.
	High hostility in middle of conversation.

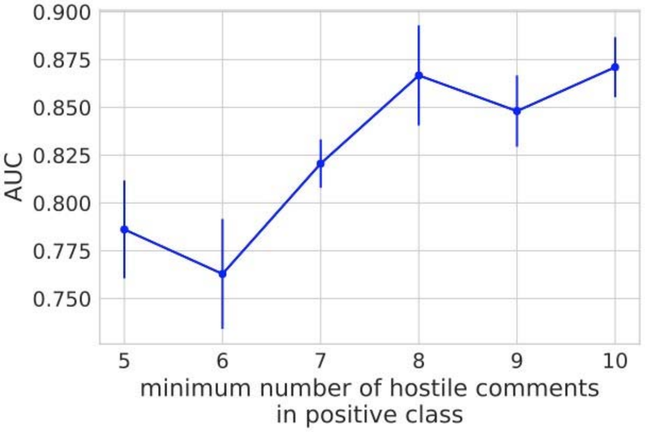
Table 2: Examining all posts containing hostile comments, we find four distinct clusters according to the volume and onset time of hostile comments.

Hostility Presence Forecasting. We implemented a machine learning model to forecast the presence of a hostile comment on an Instagram post. To do so, we experimented with a number of different sources of evidence, including linguistic information from previous comments in the post, prior commenting behavior on other posts for the users involved in the post so far, as well as trend information that quantifies the verbal trajectory of a conversation. Since our intervention protocol involves alerting the parent of an imminent threat, we conducted experiments where we varied the lead time: the time between when the forecast is made and when the first hostile comment appears. As in other forecasting domains, the accuracy of the forecast declines the farther into the future we must predict.

The figure above shows accuracy of the model with four feature sets as lead time increases (AUC is area under the ROC curve, a common metric to consider the rate of false positives and false negatives). We can interpret the results as follows: Suppose we have two Instagram posts where we observe comments up to time t. One post will receive a hostile comment ten or more hours in the future (positive example); the other post will not (negative example). The best system we have built will assign a higher threat score to the positive post 77% of the time. We can see that this best system improves over our initial baseline (the red dashed line) by 8% absolute, indicating that the linguistic and social features have greatly improved performance. In the following section, we discuss our plan for further improving these results in Year Two.

