ChatterShield – A Multi-Platform Cyberbullying Detection System for Parents

Emergent Research Forum (ERF)

Nargess Tahmasbi
Pennsylvania State University
nvt5061@psu.edu

Alexander Fuchsberger
Bucknell University
af033@bucknell.edu

Abstract

Cyberbullying (CB) is one of the major cyber-issues among adolescents. In several cases, the effect of CB on victims was so severe that victims ultimately committed suicide. Despite the prevalence of automated CB detection studies using computational approaches, CB detection research still lacks empirical studies that act upon the detected CB instances in a cross-platform environment. In this paper, we propose a multi-platform, incremental self-training system that uses a decentralized learning approach to automatically detect cyberbullying instances on a minor’s extended online social network. To improve the self-training model, the crowdsourced feedback of human moderators (guardians) is used. We first point out the major challenges in CB detection research and then explain our proposed design to address the discussed challenges. We conclude with the contribution to practice, and our plans for implementing the solution.

Keywords

Cyberbullying, automated detection, empirical design

Introduction

The prevalence of online social networks today has brought several social and ethical concerns. One of the major issues that targets mostly adolescents is cyberbullying (CB). Cyberbullying by definition shares the same components as traditional bullying except that it is conducted through an internet medium. It is defined as “an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who cannot easily defend him or herself” (Smith et al., 2008). While several detection models have been proposed to automatically detect CB, few have integrated approaches to prevent CB. Inappropriate content filtering is increasingly demanded by politics, platform users, and law. Preventing CB requires the ability to not only detect but also intercept CB content. Unless social media platforms do allow such functionality in the future we do not see direct CB prevention by blocking content currently as a realistic option. It is more likely that social media platforms implement their own measures to filter inappropriate content as seen in online computer games and chat rooms, and recently Facebook. Attempts have been made to provide apps as an informational basis for parents (Silva, Rich, & Hall, 2016) or as a nourishing environment to foster positive behavior (Fan, Song, Shibasaki, & Adachi, 2015). But the current solutions do not let parents see actual content and report it. Empirical approaches need to act in real-time and provide timely preventive solutions to let the human moderator act upon cyberbullying contents before they find their way to the child victim and make their harmful impacts.

In this paper, we address the research gaps mentioned above by proposing a CB detection and prevention application that integrates multiple social media platforms and is tailored to protect the demographic group most affected by CB: children and teenagers. In its basic form, users can grant our application, “ChatterShield”, access to various social media accounts and share potential bullying data with guardian users - typically parents. An individual database is created consisting of users in the primary user’s extended network based on the information available. Users maintaining accounts in multiple platforms are intelligently merged to reduce duplicates and provide more accurate information. The app monitors incoming messages, tags, and feed entries that may be associated with the primary user in a CB context. Once detected, the guardian users are informed of the incident and are provided information on both the
offending user and the message. The proposed system is superior to previous approaches, as it helps parents and guardians to get a more direct, complete and timely CB detection and reaction. The multi-platform detection system features decentralized customized learning modules that use incremental learning and self-training algorithms. Damage mitigation takes a more proactive approach by involving parents or guardians. In the next section, we propose and explain our solution, followed by a discussion on major contributions, challenges of the project, and project requirements.

Background

Cyberbullying is not a new phenomenon. Several studies reported cyberbullying prevalence, however, assessing severity and range remains difficulty. Because of the subjective nature and often inaccessible data or insufficient measurement tools, reports are inconsistent. Overall CB victimization prevalence has been reported as high as 72% in US middle and high school aged students (Selkie, Fales, & Moreno, 2016). The scope of harm on victim is larger in CB than that of traditional bullying. Online contents can reach a broader audience in a short span of time, especially when the bullying user has many followers or friends on social media. In several cases, the effect of CB on victims was so severe that victims ultimately committed suicide (Nikolaou, D., 2017). In less fatal cases, character development of young victims can be negatively impacted, leading to personality disorders, lack of self-confidence and social misbehavior (Extremera, Quintana-Orts, Mérida-López, & Rey, 2018; Nikolaou, 2017). It is therefore imperative to explore approaches of CB detection and prevention in an ever-increasing digital communication age.

