Detecting the Presence of Cyberbullying Using Computer Software

Jennifer Bayzick

Department of Mathematics and Computer Science

Mentor: Dr. April Kontostathis

Spring 2011

Submitted to the faculty of Ursinus College in fulfillment of the requirements for

Distinguished Honors in Computer Science

Distinguished Honors Signature Page
Jennifer Bayzick
Detecting the Presence of Cyberbullying Using Computer Software

Advisor(s): ____
April Kontostathis

Committee members:

____
April Kontostathis

____
Akshaye Dhawan

____
Lynne Edwards

Outside Evaluator:

____
John P. Dougherty

Approved: ____
Roger Coleman

Abstract

Cyberbullying is willful and repeated harm inflicted through the medium of electronic text. The goal of this project is to design computer software to detect the presence of cyberbullying in online chat conversations. Though there are some commercial products that profess to detect cyberbullying [6], none are research-based, and this project provides the foundations of research that could be used in home monitoring software.

What constitutes cyberbullying? What categorizes a chat post that contains cyberbullying? What are the defining features of each category? The types of cyberbullying dealt with in this project included the following: flooding, masquerade, flaming, trolling, harassment, cyberstalking, denigration, outing, and exclusion. Rules based on a dictionary of key words are used to classify a window of posts.

Once the program implementing these rules was built, its decisions had to be run against a truth set determined by human hand coders. We developed a truth set by giving each window to three human coders. The final truth set was determined by a voting system, where, if at least two of the three coders labeled a window as containing cyberbullying, it was labeled as cyberbullying in the truth set.

The program 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. Overall, it decides correctly 58.63% of the windows. This suggests that our coding rules must be refined to not falsely flag so much innocent conversation.

Table of Contents

I Introduction

Whatever happened to the standard, "Give me your lunch money, kid, or I'll pummel you?" Or for girls, how everyone in the seventh grade tolerated Sally at lunchtime, but as soon as she left the table, the topic of conversation moved to how her skirt did not match her blouse, how she should never do her hair like that again, and how she clearly has no friends in the world. These are classic examples of traditional bullying, happening in person between people who know each other. Though situations like these still take place in almost every elementary and secondary school in the world, the availability of the Internet has allowed today's youth to extend the harassment of peers to the virtual world as well. Cyberbullying between classmates, as well as between Internet users who are strangers offline, is increasing at an alarming rate.

In light of the increasing issue of cyberbullying today, we developed a program, BullyTracer, to detect the presence of cyberbullying in online conversation. It is a rule-based system and was compared against human-coded truth sets. 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 yet, this research provides background knowledge about the components of cyberbullying conversation and offers theories on what would improve the accuracy of the program.

II Cyberbullying Defined

Patchin and Hinduja define cyberbullying as “willful and repeated harm inflicted through the medium of electronic text [3].”

For a conversation to be considered bullying, the intent of harm (physical, emotional, social, etc.) must be present. This includes:

- Threats of physical harm or violence, most often used by males
    - “I’m gonna beat the crap out of you after school tomorrow”
    - “Ill kick ur face in punk”
- Any conversation meant to hurt the victim’s feelings. This is more commonly employed by girls, who focus on appearance and promiscuity; boys use insults too, although they focus more on the victim’s masculinity, including sexuality, strength, and athleticism.
    - “ur so fat and ugly no wonder you cant get a date”
    - “ur such a fuckin homo go fuck ur brother”
    - “omg your so pretty can i be you....not.”[1]
- Any comments meant to damage the reputation or social standing of the victim in an online or offline community (a chat group, a sports team, a class in school, etc.) This includes comments not made to the victim directly, but to someone the victim likes and respects, or comments put in public domain where the victim and others will have access
    - “Did she actually fuck him last night? Shes such a fucking slut, she’s up to 8 in that frat”

[1] This example was taken from the personal page of user Alias1 on formspring.me.

