Towards User Modelling in the Combat Against Cyberbullying

Dadvar, M.; Ordelman, R.; de Jong, F.M.G.; Trieschnigg, D.

published in
Proceedings of the17th International Conference on Applications of Natural Language to Information Systems, NLDB 2012
2012

document version
Peer reviewed version

Towards User Modelling in the Combat against Cyberbullying

Maral Dadvar[1], Roeland Ordelman[1], Franciska de Jong[1], and Dolf Trieschnigg[2]

[1] Human Media Interaction Group
[2] Database Group, University of Twente,
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{m.dadvar,f.m.g.dejong,r.j.f.ordelman,
r.b.trieschnigg}@utwente.nl

Abstract. Friendships, relationships and social communications have all gone to a new level with new definitions as a result of the invention of online social networks. Meanwhile, alongside this transition there is increasing evidence that online social applications have been used by children and adolescents for bullying. State-of-the-art studies in cyberbullying detection have mainly focused on the content of the conversations while largely ignoring the users involved in cyberbullying. We hypothesis that incorporation of the users' profile, their characteristics, and post-harassing behaviour, for instance, posting a new status in another social network as a reaction to their bullying experience, will improve the accuracy of cyberbullying detection. Cross-system analyses of the users' behaviour - monitoring users' reactions in different online environments - can facilitate this process and could lead to more accurate detection of cyberbullying. This paper outlines the framework for this faceted approach.

1 Introduction

Young people have fully embraced the internet for socializing and communicating. It first started with a simple two-way stream of communication between two people. For example sending and receiving emails. Later on it expanded by having several people communicating at the same time in a particular online environment, such as a chat room, a discussion forum or commenting on the same video. The rise of social networks in the digital domain has led to a new definition of friendships, relationships and social communications. People interact through different social networks, such as Facebook, Twitter, MySpace, and YouTube at the same time. A comment is posted on a friend's video on YouTube, and a reply is received on Twitter. Meanwhile, alongside this vast transition of information, ideas, friendships and comments, an old troubling problem arises with a new appearance in new circumstances: cyberbullying, or online bullying. There is increasing evidence that social media have been used by children and adolescents for bullying [1]. Cyberbullying is defined as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact (e.g. email and chat rooms) repeatedly or over time against a victim who cannot easily defend themself [2]. Cyberbullying can have deeper and longer-lasting effects

G. Bouma et al. (Eds.): NLDB 2012, LNCS 7337, pp. 277–283, 2012.

compared to physical bullying. Online materials spread fast and there is also the persistency and durability of online materials and the power of the written word [1]. The victim and bystanders can read what the bully has said over and over again. Bullying can cause depression, low self-esteem and there have been cases of suicide among teenagers [3].

Cyberbullying is a well-studied problem from the social perspective [1, 4] while few studies have been dedicated to automatic cyberbullying detection [5, 6]. The main focus of these studies is on the content of the text written by the users rather than the users' information and characteristics. State-of-the-art studies have investigated the detection of bullying in a single environment at a single time without considering the further effects and reactions of the user toward this act in other social networks. For instance, if someone gets bullied on Facebook, later on, their Twitter postings can be an indication of their feelings and their state of mind. Focusing on the text itself and finding harassing sentences is not enough to conclude that the act of bullying has taken place. Profanities in a discourse do not necessarily mean that they are being used to bully someone. There are many foul words that are used among teenagers just as a sign of friendship and close relationships. Moreover, being bullied and becoming a victim of cyberbullying is also dependent on the personality of each person. One person may feel bullied, threatened and depressed by sentences that do not cause any bad feelings for someone else. Therefore, even if a sentence is harassing and is used with the intention of bullying someone, it does not necessarily mean that the other party was offended or felt bullied. These subtleties complicate the differentiation of "bullying" detection from "harassment" detection.

Use Case and Applications

The main role of an effective cyberbullying detection system in a social network is to prevent or at least decrease the harassing and bullying incidents in cyberspace. It can be used as a tool to support and facilitate the monitoring task of the online environments. For instance, having a moderator specially in the forums that are mostly used by teenagers is a common thing. But because of the volume of entries in these fora it is impossible for moderators to read everything. A system that gives warnings if something suspicious is detected would greatly help the moderator to focus only on these cases instead of randomly reading the fora. Moreover, cyberbullying detection can be used to provide better support and advice for the victim as well as to monitor and track the bully. Tracking the behaviour of the users involved in an incident, across different social networks in a time frame can help to conclude that whether there is a victim or a bully.

