Detecting the Presence of Cyberbullying Using Computer Software

Jennifer Bayzick
Ursinus College
601 E. Main Street
Collegeville PA 19426-1000 |
1.610.409.3000
jennifer.bayzick@gmail.com

April Kontostathis
Ursinus College
601 E. Main Street
Collegeville PA 19426-1000 |
1.610.409.3000
akontostathis@ursinus.edu

Lynne Edwards
Ursinus College
601 E. Main Street
Collegeville PA 19426-1000 |
1.610.409.3000
ledwards@ursinus.edu

ABSTRACT

Cyberbullying is willful and repeated harm inflicted through the medium of electronic text. Computer software was developed to detect the presence of cyberbullying in online chat conversations. Rules based on a dictionary of key words are used to classify a window of posts. A truth set of MySpace threads was created. The software was found to correctly identify windows containing cyberbullying 85.30% of the time, and it identifies innocent windows 51.91% of the time. The overall accuracy is 58.63%. This suggests that our coding rules must be refined to not falsely flag so much innocent conversation.

Categories and Subject Descriptors

K.4.2 Social Issues

General Terms

Algorithms, Legal Aspects

Keywords

Cyberbullying, rule-based method, truth set.

1. INTRODUCTION

We developed a program, BullyTracer, to detect the presence of cyberbullying in online conversation. It is a rule-based system and we also developed a truth-set to determine the accuracy of the system. Our program was found to correctly identify 58.63% of the windows in the dataset. Though this is not accurate enough to implement in monitoring software, this research has provided background knowledge about the language used in cyberbullying conversation, as well as the first truth-set that can be used to test cyberbullying detection algorithms.

2. CYBERBULLYING DEFINED

Patchin and Hinduja define cyberbullying as “willful and repeated harm inflicted through the medium of electronic text [3].” Willful harm excludes sarcasm between friends comments meant to criticize or disagree with an opinion but not meant to attack the individual. Nine different types of cyberbullying were identified [1][2][4]:

Flooding consists of the bully monopolizing the media so that the victim cannot post a message [2].

Masquerade involves the bully logging in to a website, chat room, or program using another user’s screenname to either bully a victim directly or damage the victim’s reputation [4].

Flaming, or **bashing**, involves two or more users attacking each other on a personal level. The conversation consists of a heated, short lived argument, and there is bullying language in all of the users’ posts [4].

Trolling, also known as **baiting**, involves intentionally posting comments that disagree with other posts in an emotionally charged thread for the purpose of provoking a fight, even if the comments don't necessarily reflect the poster’s actual opinion [1].

Harassment most closely mirrors traditional bullying with the stereotypical bully-victim relationship. This type of cyberbullying involves repeatedly sending offensive messages to the victim over an extended period of time [4].

Cyberstalking and **cyberthreats** involve sending messages that include threats of harm, are intimidating or very offensive, or involve extortion [4].

Denigration involves gossiping about someone online. Writing vulgar, mean, or untrue rumors about someone to another user or posting them to a public community or chat room or website falls under denigration [4].

Outing is similar to denigration, but requires the bully and the victim to have a close personal relationship, either online or in-person. It involves posting private, personal or embarrassing information in a public chat room or forum [4].

Exclusion, or ignoring the victim in a chat room or conversation, was the type of cyberbullying reported to have happened most often among youth and teens [3].

3. DATASET

Our dataset consists of thread-style forum transcripts crawled from MySpace.com, where a general topic is included in the creation of the thread. Many users can post in the thread, and conversation usually deviates from the starting topic. When working with these conversations, we considered a post to be a single body of chat text posted by a user at one time. The body of text could contain multiple sentences or even multiple paragraphs and still be considered a single post. Because of the interactive nature of cyberbullying, the conversations were processed using a moving window of 10 posts to capture context.

Undergraduate research assistants developed a truth set for testing our algorithms. Three individuals reviewed each window and indicated whether or not cyberbullying is present, a window was considered to contain cyberbullying if two or more humans flagged it as such.

WebSci '11, June 14-17, 2011, Koblenz, Germany.