Several studies on CB detection have been conducted, using data originating in social media platforms. A few studies have analyzed multiple platforms and compared their findings to see if a single detection model works for all examined platforms (Nahar, Al-Maskari, Li, & Pang, 2014; Squicciarini, Rajtmajer, Liu, & Griffin, 2015). However, CB detection research lacks cross-platform studies that integrate data from multiple platforms into a single large network. A CB incident can be comprised of a chain of events that occur across platforms; the interrelation of contents and actors across different platforms can significantly improve the accuracy of detection models. A few empirical studies have taken practical approach to CB detection and prevention, as recently demonstrated by Silva et al. (2016). They designed a mobile app, “BullyBlocker”, that provides parents a summary of the child’s risk to cyberbullying based on the frequency and severity of CB occurrences targeting the child in their own network. The app uses posts and comments directed to the child on Facebook to identify negative and insulting contents. The app does not however provide the actual identified content to parents nor it provides immediate prevention or a mean to act upon the incident. In this manner, it is more of an informational tool for parents. Moreover, it only uses the comments directed to the child under the child’s posts on Facebook. Cyberbullying can have a broader scope and go beyond one’s immediate network. For example, kids might be affected by an insulting public content posted by a user in their extended network (friends of friends). Not all cyberbullies befriend their victim on social media. In addition, CB can target a user on different social media platform, not necessarily Facebook. Detecting CB content is non-trivial as it is context specific and subjective. CB cases that involve well-known individuals such as celebrities or public figures have a broader scope and different context than cases that involve an individual teen. For example, in public CB, a crowd of users may attempt to cyberbully a celebrity by just trending a hashtag on Twitter without much of negative textual content (Tahmasbi & Rastegari, 2018). However, for an individual teen who is being cyberbullied, the CB instances may include more negative tone or different textual features such as personal pronouns or name calling. Therefore, a single umbrella approach to detect CB seems incomplete. Human intervention seems also less suitable because of the large volume of data. Hence, a decentralized customized learning module would be more suitable.

Empirical Design – ChatterShield

In this paper, we propose an empirical solution in the form of a mobile software application called ChatterShield that integrates various social media platforms to identify and act upon the CB contents in social networks. Our proposed solution is focused around the immediate and extended network of one’s social media network to provide a customized and not “one for all” detection model. This approach takes advantage of the small sample of data that is labeled by a parent who acts as a moderator. The verified CB instances are fed back to the self-training model to further improve the accuracy of the model. Figure 1 depicts the integration of the required system components explained below.
**Authorization and Setup.** The first step in using this system is for the child (primary user) to authorize a parent/guardian to access the contents shared by the child’s network. Many social media platforms provide Application Programing Interfaces (APIs) to allow developers access public and private user data. The authorization step must be done for each social network platform independently but is a one-time setup process. The child can then choose to add multiple moderators (guardian users) who have the privilege of reading detected CB content in the child’s network.

**Data Extractor.** The data extractor module scans the child’s immediate and extended network for new contents on all registered platforms every \( n \) minute. Where \( n \) is initially indicated by the parent in the account settings but will be adjusted periodically based on dynamic analysis of the user’s activity rate. The type of contents that the data extractor scans include the following: (1) New comment on child’s posts (from any user); (2) New private message to child (from any user); (3) New public/private post from child’s immediate friends; (4) New public (visible by child) post from child’s extended friends. These newly collected data are sent to the processing module periodically. The reason we include the extended network in our analysis is that even if the perpetrator and the victim are not directly connected through social network (e.g. not following each other on Instagram), public posts by the cyberbully may find their way through to reach the victim (e.g. a public post liked by a friend of the victim will be visible to the victim).

**Data Processing Module.** This module takes the input from the extractor module and processes it to extract features that were explained in the previous studies. The major tasks of the processing module are:

1. **Update the network structure:** The processing module extracts the new interactions among the users in the immediate and extended networks. Examples of different types of interactions include: a user likes or favorites a content posted by another user on Instagram or Twitter; a user retweets from another user on Twitter; a user comments on another user’s post. Based on the extracted interaction pattern, the processing module updates the current network structure. In this network, a node indicates a user and an interaction between two users is indicated by a link between the two users. Each link has a weight that indicated the strength of the interaction between two users. Further interactions will increase the weight of associated links.

2. **Update users’ cyberbullying risk score:** Another stream of input to the processing module is from the moderator. A guardian user can provide input to the system in two ways. First, in response to the inquiry from the learning module to verify the labeled instance as cyberbullying, a moderator can mark the instance as verified or not verified. In any case, the response is submitted to the processing module. The processing module uses the newly verified CB instance to update the risk factor associated with the poster of the content and any user (that belongs to the network) who has liked/favorited the content. Second, a moderator can manually submit incidents that occurred outside the social media context, such as a CB incident that targeted their child at school. This information is submitted to the processing module and can be used to further update the risk factor of the bully (in case the bully has an online presence in the network).