2 State-of-the-Art

For several topics related to cyberbullying detection, research has been carried out based on text mining paradigms, such as identifying online sexual predators [7] and spam detection [8]. However, very little research has been conducted on technical solutions for cyberbullying detection, for which lack of sufficient and appropriate

training datasets, privacy issues and ambiguities in definition of cyberbullying can be some of the reasons. The related studies provide some inspiration for cyberbullying detection but their approaches are not directly suitable for this problem. For instance, the main difference between a spam message/email and a harassing one is that the former is usually about a different topic than the topic of discussion. In a recent study on cyberbullying detection Dinakar et al. [6], applied a range of binary and multiclass classifiers on a manually labelled corpus of YouTube comments. Their findings showed that binary classifiers can outperform the detection of textual cyberbullying compared to multiclass classifiers. They have illustrated the application of common sense knowledge in the design of social network software for detecting cyberbullying. The authors treated each comment on its own and did not consider other aspects to the problem as such the pragmatics of dialogue and conversation and the social networking graph. They concluded that, taking such features into account will be more useful on social networking websites and can lead to better modelling of the problem. Yin et al. [5] used a supervised learning approach for detecting harassment. They used content, sentiment, and contextual features of documents to train a support vector machine classifier for a corpus of online posts. In this study only the content of the posts were used to determine either a post was harassing or not, and the characteristics of the author of the posts were not considered. Yin et al. [5] have used the combination of these three features. Their results show improvements over the baselines. In another study with the same dataset the authors tried to identify clusters containing cyberbullying using a rule-based algorithm [9]. A newly emerging field of work, which will be integrated into our study, is the issue of identifying users via interaction over the web. While providing profile information for social networks, browsing the web, users leave large number of traces. This distributed user data can be used as a source of information for systems that provide personalized services for their users or need to find more information about their users [14]. Connecting data from different sources has been used for different purposes, such as standardization of APIs (e.g. OpenSocial[1]) and personalization [10]. In another study authors evaluated cross-system user modelling and its impact on cold-start recommendations on real world datasets from three different social web systems [11]. The social connections of an individual user across different services can be obtained by Google's Social Graph API [2].

3 Proposed Approach

We propose that the incorporation of the users' information, their characteristics, and post-harassing behaviour, alongside the content of their conversations, will improve the accuracy of cyberbullying detection. We will investigate the cyberbullying detection from two perspectives. First, which is the conventional way, the users' behaviour will be considered only in a single environment, for instance, the user's comments on a video on YouTube. We envision an algorithm that would go through the comments' body and would classify them as either bullying or non-bullying. At this phase of the

[1] http://code.google.com/apis/opensocial/
[2] http://socialgraph.apis.google.com

experiment we hypothesize that including the users' characteristics – either the bully or the victim - such as age and gender, will improve the detection accuracy. Social studies show that there are differences between males and females in the way that they bully each other. For instance, Argamon et al. [12] found that females use more pronouns (e.g. "you", "she") and males use more noun specifiers (e.g. "the", "that").

Our second perspective would be cross-system analysis of users' behaviour. As we mentioned earlier, a content-based approach is not sufficient to classify a sentence as a bullying one. It also depends on the impact of the content on the person that it has been directed to. One way to understand how a person feels is to study the way that they respond and react to the harassing sentences. This can be the user's next reply to the comment in the same environment or it can be in another form of reaction in another environment, for example, it can be a new tweet on the same user's Twitter profile. By identifying the same users in different social networks we can monitor their behaviour and see how they react after a case of harassment in one initial starting point and whether the harassment has led to a bullying case or not. The above mentioned facts motivated us to concentrate more on the information and behaviour of the users involved in the conversation (bully or victim) rather than only the content of the conversation itself. Our approach is the first attempt to incorporate social networking graphs and user information into both cross-system and single system automatic cyberbullying detection (see Figure 1).

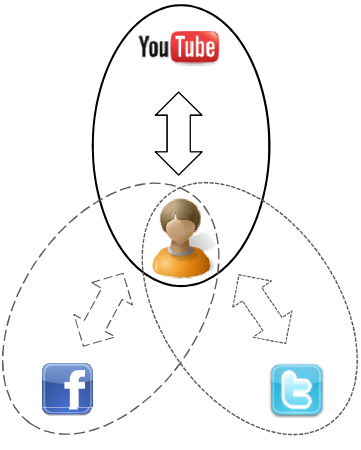

Fig. 1. The conceptual model of the proposed approach. Incorporation of user information into automatic cyberbullying detection both within a particular system and multiple systems.

Data Collection

One important shortage in this field of study is the lack of a standard labelled dataset. As in the state-of-the-art studies the personal information of the users such as age and gender is not taken into account, the currently available datasets are not suitable for our study. Not having a precise definition of the phenomena referred to as cyberbullying is also an obstacle for agreement on what could be a useful dataset for training.

Moreover, the ratio of bullying to non-bullying comments/posts is small, therefore, collecting training data to evaluate our approach is a challenging and time-consuming task. The most urgent need is a text dataset, such as comments or discussions, which contains a sufficient number of bullying posts. In order to study the effect of user's information on a better detection, the dataset authors should consist of different gender groups as well as age groups. Meanwhile, to train the classifier we need to have a labelled dataset. One avenue we are currently exploring to label a data set is the use of crowd sourcing. By using manual annotations obtained through platforms such as MTurk or Crowdflower, we will be able to label large amounts of data. For developing the dataset which will be used for cross-system analysis, first we will manually identify a set of users who are involved in conversations which are considered to be bullying, in a network such as YouTube. Then, using applications such as Google Social Graph API and Mypes[3], we will select those who also have a publicly accessible profile in another network, such as Twitter. By using the networks' streaming API, the selected users' posts and activities will be traced and collected for a period of 6 to 9 months, to make sure that all types of posts (bullying and non-bullying) are collected.