4. BULLYTRACER

BullyTracer is a program designed to detect the presence of different types of cyberbullying in a chat room conversation. We describe its first version in this section. This paper lays the groundwork for further examination of the linguistic components of a cyberbullying conversation, the distinction between various types of cyberbullying, and the algorithms used to detect the presence of cyberbullying language. BullyTracer analyzes all files in given directory using a rule-based Algorithm.

BullyTracer uses a dictionary of code words that fall into the categories: insult word (retarded, dumb), swear word (bitch, fucker), and second person pronouns (you, your). BullyTracer marks each post in a window with the category of any words found in the dictionary. These categories were chosen because they seemed to have the highest correlation to the presence of cyberbullying in a chat post. Insults and swear words indicate hostility and mean-spiritedness from the user who posted them. Second person pronouns help to distinguish the object of the nasty words.

Another indication of hostile language is the use of many capital letters. General use of capitals at the beginnings of sentences or sparingly is normal, but if the percentage of capital letters to lowercase letters is greater than 50%, the post is considered to contain cyberbullying.

A window is labeled as containing cyberbullying if it contains any cyberbullying posts.

5. EVALUATION AND RESULTS

Evaluation of BullyTracer's coding rules was done by comparing the program's decisions to the human-defined truth set for each window. The program counts the number of correctly identified windows, as well as the number of windows that are innocent but identified as containing cyberbullying (false positives) and the number of windows that contain cyberbullying but were incorrectly identified as innocent conversation (false negatives).

Overall, the BullyTracer coding decisions match the human truth set 58.63% of the time (Table 1). Percentages of correct coding vary by packet between 32.32% and 83.97%. Of the 415 windows that actually contain cyberbullying, BullyTracer labeled 354 of them correctly. The program rarely incorrectly identifies a window that the truth set labels as containing cyberbullying, which says that our coding rules seem to capture the essence of cyberbullying conversation. The program is less able to identify innocent conversation. Of the 1647 innocent windows, BullyTracer codes 855 of them correctly. This suggests that our coding rules are too broad and need to be refined.

6. CONCLUSIONS

This project defines nine types of cyberbullying and proposes methods to detect the presence of cyberbullying in online chat conversations. It also describes the first implementation of algorithms to detect the presence of cyberbullying. BullyTracer was found to correctly identify windows containing cyberbullying 85.30% of the time, and it identifies an innocent window correctly 51.91% of the time. Further research is necessary so as to not falsely flag so much innocent conversation.

The most important contribution of this work was the creation of the MySpace cyberbullying truth set. There was previously no truth set in existence for cyberbullying.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0916152. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] *Glossary of cyberbullying terms*. (2008, January). Retrieved from http://www.adl.org/education/curriculum_connections/cyberbullying/glossary.pdf

[2] Maher, D. (2008). Cyberbullying: an ethnographic case study of one australian upper primary school class. *Youth Studies Australia*, *27*(4), 50-57.

[3] Patchin, J., & Hinduja, S. "Bullies move beyond the schoolyard; a preliminary look at cyberbullying." Youth violence and juvenile justice. 4:2 (2006). 148-169.

[4] Willard, Nancy E. Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Champaign, IL: Research, 2007. Print.

Packet Number	**Number of Windows in Packet**	**Correctly Identified Bullying**	**False Negatives**	**Correctly Identified Innocent**	**False Positives**	**Pct of windows containing Bullying**	**Percent Correct**
1	131	19	3	91	18	16.793	83.969
2	148	32	1	29	86	22.297	41.216
3	226	77	8	55	86	37.610	58.407
4	207	0	0	136	71	0.000	65.700
5	196	10	1	130	55	5.612	71.429
6	199	39	0	42	118	19.597	40.704
7	212	57	2	81	72	27.830	65.094
8	169	0	8	58	103	4.733	32.320
9	210	31	14	105	60	21.428	64.762
10	178	42	12	52	72	30.337	52.809
11	186	47	12	76	51	31.720	66.129

Table 1: BullyTracer Results