  - “Dude totally fucked the gay guy last night, flaming homo”

Willful harm excludes:

- Sarcasm between friends, in which one chat line taken out of context looks worse than it is. These examples are taken from chat conversations in which the users are known to be close friends.
  - “yea yea whatever. Hobag”
  - **“**I got your e-mail, slut, I can't believe someone so horrible could be so thoughtful”
  - **PersonA:** “i like my away message more than i like you”

    PersonA: “just thought i'd tell you”

    PersonB: “well I'm glad, so then whenever I see you I make you miserable, which makes my day”

    PersonA: “ hahahah”

    PersonA: “you're such a whore”
- One or more comments meant to criticize or disagree with an opinion but not meant to attack the individual. A comment like this can vary in degrees of “severity” or vulgarity, but much of this is dependent upon the reaction of the victim. This is difficult to quantify because in many cases, it is not clear whether the bully means to cause harm or if the victim is taking something too personally. For instance, a conversation like:

  PersonA: “I'm so psyched I got tickets to the Phillies game tomorrow! ☺”

  (a few lines later, after conversation from others)

  PersonB: “The Phillies suck, I can't believe anyone would want to see them play! lol”

> Though this may seem insulting, this would not be considered cyberbullying because it does not seem to directly and personally attack PersonA; it merely expresses a different opinion.

What distinguishes a student who makes one nasty comment from a full-fledged bully, either in-person or online? A bully meaning to cause harm usually does not stop with a single malicious remark to the victim and may target more than one victim. If a bully steals your lunch money once, it is likely that he will steal it again, and he may even move on to your nerdy friends. Similarly, a key element of cyberbullying is the fact that insults happen repeatedly over time.

By definition, cyberbullying must take place using an electronic medium. Patchin and Hinduja specify that the communication be through electronic text (as opposed to talking on the phone or sending pictures, videos, or other types of electronic media). Within the textual data, other mediums sometimes become relevant (pictures, video, etc.) and can be used to determine context, but the vast majority of the data is in text format.

One could also consider the definition of cyberbullying from the viewpoint of the victim, as opposed to the bully. Essentially, if one feels bullied, one is being bullied, regardless of the intentions of the bully. However, sentiment of intent is difficult to determine online, so we focus on specific language and terminology.

III Types of Cyberbullying

The definitions above provide a theoretical description of cyberbullying. The goal of this project is to detect the presence of cyberbullying in chat transcripts, so more specific descriptions are needed for the development of software. Nine different methods of cyberbullying were found [1][2][4].

i. Flooding

Flooding consists of the bully repeatedly entering the same comment, nonsense comments, or holding down the enter key for the purpose of not allowing the victim to contribute to the conversation. For example:

Luke:
Luke:
Luke:
Luke:
Luke:
Luke:
Luke:
Luke:
Luke:
Luke:
Jack: Don't

Luke uses flooding against Jack and everyone else in the chat room as a form of intimidation. Jack's response includes a protesting comment, "don't." "Please don't do that," "stop," "this really sucks," etc. Then, the presence of multiple lines of blank or repeated comments by the bully, followed by a protesting comment by another user, the victim, would constitute flooding [2].

ii. Masquerade

Masquerade involves the bully logging in to a website, chat room, or program using another user's screenname. A masquerading bully can have a number of different motivations.

- BullyA logs in under VictimA’s screenname to post inappropriate or embarrassing comments about VictimA and make it seem as though VictimA is talking about himself or herself.
- BullyA logs in under VictimA’s screenname to harass VictimB and make it seem as though VictimA is the bully.
- BullyA creates a new screenname for an imaginary UserC, then harasses VictimA so that VictimA does not know who the comments are coming from.

This technique is inherently difficult to detect, because online, if you sign into a forum or chat room with a person’s screenname, you essentially are that person. There would have to be a conversation separate from the bullying medium (the particular chat room or website) between the victim and an outside observer in which the victim states that someone else had logged in under their name and posted the comments.

