Facebully: Towards the Identification of Cyberbullying in Facebook

Lisa M. Tsosie
Arizona State University
4701 W. Thunderbird Road
Glendale, AZ 85306, USA
lmtsosi1@asu.edu

Yasin N. Silva
Arizona State University
4701 W. Thunderbird Road
Glendale, AZ 85306, USA
ysilva@asu.edu

1. INTRODUCTION

The Problem. In the past year, one million children were victims of cyberbullying on Facebook. In fact, about 20-40% of all youths have experience cyberbullying at least once in their lives [1]. The goal of our research work is to study, design and implement a model to identify cyberbullying by exploiting social media data. Once the application is ready for deployment, it can be used for parents or guardians to monitor their children via their social network and give them knowledge of whether their child is a victim of online aggression or not. The techniques of our work could also be used to identify depression or self-destructive tendencies in youth as well.

The Motivation. There has not been sufficient research in identifying cyberbullying behavior in social networks. Rather, the emphasis of the study of aggression has been on traditional bullying, and cyberbullying through venues such as mobile or chat. Particularly, to the best knowledge of the authors, no previous work has proposed an automated model to identify cyberbullying in social networks like Facebook. Facebook is ranked the second most popular site in the United States, with more visits than Google last year [2], providing an appealing environment for the implementation of the type of application we wish to create.

Our Contributions. The key contributions of our work are:

- The design of a model to identify cyberbullying that builds on previous research findings in the areas of traditional and cyberbullying in adolescents. Our model uses social media data, e.g., content of messages, posts, picture comments, etc. to compute a *Bullying Rank* function.
- The design and implementation of an application that used our proposed model to identify cyberbullying. We report in this paper the current state of the implementation of our Facebook app: *Facebully*.

2. BACKGROUND AND RELATED WORK

Most of the work in identifying bullying among adolescents has focused on traditional bullying, or cyberbullying via mobile or chat-based venues [3, 4, 5, 6, 7, 8]. To identify a foundation for the proposed model, we first identified the previous findings that were relevant to the identification of cyberbullying on social networks.

Several studies about cyberbullying have been proposed in the literature, e.g., whether parental perception of adolescents' online behavior is causational with adolescents' vulnerability to cyberbullying [3, 4], probability of victimization [5] and emotional impact [6, 7] for age-gender groups, and measuring the severity of online aggression in correlation to the number of bullies involved [8].

The first task of our project was to identify the warning signs and states of vulnerability studied in the previous work that can be identified using social media data. For instance, the survey-driven work in [6, 7] studied the frequency and emotional impact of cyberbullying in different age-gender groups. From these studies, we not only identified that gender and age are two characteristics to consider in the identification of cyberbullying, but also the probability of bullying for each group. Similarly, the work in [5] concluded that cyberbullying victims are typically adolescents on the "fringe" of various peer groups, e.g., newcomers, ethnical, physically/mentally handicap, etc. Social media data can be mined to identify if a user belongs to these groups.

3. AUTOMATED IDENTIFICATION OF CYBERBULLYING IN FACEBOOK

The main components of Facebully, our application to identify cyberbullying in Facebook is presented in Figure 1. The application is designed to interact with both the user, i.e., minor, and parental figure. Before the application can extract from the user's account, the user must grant various permissions requested by the application in order to obtain all the required data. The cyberbullying identification module is in charge of using this data to determine if the user is a victim of online aggression. To this end, the application computes a *Bullying Rank* (*BR*) expression that is based on warning signs and states of vulnerability identified in the first stage of our project. This rank is used to normalize the intensity of cyberbullying and is computed as follows:

$$S = min(N, w_1 \cdot IWC + w_2 \cdot IMC + w_3 \cdot CEP)$$

$$V = (w_3 \cdot NSF) \cdot (w_4 \cdot NNF) \cdot (w_5 \cdot AGF)$$

$$BR = 10 * S * V$$

where $0 \leq BR \leq 59$.

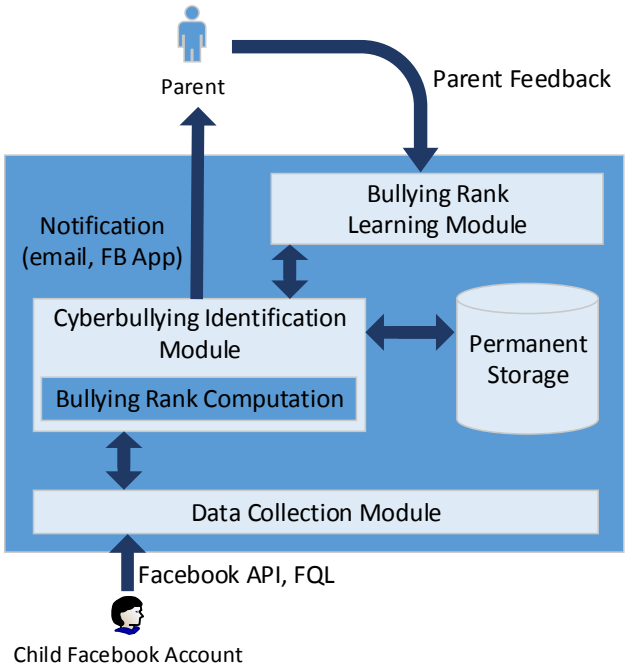

Figure 1. Facebully Architecture.

The Bulling Rank and its main components are graphically depicted in Figure 2.

Properties studied from previous research findings are interpreted as conditions which fall into one of two categories: warning signs (*S*) and states of vulnerability (*V*), where each sub-factor is given an appropriate weight *w* and the product of the two main factors produces the Bullying Rank *BR*. The Bulling Rank can fall into any of three pre-defined levels with the respective intervals: low risk (0-19), moderate risk (20-39), and severe risk (40-59).

