7-21-2016

Detection of Cyberbullying in SMS Messaging

Bryan W. Bradley
Ursinus College, brbradley@ursinus.edu

Detection of Cyberbullying in SMS Messaging

Bryan Bradley
Dr. April Kontostathis
July 22nd, 2016

Abstract

Cyberbullying is a type of bullying that uses technology such as cell phones to harass or malign another person. To detect acts of cyberbullying, we are developing an algorithm that will detect cyberbullying in SMS (text) messages. Over 80,000 text messages have been collected by software installed on cell phones carried by participants in our study. This paper describes the development of the algorithm to detect cyberbullying messages, using the cell phone data collected previously. The algorithm works by first separating the messages into conversations in an automated way. The algorithm then analyzes the conversations and scores the severity and frequency of the bullying words. A scoring threshold is used to predict whether or not a message or a conversation contains cyber bullying. Over four different data sets, the algorithm precisely found, 98.01% of all conversations. We achieved results of a precision of 62.50% and recall of 55.56%.

1. Introduction

Social media and the use of cell phones has grown rapidly in the past decade, especially with adolescents. With this rise in usage of social media and technology, there is also a rise in cyberbullying. Cyberbullying is defined as "the act of harassing someone online by sending or posting mean messages, usually anonymously" ("cyberbullying", 2016). In short, this means that cyberbullying is using technology to post or send mean or hurtful messages. Previous research has shown that 19% of teens reported that they have been cyberbullied (Chen, Zhu, Zhou, & Xu, 2012).and new types of social media platforms are being developed all of the time. To combat these rapidly emerging issues of cyberbullying, a program was developed to detect and flag this cyberbullying so we can protect children, teens, and even adults.

We developed an algorithm with two primary functions: to sort Short Message Service (SMS) messages into conversations and to use a dictionary of manually compiled bully words to determine which messages and conversations are cyberbullying. This program can sort through and score words, messages, and conversations for cyberbullying content. As the program sorts through the messages, it is able to flag the components of the conversation which could be cyberbullying. The goal of the program is to detect cyberbullying at both the single instances level and the conversation level.

2. Background Information and Related Work

Despite the high prevalence of cyberbullying, few other research teams have worked on the detection of cyberbullying. One of the more notable research projects was conducted by Reynolds et.al (Reynolds, Kontostathis, & Edwards, 2011). They analyzed data from Formspring.me and utilized a machine learning program call Weka (The WEKA Data Mining Software: An Update, 2009). Weka was used to train the program to recognize cyberbullying. In the end of the paper, they discuss how the machine learner was able to correctly detect 78.5% of cyberbullying posts. This is significant progress in the field of cyberbullying detection and can be used as an example of importance of detecting cyber bullying by showing that with certain attributes, it can be done.

Earlier work by Bayzick, et. al used MySpace data to try and detect cyber bullying. The project was called BullyTracer and used human-labeled data as a comparison to the algorithm that was created. Along with this, a MySpace truth set was created for research purposes, and was the first one to be created in this domain. The results of the research was that cyberbullying was

correctly flagged 85.30% of the time. However, it incorrectly labeled 51.91%, making the total accuracy 58.63% (Bayzick, Kontostathis, Edwards, 2011).

3. Methods

There was two main methods to this research, finding the conversations, and detection the cyberbullying.

3.1 Data

For this research we used real SMS messages from different participants. In previous research done by our lab, we distributed eleven cell phones to youth who agreed to participate in this study. The participants were ages ten to fourteen and were given cellphones under the assumption that we would collect their SMS messages. Within the last year, we have collected over 80,000 messages that were used for this research.

3.2 Conversation Detection

To sort messages into conversations, we first take three pieces of data, sender, receiver, and time stamp. Next, we take the sender and receiver and match it with the message before, or a conversation before, to see if they match, if they do not a new conversation is created. Then the message is checked with the message before to see if it is within a certain time threshold (it was determined at 140 minutes). If it is that time threshold it is added to the conversation, if not a new conversation is created.

Figure 1. Interleaved SMS Messages from Cell Phone Data

408		Yea	10/18/2015 14:47	1
409		It\'s ok	10/18/2015 14:47	1
410		Please let me know when you get back in the house. Thanks	10/18/2015 14:51	18
411		Cya bye	10/17/2015 17:54	9
412		Hi Ben. I am glad you are home!	10/18/2015 15:00	19
413		Hey it worked:-)	10/18/2015 15:01	19
414		Ok :)	10/18/2015 13:27	14
415		R u going to youth group	10/18/2015 15:18	20
416		No	10/18/2015 15:19	20
417		Can u send me a pic of the reign of terror page bot parts	10/18/2015 15:28	17
418		Okay	10/18/2015 15:30	1
419		No	10/18/2015 15:19	20
420		No	10/18/2015 15:19	20
421		Are you back in the house ?	10/18/2015 15:39	18
422		Did you get home ok?	10/18/2015 15:40	21
423		Yea	10/18/2015 15:49	1
424		Yes	10/18/2015 15:49	21
425		Yes	10/18/2015 15:49	17
426		In a minute	10/18/2015 15:49	17
427		Yup	10/18/2015 15:50	1
428		So	10/18/2015 15:51	1

Once we believed that the algorithm was working correctly, we started to find the statistics of the conversations. Among these statistics were how many true positives, false positives, and false negatives there were in the conversations. We did this by creating a different set of conversations and manually labeled them. Then we compared the actual conversations to the predicted conversations and gathered the data. From the data, we could calculate four statistics, true positive, false positive, true negative, and false negative. True positive means that the algorithm detected it and it was cyberbullying. False positive is when the algorithm detects it but, it is not cyberbullying. False negative means that it was not found by the algorithm but was cyberbullying and true negative says that the algorithm did not find it and it was not cyberbullying. From this data we calculated two more important statistics, precision and recall.