In the case where the bully creates a new user account to bully from, it is even more difficult to detect masquerade. Having to enter minimal information to create an account and allowing users’ information to be private makes it easy to create an anonymous account so the victim cannot tell from whom the harassment is coming. An alternative for the bully would be to lie about key information so the bully’s actual information is not revealed. Some websites (such as formspring.me) even allow the posting of comments entirely anonymously, making it very easy to harass someone.

When looking to detect masquerade, there is a bright side. Although the bully is harassing the victim from another screenname, the bully uses the same types of methods and language any other bully would use. Then, to detect masquerade, the first step is to find bullying language in the way one would detect any other type of bullying. Once it is determined that cyberbullying has occurred, it can be categorized as masquerade in two ways. If it is found that the bully logged in under a stolen screenname, then clearly masquerade has occurred. The other option would be to examine chat lines posted by the stolen screenname after the bullying lines to see if there is some sort of confrontation or accusation of stolen identity from the stolen screenname. The latter is not foolproof, but can provide evidence for the possible presence of masquerade [4].

iii. Flaming (bashing)

Flaming, or bashing, involves two or more users attacking each other on a personal level. In this form of cyberbullying, there is no clear bully and victim, and there is little to no power differential between the participants. The conversation consists of a heated, short lived argument, and there is bullying language in all of the users' posts [4]. For example,

PersonA: "I like hats!"
PersonB: "PersonA, this is the dumbest blog post on earth. You're the dumbest person on earth. You have no friends. Quit the Internet."
PersonA: "Really, well maybe I would, but that would leave only your blog for people to read, and since you have no friends to read it, the Internet would collapse. Fuck your mother."
PersonB: "chinga tu madre, coño!" [Translation: "Fuck your mother, mother fucker!"]

iv. Trolling (baiting)

Trolling, also known as baiting, involves intentionally posting comments that disagree with other posts in the thread for the purpose of provoking a fight, even if the comments don't necessarily reflect the poster's actual opinion. The poster intends to arouse emotions and incite an argument, although the comments themselves are not necessarily personal, vulgar, or emotional. For instance, a teenager finds a Jewish message board and posts, "The Holocaust never happened," or joins a pro-life thread and says, “well clearly abortion is ok because the human life doesn’t begin until a baby is born, it’s just a bunch of cells [1].”

v. Harassment

Harassment is the type of conversation that comes to mind when the term “cyberbullying” is mentioned, and it most closely mirrors traditional bullying with the stereotypical bully-victim relationship. With harassment, the bully and victim roles are clear, and there is a definite power differential (physical, social, etc.) between the two participants. This type of cyberbullying involves repeatedly sending offensive messages to the victim over an extended period of time. A harassment conversation is generally not much of an argument because the victim plays a fairly passive role; sometimes the victim may send vulgar or offensive messages back to the bully, but this is for the purpose of stopping the harassment, and the victim rarely, if ever, initiates conversation [4].

vi. Cyberstalking and cyberthreats

Cyberstalking and cyberthreats involve sending messages that include threats of harm, are intimidating or very offensive, or involve extortion. Threats of harm, murder, or suicide, as well as “distressing material,” which may indicate that the person is upset and considering hurting themselves or someone else, all fall under cyberstalking. The line where harassment becomes cyberstalking is blurred, but one indicator might be when the victim begins to fear for his or her safety or well-being, then the act should be considered cyberstalking [4].

vii. Denigration (putdowns)

Denigration involves “dissing” or 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. The purpose is to harm the victim in the eyes of others; not only does the victim hear the insults, but so does the rest of the community. There are entire websites that are devoted to denigration, particularly in the college setting, such as juicycampus.com (shut down as of February 5, 2009) and collegeacb.com (still active). These sites are set up so users can post conversation topics in new threads and other users can comment, all anonymously, and they have become excellent sources of examples for this project, though they have not yet been used as primary data [4].