Hostility Intensity Forecasting

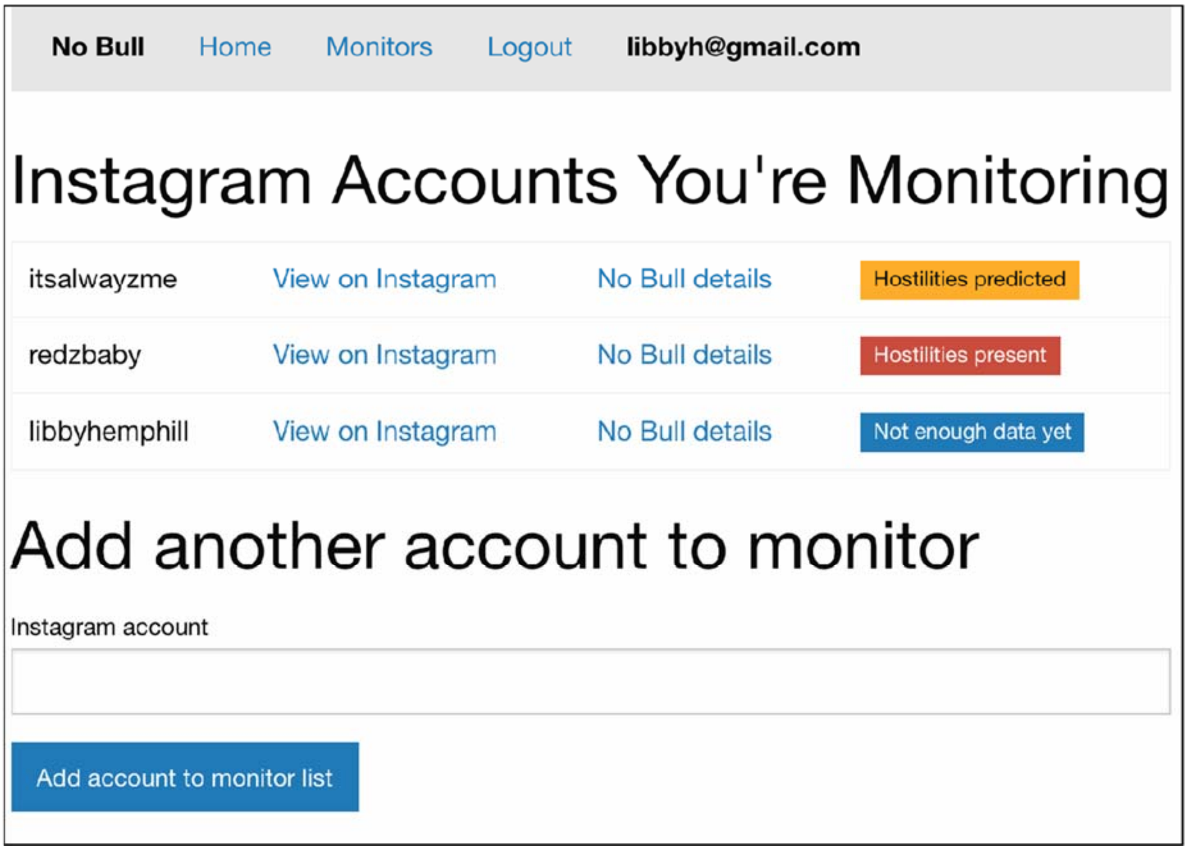
In some conversations a single hostile comment is ignored, while in others it sparks a rapid escalation in hostilities. In this second task, we observe comments up to and including the first hostile comment on a post. We then forecast whether the post will receive more than *N* hostile comments in total. We use similar sources of evidence as in the prior task, but we also include features of that first hostile message. Thus, we attempt to isolate qualities of hostile comments that lead to escalations. The adjacent figure shows AUC as we vary the value of *N*. E.g., for *N=8*, our model predicts whether a post will receive one hostile comment versus eight or more hostile comments. We can see that the model is more accurate for this task, indicating that there are distinguishing characteristics of high intensity conversations. We also investigated the model coefficients for different features to determine what the most predictive indicators are. We find that features about the *author* of the post are most predictive—that is, users who have received hostile messages in the past are likely to receive them in the future. Furthermore, we identified several indicators of high intensity conversations: e.g., terms like “stfu” (“shut the f**k up”) are often used just prior to escalations. Finally, we find that female-specific insulting terms are associated with higher intensity. Taken together, these results suggest that our forecasts can be improved by including a richer user representation (e.g., demographic attributes), which we describe more in Section 3.



User Interface

Every parent we talked to already had conversations with their children about what was happening at school and how they were using social media, and they also had access to their children's phones and social media accounts. They thought (a) the predictions our system provided and (b) information about their children's networks would be useful for talking to their children about what was happening in their social circles. We built a working prototype of the parent interface that has two main components: (1) a list of accounts the parent/user is monitoring and (2) detailed information about each account. The prototype is accessible at http://nayar.casmlab.org with the username "libbyh@gmail.com" and password "password".

In each view, we indicate whether hostilities are present (our model detected them already), hostilities are likely (our model indicates at least one hostile comment will appear soon), or if a conversation is benign. We do not provide details about specific posts or users because parents thought that was too much information for their needs and was potentially problematic if predictions or labels were wrong (and hostilities weren't actually present or imminent).