Measuring Warning Signs. The current model accounts for the number of insulting content (as defined by a library of insults) on the user's wall or in the user's inbox, as well as the probability of perceived victimization for identified age-gender groups [5]. The warning signs have numeric values, and represent the number of occurrences of: Insulting Wall Content (*IWC*), Insulting Message Content (*IMC*), and Comments on Embarrassing Photos (*CEP*). Our approach does not include a separate component for insulting chat content since Facebook implicitly integrates chat messages inside of a user's inbox. Insulting chat messages are included in *IMC*. We account for the *Group Effect*, as identified in [8], where the number of *CEP*s increases the severity of perceived victimization. Any *S* value above *N* is classified as severe, thus we compute the minimum between the summation of the warning signs' sub-factors and *N*. Our application currently use *N*=100.

Measuring Vulnerability. The states of vulnerability have asymmetric binary values and include: New School Factor (*NSF*), New Neighborhood Factor (*NNF*), and Age/Gender Factor (*AGF*). When the value of a component is true, the numeric value of this component is a number that

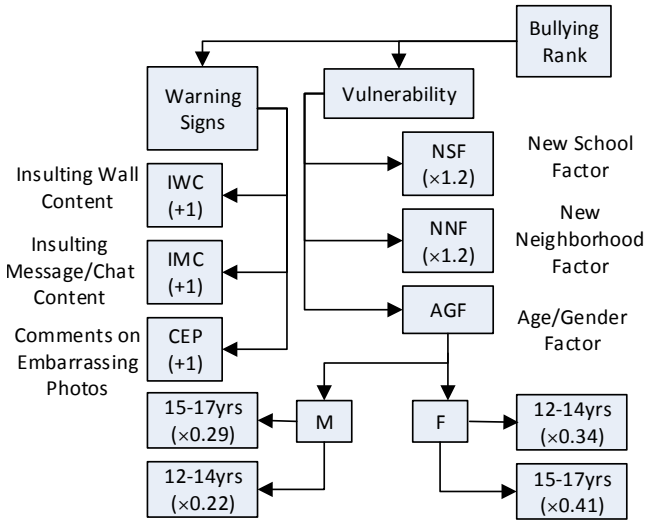

Figure 2. Bullying Rank Factors.

portrays the probability of perceived victimization. These probabilities have been derived from the statistical results of [5]. When the value of a component is false, its numerical value is set to 1.

When the application extracts from the user's profile for the first time, only data during a pre-defined interval (currently the previous 6 months) is acquired for a comprehensive analysis. Some of that data is stored in a permanent storage because certain information must be: (1) monitored to identify changes, e.g., a change of current location (*NNF*) or school (*NSF*), or (2) stored for future computations, e.g., the value of *S* from the last update. From that point forward, incremental updates extract data from the latest user activity since the last update. After every update, the parent is notified of the user's current *BR* via their Facebook account or e-mail.

Learning Module. One of our design goals is the inclusion of mechanisms that enable improving the accuracy of our cyberbullying identification model. To this end, parents who use the application can provide feedback about the accuracy of the predictions and this information can be analyzed by both our domain experts and our team to improve the accuracy of the model, e.g., by adding new factors or modifying the weight values and probabilities. The inclusion of machine learning techniques to improve the accuracy of our classification model is one of our future tasks.

4. FUTURE WORK

The research work presented in this paper is currently in progress. Our future tasks include: (1) studying mechanisms to integrate machine learning models to improve the accuracy of cyberbullying identification which can be seen as a classification problem, and (2) studying mechanisms to include new cyberbullying factors that cannot be directly extracted from Facebook data, e.g., ethnicity, physical and mental disabilities, etc.

5. REFERENCES

[1] Poremba, Sue. 2011. How to tell if your child's being cyberbullied. *MSNBC, Web*. 14 Feb. 2012.

[2] Russell, Matthew. 2011. *Mining the Social Web*. Sebastopol: O'Reilly Media Inc.

[3] Ang, R. P., Chong, W. H., Chye, S., and Huan, V. S. Loneliness and generalized problematic Internet use: Parents' perceived knowledge of adolescents' online activities as a moderator. *Computers in Human Behavior, 28* (4), 1342-1347. DOI = http://dx.doi.org/10.1016/j.chb.2012.02.019.

[4] Law, D. M., Shapka, J. D., and Olson B. F. To control or not to control? Parenting behaviours and adolescent online aggression. *Computers in Human Behavior, 26* (6), 1651-1656. DOI = http://dx.doi.org/10.1016/j.chb.2010.06.013.

[5] Piazza, J., and Bering, J. M. Evolutionary cyber-psychology: Applying an evolutionary framework to Internet behavior. *Computers in Human Behavior, 25* (6), 1258-1269. DOI = http://dx.doi.org/10.1016/j.chb.2009.07.002.

[6] Ortega, R., Elipe, P., Mora-Merchin J. A., Calmaestra, J., and Vega, E. The emotional impact on victims of traditional bullying and cyberbullying: A study of Spanish adolescents. *Journal of Psychology, 217* (4), 197-204. DOI = http://dx.doi.org.ezproxy1.lib.asu.edu/10.1027/0044-3409.217.4.197.

[7] Waasdorp, T. E. and Bradshaw, C. P. Examining student responses to frequent bullying: A latent class approach. *Journal of Psychology, 103* (2), 336-352. DOI = http://dx.doi.org.ezproxy1.lib.asu.edu/10.1037/a0022747.

[8] Dooley, J. J., Pyżalski, J., and Cross, D. Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. *Journal of Psychology, 217* (4), 182-188. DOI = http://dx.doi.org.ezproxy1.lib.asu.edu/10.1027/0044-3409.217.4.182.