viii. Outing

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. This information can include stories heard from the victim, or personal information such as passwords, phone numbers, or addresses. Emotionally, this is usually more harmful to the victim because the victim confided information to the bully and the bully betrayed that trust. That the personal relationship between the bully and victim gives credibility to the information the bully disseminates. Sometimes trickery can be involved; the bully pretends they are alone when having a “private” online conversation, when others are reading the conversation as well, or saying the conversation is in confidence, then forwarding or posting all or parts of the conversation. In terms of the language used, outing is generally similar to denigration, although usually less vulgar and volatile [4].

ix. Exclusion

In the survey conducted by Patchin and Hinduja, 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]. However, detecting exclusion without contact from the victim outside the chat room is not straight forward. If the victim reacts in a way that indicates they think they are being ignored (i.e. “hey guys, are you listening???”), then it can be concluded that exclusion has happened. However, if the victim brings up a topic or asks a question, but the subject is never responded to or brought up again, this constitutes exclusion but may not be as easily distinguished [4].

IV Dataset

Our dataset consists of chat transcripts crawled from MySpace.com. The conversations are thread-style forums, 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 from the Media and Communications Studies Department developed a truth set for testing our algorithms. These individuals were given directions for coding found in the codebook in Appendix A along with the definitions found in sections III and IV. They reviewed each window and indicated whether or not cyberbullying is present. Three assistants coded each transcript, and a voting method was used, where a window was considered to contain cyberbullying in the truth set if at least two humans flagged it as such. These labelers also identified the type of cyberbullying and the lines that are involved in the instance of cyberbullying, but this information is not currently being used. Interestingly, though the coders generally agreed on the presence or lack of cyberbullying in a window, they varied widely in their classification of an instance of cyberbullying. This can suggest either that our category definitions need clarification or that the language is inherently ambiguous. In later stages of the project, the algorithms will be fine-tuned in order to detect the distinction between different categories. It should be noted that we created our own truth set because there was no set of transcripts coded for cyberbullying previously in existence.

V ChatCoder

The general approach we used for BullyTracer was based upon another program, ChatCoder, which was build to classify predatory language from chat conversations into categories [5]. ChatCoder labels chat lines (single posts) in an instant message conversation into one of three classes: personal information (the predator asks questions or volunteers information), grooming (any inappropriate or sexual language), or approach (talking about meeting the victim in person or talking on the phone).

ChatCoder contains a dictionary of relevant words separated into subcategories based on parts of speech and the three coding classes. For instance, "I" is a first person pronoun, "come" and "see" are approach verbs, and "you" is a second person pronoun. ChatCoder highlights and saves all instances of words in the dictionary found in the body of a line of chat. Each line "remembers" which subcategories it contains words from.

Once all words in the dictionary are highlighted, ChatCoder examines each line of chat based on the subcategories it contains, classifies it if appropriate, and highlights the username of the lines classified as containing predatory language. For instance, the line by user goodboyb88 : "I will come see you over and over" was classified as containing approach language because it contained a first person pronoun, a second person pronoun, and an approach verb. This is only one of twelve rules ChatCoder uses to classify posts.

VI BullyTracer

BullyTracer is a program meant to detect the presence of different types of cyberbullying in a chat room conversation. We describe its first version in this section. Due to the conversational nature of cyberbullying, threads were split into moving windows of 10 posts each. Labeling is done at the window level currently, and later stages of the project will distinguish between types of cyberbullying, and the lines involved in cyberbullying. BullyTracer analyzes all files in given directory using a rule-based algorithm. The program evaluates each window on a number of different levels.

A. Post level

The dictionary of code words used by BullyTracer appears in Appendix B. It includes terms in 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. For instance, the post "you are so dumb you make me want to kill someone" should be distinguishable from the post "that test was so dumb it was killing me." Even though the two posts have many words in common, the former should be coded as cyberbullying and the latter should not.

As seen in section IV, some of the categories of cyberbullying include insults directed at a third party, such as, "she's so fat." However, the MySpace data overwhelmingly focuses on direct bully-to-victim cyberbullying. Any instances of an insult to a third party were directed at celebrities in pop culture, politics, or some other realm of public interest.