To find *precision*, we used the formula $\frac{TruePostive}{TruePositive+FalsePositive}$. To find *recall* we used the formula $\frac{TruePositive}{TruePositive+FalseNegative}$.

3.3 Cyberbullying Detection

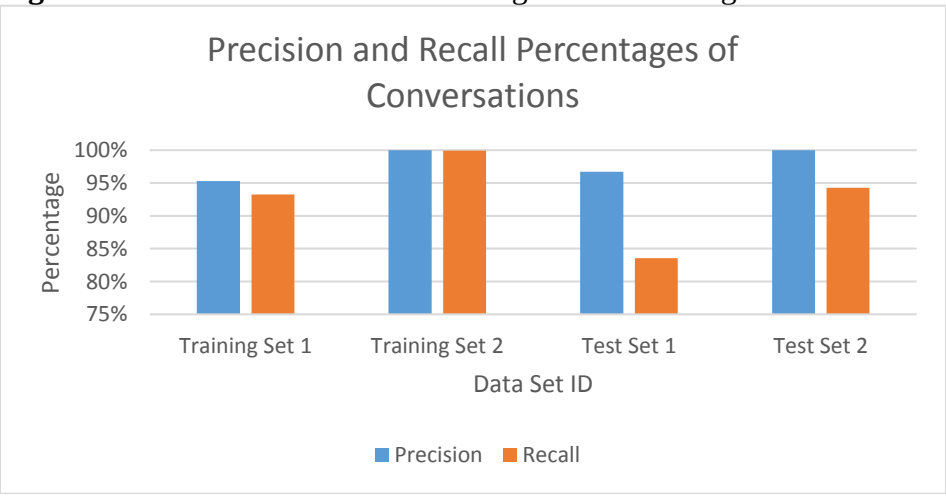
When the algorithm starts, it takes each message and breaks it down to each individual word. Each word is compared to each word in the bully dictionary. If it matches one of the words, it takes the score of the bully word and adds it to the total messages score. If the message score passes a determined threshold then the messages is flagged as cyberbullying. Once all the messages are scored, then the algorithm adds all of the message scores in a conversation and if the total conversation score reaches a threshold, then the entire conversation is declared as cyberbullying.

After running the algorithm with different thresholds, the highest statistics I could get were a precision of 64.58% and a recall of 87.50%. Although these statistics are promising, they are not as high as we wanted. Because of this, we decided to create our own bully dictionary. After manually looking through what we labeled as cyberbullying, we created our own dictionary of bullying words and got much better results.

4. Results

After finding statistics on the first dataset, we manually labeled three more sets of data and ran them through the algorithm. We found that the recall was a little low so we experimented with how long between messages should count as a conversations, and we adjusted the time to get the highest recall percentage. When running the statistics of conversations on all four datasets, it produced the graph in Figure 2. We found that the algorithm can continuously reach a precision of at least 90% and recall of at least 83%. In most conversations, the time between text messages fell in under 140 minutes. Anything that isn't detected correctly, either the messages were more than 140 minutes apart or could have been mislabeled.

Figure 2. Precision and Recall Percentages Based on Algorithm



When we ran the algorithm with the bully dictionary made in previous research we got very high precision but low recall. In the training set we got a precision of 91.67% and a recall of 56.46%. In the test set, we got a precision of 100.00% and a recall of 30.56%. Although this

sounds good it is not. Since the recall is so low that means it is not picking up enough of the bullying messages but, they are correctly identifying the few it does find.

After creating our own dictionary of bullying words we got better statistics. On the training set, we optimized the threshold on the training set and got a precision of 95.83% and a recall of 44.22%. Then on the test set we got precision of 62.50% and recall of 55.56%.

With both of these experiments we have the same problem - low recall. It is good that we have a high precision, but since we have a low recall, which means it is not finding a lot of the cyberbullying

5. Conclusion

We found that this algorithm can be effective in detecting cyberbullying in SMS messages. Even though the statistics are low, we know that when we get the recall up, we will to be able to detect more cyberbullying messages.

5.1 Future Work

The next steps in the research involve raising the precision. This can be done in multiple different ways. These can include optimizing the dictionary, looking for different speech patterns, and checking the messages before and after the bullying to see the reaction of the person being bullied.

I will also looking into bullying in the entire conversation. Especially if we can find that if someone is starting to bully and we know they will continue in the conversation, we can stop them before it gets too bad.

6. Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grant Nos. 0916152 and 1421896. 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.

I would also like to thank my advisor Dr. April Kontostathis and my mentee Serena Schaefer.

7. References

Bayzick, J., Kontostathis, A., & Edwards, L. (2011). Detecting the presence of cyberbullying using computer software.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012, September). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)* (pp. 71-80). IEEE.

cyberbullying. (n.d.). *Dictionary.com Unabridged.* Retrieved July 21, 2016 from Dictioary.com website http://www.dictionary.com/browse/cyberbullying

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on* (Vol. 2, pp. 241-244). IEEE.