Auxiliary Information and Cross-System Analysis

We will use a supervised learning approach to train a classifier for detecting online bullying. We will employ a Support Vector Machines (SVM) model in WEKA [13] as classification tool. The target language for this experiment is English. To develop our model and train our classifier we will use several types of features. We will employ the TFIDF value of profane words in each post including their abbreviations and acronyms. The other feature is the TFIDF value for personal pronouns used in each post, grouped into second person and other pronouns. We will also make use of user features, such as age (adult versus teenager) and gender. The classifiers will be trained separately for each group. A plausible way to monitor post-harassing behaviour of the users and track their reactions in other systems is users' cross-system modelling. Aggregation of users' profiles from different systems can provide us with more information. Mypes is one of the available services that allow this [14]. By finding and tracing the users' behaviour over a period of time we can gain more accurate information about whether a user is a real bully or a victim. Over time it may become clear whether someone uses a vulgar behaviour towards everyone and in all conversations or has only targeted a specific person to bully. Similarly, by analysing a user's posts and activities, we can find how they react to other people's comments and behaviours and whether they are being victimized.

4 Discussion and Future Work

The main focus of the technical studies which have been conducted so far on cyberbullying detection is mainly on textual content and there is one single system under

[3] http://mypes.groupme.org/mypes/

study at a time. We propose the incorporation of the users' profiles, such as age and gender, their characteristics, and post-harassing behaviour, to improve the accuracy of cyberbullying detection and to take into account the social networking graph. One of the main shortcomings in this field of study is the lack of access to suitable datasets, mostly because of privacy issues. The bullying comments and conversations that happen in public access networks, such as YouTube, are usually not continuous and their negative effects are softened by the support of other users who fight back the bullies and are against hate messages. There are also instances where the victim even becomes a hero for other users because of being a bullying target. Therefore, the real bullying incidents that are a match to our definition of bullying, mostly happen in private conversation such as chat logs and private messages on network profiles. So far we have investigated the gender-based approach for cyberbullying detection in a particular system, in which we observed improvements in classification. We are also going to investigate the age groups differences in cyberbullying. The next phase of our research aims to identify the victim of cyberbullying through monitoring the user's behaviour after a potential case of bullying across other systems. This means that after detecting a harassing post, we investigate the effect of the sentence on the targeted user and whether it had caused the user to feel bullied or threatened. To do so we can employ cross-system user modelling and make use of services that allow the aggregation of users' profiles.

References

1. Campbell, M.A.: Cyber bullying: An old problem in a new guise? Australian Journal of Guidance and Counselling 15, 68–76 (2005)
2. Espelage, D.L., Swearer, S.M.: Research on school bullying and victimization. School Psychology Review 32, 365–383 (2003)
3. Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., Tippett, N.: Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry 49, 376–385 (2008)
4. Kowalski, R.M., Limber, S.P., Agatston, P.W.: Cyber bullying: Bullying in the digital age, p. 224. Blackwell Publishing (2008)
5. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on Web 2.0. In: Proceedings of CAW2.0, Madrid, April 20-24 (2009)
6. Dinakar, K., Reichart, R., Lieberman, H.: Modelling the Detection of Textual Cyberbullying. In: ICWSM 2011, Barcelona, Spain, July 17-21 (2011)
7. Kontostathis, A.: ChatCoder: Toward the tracking and categorization of internet predators. In: Proceedings of SDM 2009, Sparks, NV, May 2 (2009)
8. Tan, P.N., Chen, F., Jain, A.: Information assurance: Detection of web spam attacks in social media. In: Proceedings of Army Science Conference, Orland, Florida (2010)
9. Chisholm, J.F.: Cyberspace violence against girls and adolescent females. Annals of the New York Academy of Sciences 1087, 74–89 (2006)
10. Carmagnola, F., Cena, F.: User identification for cross-system personalisation. Information Sciences 179, 16–32 (2009)

11. Abel, F., Araújo, S., Gao, Q., Houben, G.-J.: Analyzing Cross-System User Modeling on the Social Web. In: Auer, S., Díaz, O., Papadopoulos, G.A. (eds.) ICWE 2011. LNCS, vol. 6757, pp. 28–43. Springer, Heidelberg (2011)
12. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. Text - Interdisciplinary Journal for the Study of Discourse 23, 321–346 (2003)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Newsletter 11, 10–18 (2009)
14. Abel, F., Henze, N., Herder, E., Krause, D.: Linkage, aggregation, alignment and enrichment of public user profiles with Mypes. In: Proceedings of I-SEMANTICS, Graz, Austria, pp. 1–8 (September 2010)