Within the dictionary, some formatting issues needed to be addressed. For instance, spacing is particularly important. The word “ass” as a swear word is important to the dictionary, but it should not be considered a swear word when used in the word “pass.” Therefore, the entry in the dictionary is “ ass ” and words such as “asshole” and “dumbass” needed to be included separately. Similarly, in online conversation, users frequently use Internet shorthand when typing. So in the line “u want 2 go 2 the mall?”, “U” is a second person pronoun. To cover these cases without identifying each instance of the letter u in the body of the post, the dictionary includes entries for “ u “ and “ ur ” (you’re).

Additionally, some words fall into multiple categories. For instance, words like “bitch” and “asshole” are considered both insults and swear words, while “idiot” is considered an insult and “fuck” is only a swear word.

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.

Currently, a single post is considered to contain cyberbullying if it contains an insult word and a second person pronoun.

B. User level

Each user is often associated with multiple posts in a particular window, and knowing which posts belong to which user can help BullyTracer gain information about a user’s conversational style and the context of the window’s conversation.

For example, the severity of a swear words can be examined by looking at the number of swear words a particular user uses in general conversation. This will be more helpful when enough data has been collected that a "user history" can be stored and the vulgarity of a user can be examined over time. Currently, the number of swear words a user types in a post divided by the number of swear words in the entire window determines the vulgarity of a post. The connection between the vulgarity of a post and the presence of cyberbullying has not yet been established.

A user is considered to be a cyberbully if one or more lines of chat posted by the user in a thread are considered cyberbullying.

C. Window level

A window is labeled as containing cyberbullying if it contains any posts that are labeled as containing cyberbullying. As described in section III, a single line of nasty conversation may not be considered cyberbullying. The window level was conceptualized to be primarily concerned with distinguishing between types of cyberbullying. BullyTracer was designed to be expandable in later stages of the project.

It should be noted, however, that while development has begun on an implementation of ChatCoder as a research-based Internet monitoring tool, BullyTracer is far more preliminary. 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.

VII BullyTracer Design Details

This section will examine the specific design details of BullyTracer. The program is written in the Java programming language. A class diagram is seen in Figure 1.

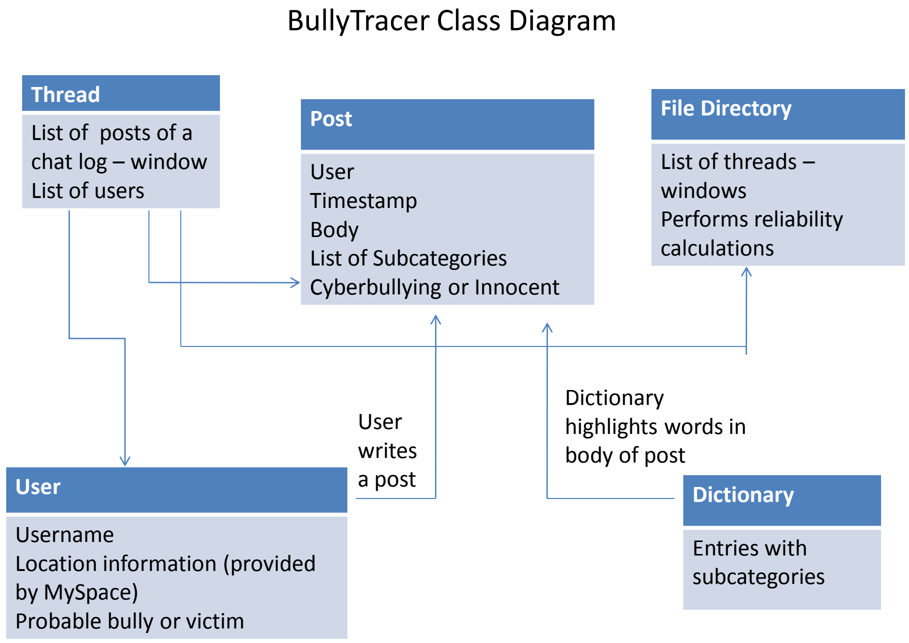
BullyTracer Class Diagram
Thread
List of posts of a chat log – window
List of users
Post
User
Timestamp
Body
List of Subcategories
Cyberbullying or Innocent
File Directory
List of threads – windows
Performs reliability calculations
User writes a post
Dictionary highlights words in body of post
User
Username
Location information (provided by MySpace)
Probable bully or victim
Dictionary
Entries with subcategories