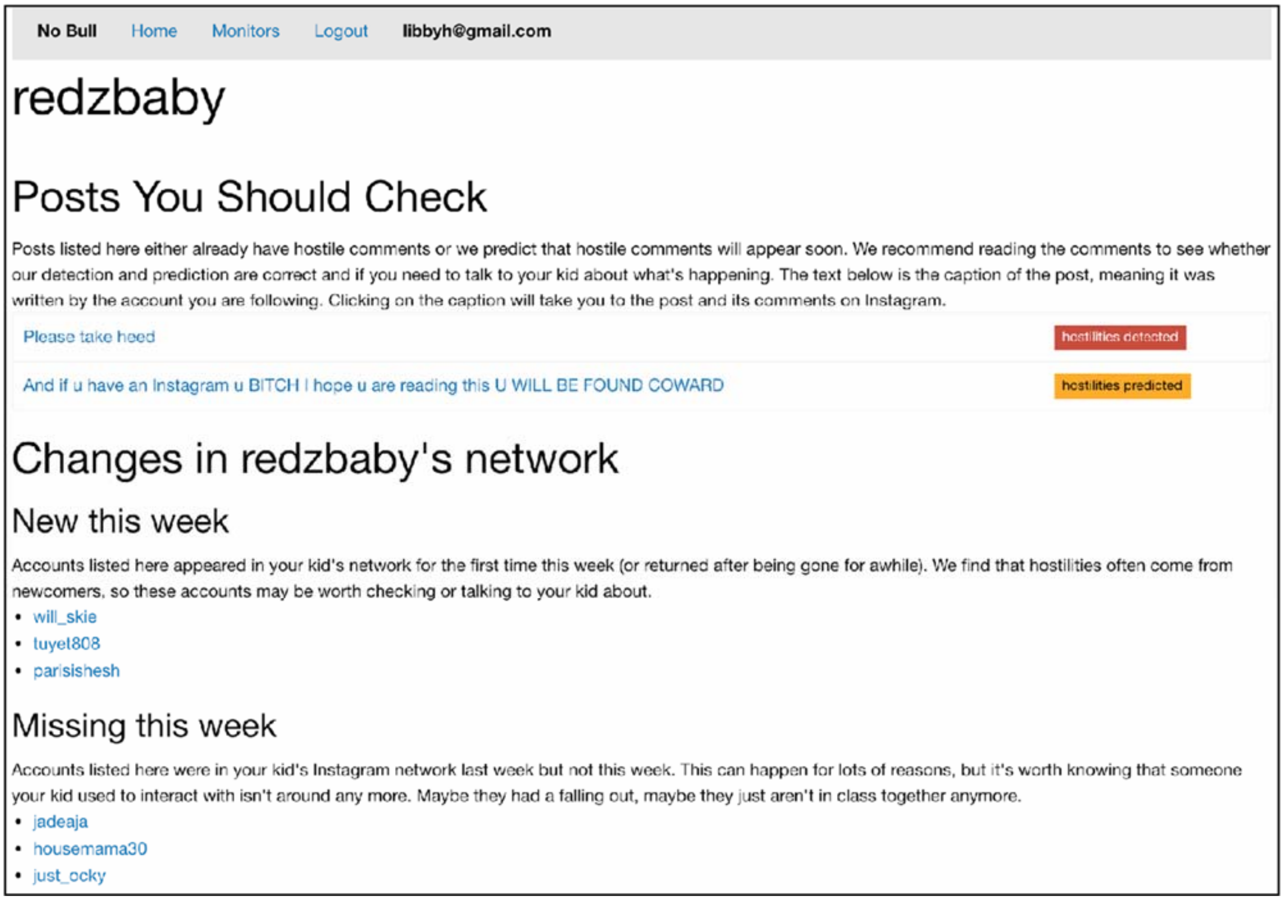


Instagram Accounts You're Monitoring

itsalwayzme	View on Instagram	No Bull details	Hostilities predicted
redzbaby	View on Instagram	No Bull details	Hostilities present
libbyhemphill	View on Instagram	No Bull details	Not enough data yet

Add another account to monitor

Instagram account



redzbaby

Posts You Should Check

Posts listed here either already have hostile comments or we predict that hostile comments will appear soon. We recommend reading the comments to see whether our detection and prediction are correct and if you need to talk to your kid about what's happening. The text below is the caption of the post, meaning it was written by the account you are following. Clicking on the caption will take you to the post and its comments on Instagram.