**Integration Module.** Users may have presence in multiple social network platforms. Thus, we collect data from multiple platforms. The main challenge is to integrate the users’ networks into one comprehensive network. For this task, we need to match user profiles across different platforms. Users may go by different usernames and profile bio in different social platforms. Therefore, the job of the integration module is to use machine-learning techniques to identify the profile matches among different platforms based on the similar patterns of contents posted by the user in different platforms and the user’s meta-data, as well as user’s network features similarities. Moreover, this module can periodically ask the moderator or the child for verifying the classified instances of user profile matches. It then uses the verified labels in its training component. Users (child or moderator) may also submit the matching manually to the integration module. The output of this module would be a profile names resolution table in the central database that includes in each instance, a list of user accounts aliases from different platforms that belong to the same person. The processing module have access to the profile-matching table. Thus, when updating the network structure, or a user’s risk factor, the same information will be updated for all accounts belonging to the same user.

**Incremental Self-training CB Detection Module.** This module is the core of ChatterShield. It uses machine learning and Natural Language Processing (NLP) techniques to classify the new contents into CB or non-CB instances. It uses the textual features extracted by the processing module as well the network measures and users’ meta-data of the poster (including risk score) in the classification process. To build an accurate self-training model, we first go through an initial learning phase in which the manually labeled contents are generated gradually by the moderator. The model incrementally trains itself until it reaches a
reasonable threshold of accuracy. After this initial phase, the model starts predicting the new instances. Meanwhile, the verified instances are fed to the module that are further used in continuous training.

**Central Database.** This database stores the output from processing module, integration module, and the self-training CB detection module. It is a central unit that records the data for all users from all platforms. It keeps the labeled data that are being used for continuous training of the model for a limited period. However, the user-related information including users’ metadata and network features never expire. Since we are using self-training models, a labeled data set is needed to initially train the model. Once enough labeled data is collected, the model can use the small labeled sample to start the training. The initial labels of CB contents are manually created by the moderator. Therefore, at this initial phase, a higher involvement of the moderator is expected until sufficient number of CB cases have been labeled. After this phase, the self-training model will just notify the moderator of a potential CB instance and the verified labeled instances are entered back to the model again and are used in continuous training for improved accuracy of the model. Thus, after the initial phase, the moderators will not have to have extensive engagement as they will be notified occasionally by the system to verify or deny an automatically labeled instance.

### Discussion

The proposed design addresses several CB detection issues that were mentioned in this paper. Our next step to complete this research is to develop the actual application. We anticipate that new challenges will arise once the actual system is in development. One of these challenges would be finding a trade-off between sensitivity of the model and its accuracy. If the detection model is too sensitive (high false positive), parents may gradually lose engagement since they may not rely on the app as before. If the model yields high false negatives, it may result in kids getting hurt. The verification feedback we receive from the moderators will inform the self-training module and potentially increase its accuracy and sensitivity. However, an appropriate evaluation measure will also be beneficial. Our suggested evaluation index is the ratio of verified labels by the moderators to the total identified instances by the detection module. This index can be combined with the ratio of missed positive instances that are manually reported by the parents to the total identified and verified instances. Other than the internal challenges, are external challenges that are imposed by social network platforms. One of them is the lack of social media platforms’ API availability.
Whether it is ethical to facilitate the access to the public data of internet users or not still is debatable. Not all social network platforms provide public access to their contents. Some of the ones that do offer public API, continuously update their privacy policy, and as a result, impose more restriction on the access of content through API functions. For example, Instagram has restricted access to follower list of public accounts, relationships, and commenting on public content (Constine, 2018). With such a dynamic API landscape, it may prove difficult to maintain a multi-platform CB detection model. Where APIs are not available, crawlers, scripts that scan websites for data, could be used to collect public data. Another external challenge is that the proposed design is an external application to the social media platforms. Thus, it does not have full control over which content to be published and which one to be banned. It is very challenging to establish an agreement with the platforms providers to make decision on publishing or blocking their content that are informed by the outcome of our application. Regardless, the proposed application is potentially capable of blocking the content from being published on the platform (keep the sensitive content pending until approved or denied by the moderator). However, until such agreement is established, preventive solution is not guaranteed.

**Conclusion**

In this paper, we proposed a multi-platform, incremental self-training system that uses a decentralized learning approach to automatically detect cyberbullying instances on one’s extended online social network. To improve the self-training model, the crowdsourced feedback of human moderators (guardians) is used. This system contributes to the research by addressing several challenges in CB detection research: the computational cost resulting from large volumes of online streaming data, lack of context-sensitive, cross-platform CB detection, lack of human moderation in evaluating the detection model, and a disconnection between physical bullying incidents and cyberbullying incidents. Our future plan for this research is to implement the actual solution, evaluate the system, and address the potential challenges that may arise during the implementation.

**References**