Figure 1: a class diagram for BullyTracer

When the program begins running, it loops through all the files in a pre-defined directory, and processes all files that are in XML format. It is assumed that each xml file contains a single window because a companion program was built to split an entire thread into windows of 10 lines. Within the program, a ***FileDirectory*** contains of all the windows in the given root directory, a ***thread*** is be used to describe the entire conversation in the file containing a single window, a ***user*** contains all information about a user that posts in a given thread, and a ***post*** constitutes a single line of chat posted by a user. A post has an associated post ID, user ID, body, and timestamp, as well as a number of different attributes based on the content of the body. In the program, a ***user*** keeps track of its own user ID, username, location, which posts within the chat log have been posted by that user, as well as a number of attributes based on the bodies of the posts by the user.

VIII Evaluation and Results

Evaluation of BullyTracer's coding rules was done based on how closely the program's decisions matched the human-defined truth set for each window. As shown in Table 1, the program counts the number of correctly identified windows that contain cyberbullying and innocent conversation, 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). The windows in packets 1 and 2 were examined closely to develop the rules in BullyTracer, but this does not seem to have a significant effect on the results for those packets. Also included in the table is the percentage of windows in the packet that contain cyberbullying.

Table 1: BullyTracer Results

Packet Number	Number of Windows in Packet	Correctly Identified Bullying	False Negatives	Correctly Identified Innocent	False Positives	Percent 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

As seen in table 1, BullyTracer codes fairly consistently, even considering that some of the packets were used to create the coding rules and others were not. Overall, the BullyTracer coding decisions match the human truth set 58.63% of the time. Percentages of correct coding vary by packet between 32.32% and 83.969%. We see that, of the 415 windows that actually contain cyberbullying, BullyTracer labeled 354 of them correctly, which is correct 85.30% of the time. 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, as there is a large percentage of false positives (innocent windows identified as containing cyberbullying) and a smaller number of false negatives (windows that contain cyberbullying but are labeled as innocent). Of the 1647 innocent windows, BullyTracer codes 855 of them correctly, which is correct 51.91% of the time. This suggests that our coding rules are too broad and need to be refined. One cause of this is the nature of our data. The project studies general online conversation, and this can include multiple user threads, chat rooms, or one-on-one instant message conversations. Our current test data comes exclusively from MySpace in thread-style forums, where many users can post on a given topic. In contrast to instant message style chats, where posts are usually extremely short, users in the MySpace threads often post multiple sentences or multiple thoughts in a single post. This makes it difficult to distinguish between sentences or thoughts within a post, so sometimes associations are made incorrectly. For example, a post that says, “Did you finish reading that book for class yet? Man it’s so freaking stupid!” would be flagged as containing cyberbullying because the second person pronoun “you” and insult word “stupid” are found in the same post.

However, this is incorrectly flagged because “stupid” does not refer to “you.” Many examples like this have been found in our dataset.

Another issue is strangely worded insults or sentences that do not fit a particular pattern that can be coded for. Human language styles vary inherently between people, and the Internet setting further loosens grammatical, spelling, and vocabulary standards. For example, consider this conversation between two users:

> Black Sheep SuzieQ: Quite an ironic quote coming from someone I gather supports the Bush agenda.
>
> Agent Working MySpace: You gathered wrong. By the look of your blogs it seems this isn't the first time.