Please take heed	hostilities detected
And if u have an Instagram u BITCH I hope u are reading this U WILL BE FOUND COWARD	hostilities predicted

Changes in redzbaby's network

New this week

Accounts listed here appeared in your kid's network for the first time this week (or returned after being gone for awhile). We find that hostilities often come from newcomers, so these accounts may be worth checking or talking to your kid about.

- will_skie
- tuyet808
- parisishesh

Missing this week

Accounts listed here were in your kid's Instagram network last week but not this week. This can happen for lots of reasons, but it's worth knowing that someone your kid used to interact with isn't around any more. Maybe they had a falling out, maybe they just aren't in class together anymore.

- jadeaja
- housemama30
- just_ocky

Plans for Phase II:

We have identified three main areas for future work: improving our existing model, adapting our model to new platforms, and developing socio-technical mechanisms to respond to hostilities.

Improve Instagram Model

Our existing model achieves 77% accuracy at predicting the appearance of a hostile comment on a post ten or more hours in the future and 87% accuracy distinguishing between posts that will receive high and low levels of hostility. Those performance measures are quite good, but we recognize room for improvement in three ways. First, given the predictive power of the level of hostility an author’s previous posts received, we assume that something about the user makes their posts more likely to receive hostile comments. We propose inferring more detailed user attributes (e.g., age, gender, ethnicity) to improve the model. Second, as we mention above, four patterns emerged from the distributions of hostile comments over time. Being able to distinguish posts with cluster 1 or cluster 3 hostility patterns would enable us to develop more effective interruption mechanisms. We need additional instances of all conversation pattern types in order to improve our ability to distinguish between those clusters and therefore plan to label more conversations. Third, because Instagram is a photo and video sharing site, features of the images and videos may be useful for predicting hostilities in the comments. We will investigate image classification methods to improve our prediction models, building on the latest research in deep neural networks.

Adapt Instagram Model to New Domains

Both survey research and our interviewees have indicated that many children have accounts on multiple social networking websites. Thus, addressing cyberbullying on only one platform will have limited impact, and is one of the motivating factors for our investigation into a general purpose forecasting tool. We will develop methods that enable the forecasting tool to be applied to new data sources—e.g., our system trained on Instagram should be able to be effective on YouTube without further algorithm development. In machine learning, this is often referred to as domain adaptation or multitask learning, in which a model trained on one dataset is then applied to a new dataset drawn from a different source (Daume and Marcu 2006). Parents and school officials suggest that Twitter is the social network platform on which we should concentrate, and we will test our existing model's performance on Twitter conversations and then identify Twitter-specific features that could improve its performance there.

Expand Response Mechanisms

Our model addresses crucial steps in curbing cyberbullying—detecting and predicting hostile interactions —but it does not provide mechanisms for responding to hostilities. We plan to leverage our model's success to develop socio-technical response mechanisms by testing and improving the Monitoring UI and by experimenting with automated de-escalation strategies. For example, we will enroll parents from our interview study in a field study of the Monitoring UI. Parents will be able to monitor specific Instagram accounts, and we will survey them periodically about the performance of the prediction model and their use of the monitoring system. We are interested in whether they agree with our predictions and receive enough information from the UI to facilitate the conversations they want to have with their children. We can use their feedback on both fronts to improve the model and the Monitoring UI.

The Monitoring UI is a passive mechanism for enabling social responses to hostilities. We are interested in experimenting with active, automatic mechanisms for addressing hostilities as they occur. For instance, are there particular messages that, when posted in response to a hostile comment, reduce the likelihood of future hostilities? Are there ways to pause or eliminate comments on particular posts in order to prevent conversations from turning hostile?

We plan to investigate with various types of experiments on Instagram or other platforms.