This is considered to contain cyberbullying in our truth set because the second user is attempting to insult the first user, but the way the insult is set up does not lend itself to being classified. Research on the linguistic components of the conversation is necessary to detect this instance of cyberbullying.

Sarcasm and good natured teasing are also troublesome. The exact same body of a post can be cyberbullying in one instance and a good natured ribbing in another. More features pertaining to the contextual components of a good natured conversation needs to be done in order to be able to identify a sarcastic post as innocent.

IX Conclusion and Further Research

This project defines nine types of cyberbullying and proposes methods to detect the presence of these types in online chat conversations. It also describes the first implementation of algorithms to detect the presence of cyberbullying on a window level in the program called BullyTracer. Further research includes refining these methods and expanding them to distinguish between different types of cyberbullying. Throughout this project, the only data available for examination was a set of thread-style conversations from www.myspace.com.

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. Overall, it decides correctly 58.63% of the windows. This suggests that our coding rules must be refined to not falsely flag so much innocent conversation.

Future work on the project includes examining instances where BullyTracer codes incorrectly more closely, particularly the false positives, and refining BullyTracer's rules accordingly. The importance of capital letters and swear words should also be examined more deeply. More ambitious goals include examining the role of context in distinguishing between types of cyberbullying and developing rules to classify windows into one of the nine cyber-bullying categories we define in this work.

One of the most important contributions of this work was the creation of the MySpace cyberbullying truth set. There was previously no truth set in existence for cyberbullying, and so ours has been requested by a number of other research teams.

X Acknowledgements

The author would like to thank Dr. April Kontostathis and Dr. Lynne Edwards for all of their help and support, as well as India McGhee, Bryan Stafford, Brett Thom, Rosie Kress, Liz Kilmer, Magdalena Parks, Brett Neslen, Sarah Anne Rosner, Katie Haldeman and the Ursinus College Summer Fellows program**.** 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.

Citations

[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.

[5] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride and E. Jakubowski. Learning to Identify Internet Sexual Predation. In International Journal of Electronic Commerce. To appear.

[6] http://www.netnanny.com/

Appendix A – Cyberbullying Codebook

This is the information our undergraduate assistants were given to learn to code chat transcripts for cyberbullying. They had the opportunity to ask questions at any time of myself or Professors Kontostathis and Edwards for clarification. The reference to my Summer Fellows paper encompasses the information in sections III and IV of this paper.

Directions for Hand Coders:

Thank you so much for helping us out!!

So the plan:

My honors project is on the detection of cyberbullying in online chat conversations. My goal is to build a program – BullyTracer – that will detect the presence of cyberbullying based on a set of rules I need to develop. To decide what should and should not constitute cyberbullying (and then train the computer to recognize it), I need "correct answers" – that's where you come in. I need others' opinions on what should and should not constitute cyberbullying to develop the program, and I also need truth sets to determine if the computer is doing what it should be.

To begin, please read through my Summer Fellows paper – the file labeled BayzickBullyTracerSFFinal.docx. The whole thing will be good background information, but the only section that is necessary is section IV – Types of Cyberbullying. With research, I discovered and defined nine categories that cyberbullying conversation can fall into. Become familiar with these categories, but realize there is also much ambiguity between them, as well as between cyberbullying and innocent conversations. The first few sections of my paper give background information on cyberbullying itself, and the last few sections describe my plan for my honors project. If you have any suggestions or comments for my research, please let me know!

So, as for the actual coding,

1. *Locate the Excel spreadsheet entitled Human Data Sheet 1.xlsx. Fill in your name at the top. Note: you should have a new copy of the data sheet to fill in for each new packet of files you are working on (the packets will contain 100 – 200 files to be analyzed). When you save your edited copy of the file, please label it as "Human Data Sheet [your last name] packet [packet number].xlsx"*
2. *In each packet, there are html files that contain windows of 10 chat lines that are excerpts from longer chat conversations. All data was crawled (saved by an automated computer program) from MySpace thread-style conversations, in which*

multiple users can chat about a specified topic, though conversations do not always stay on that topic. Some windows include the entire thread, while others are some section of a much larger conversation. Locate these files.

3. *For a particular file: Open the html file in your web browser or notepad/wordpad program. Read through the conversation excerpt. Enter the file name into the spreadsheet. If there is language (a post or a group of posts) that you think constitute cyberbullying, note that in the spreadsheet in the "Is Cyberbullying Present? (Y/N)" column, and the line numbers (all files should be labeled 1-10) that are involved in the "Chat Lines Involved" column. Make a best guess at the category it falls into.* ***It is more important to identify the presence of cyberbullying than to classify it correctly.*** *However, classification is helpful too. If you would like to make any notes about the language used, the ambiguity of the category you chose, the style of a particular user, or anything else you think of, feel free to do that in the spreadsheet to the right of the designated columns. Any comments would be helpful for my analysis.*

4. *Note: most conversations you will read will probably be considered innocent and might be boring. I apologize. However, if there are any innocent conversations that look like cyberbullying but you would not consider them to be "bad" enough to be classified, notes on those examples would be particularly helpful to determine where the line falls between cyberbullying and trash talking or innocent conversation. Also, there do not have to be notes for each file (window), but if there is something interesting going on, please comment.*

5. *Once you finish documenting a particular html file, congratulations!! Pick a new one, rinse, and repeat. When you have finished all files in a packet, make sure to save your human data sheet, and start a new packet of files! Remember, there should be a separate data sheet for each packet (so your excel file doesn't get too crazy long)*

Things I'm thinking about:

1. *The difference in language patterns in cyberbullying versus innocent conversation versus trash talking, and how to distinguish between them in a systematic way.*

2. *Users' roles in the cyberbullying (bully/victim, multiple bullies harassing one or more victims, multiple users attacking each other, but with no clear victim) and how this plays a part in the classification of the window.*

3. *Ways in which a computer could begin to classify conversations (think hard and fast rules, a computer doesn't do ambiguity)*

4. *Where can we go from here?*

5. *Lots of other random things that I'm blanking on right now – if you have any suggestions, please let me know!!*

Thank you again for helping us all out!! You are a particular help to me and my honors project, so thank you! If you have any questions/comments/concerns/suggestions, feel free to let me know at jebayzick@ursinus.edu.

Thanks again!!
~Jen Bayzick

Senior, mathematics and computer science

Appendix B – Codeword Dictionary

These are all the words that are flagged as contributing to cyberbullying conversation in a chat post. Take note that some entries fall under multiple categories based on their meaning and vulgarity. Also, entries that include quotation marks (“ “) were not used in the dictionary; the quotation marks are there to highlight special spacing issues before or after the word.

Dictionary Entry	Coding Category
bullshit	swear
“shit “	swear
fuck	swear
fucker	swear
f*ck	swear
fk**	swear
f***	swear
“ ass “	swear
damn	swear
dumbutt	swear
dumbbutt	swear
dumbass	swear
jackass	swear
asshole	swear
bitch	swear
retarded	insult
stupid	insult
dumb	insult
dipshit	insult
“ tard”	insult
idiot	insult
loser	insult
whore	insult
“ ho “	insult
“ hoe “	insult
douche	insult

douchebag	insult
“ cock “	insult
slut	insult
cunt	insult
ugly	insult
bitch	insult
skank	insult
hoebag	insult
gay	insult
“ fag ”	insult
faggot	insult
“ oaf”	insult
ignorant	insult
loser	insult
pussy	insult
pathetic	insult
shithead	insult
fucker	insult
“ ass “	insult
dumbutt	insult
dumbbutt	insult
dumbass	insult
jackass	insult
asshole	insult
bitch	insult
“you “	second person pronoun
your	second person pronoun
“ ur “	second person pronoun
“ u “	second person pronoun