Budget:

The budget will support one full-time graduate student from the Computer Science Department and one full-time undergraduate student from Lewis College of Human Sciences, summer support for each PI, conference travel to present findings, and data collection and annotation costs.

Salaries and Wages

Graduate Student Research Assistant: One graduate student is budgeted for one full year. The student will be selected from the Computer Science Department and will help the PIs in data collection, software development, and software testing and validation. Additionally, the graduate student will help mentor the undergraduate student.

Undergraduate Student Research Assistant: One undergraduate student is budgeted for one semester. The student will be selected from Lewis College and will aid in data collection, annotation, and testing.

Aron Culotta, Principal Investigator (PI): One summer month is budgeted for one year. The PI will be responsible for directing and administering the proposed work, particularly the cyberbullying prediction algorithm development. The PI will also be responsible for mentoring the students.

Libby Hemphill, Principal Investigator (Co-PI): One-half summer month is budgeted for one year. The PI will be responsible for directing and administering the proposed work, particularly data collection, annotation, usability testing, and interaction with stakeholders.

Fringe Benefits

Fringe benefit rate is 23.8% for academic year salary, 7.9% for the summer month salary, 24.5% for staff salary, and 0.0% for student salary.

Travel

We have budgeted for two researchers to attend the Association for the Advancement of Artificial Intelligence (AAAI) conference, where they will present the details of our model.

Other Direct Costs – Materials and Supplies

Twitter data access: We have budgeted $11,500 to access historical Twitter data from Gnip. This historical data is needed to identify previous cases of cyberbullying that we will use to train the cyberbullying detection algorithm.

Data annotation: We have budgeted $10,000 for data annotation, which will primarily be used to support online services like Amazon Mechanical Turk to help annotate cyberbullying events. We will also use these funds to do user studies of the application once the prototype is completed.

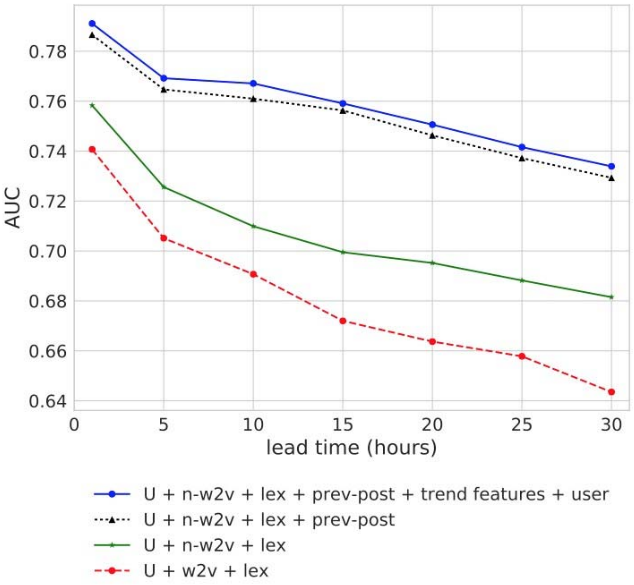
References

Daume, Hal, III, and Daniel Marcu. 2006. “Domain Adaptation for Statistical Classifiers.” *The Journal of Artificial Intelligence Research* 26: 101–26.

Duggan, Maeve. 2014. “Online Harassment.” Pew Research Center. http://www.pewinternet.org/files/2014/10/PI_OnlineHarassment_72815.pdf.

Hawker, D. S., and M. J. Boulton. 2000. “Twenty Years’ Research on Peer Victimization and Psychosocial Maladjustment: A Meta-Analytic Review of Cross-Sectional Studies.” *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 41 (4): 441–55.

“Home.” 2012. *Stopbullying.gov*. Department of Health and Human Services. February 17. http://www.stopbullying.gov/.


AUC
lead time (hours)
U + n-w2v + lex + prev-post + trend features + user
U + n-w2v + lex + prev-post
U + n-w2v + lex
U + w2v + lex