Cyberbullying Detection; A Step Toward a Safer Internet Yard

Maral Dadvar Franciska de Jong
Human Media Interaction Group, University of Twente
PO Box 217, 7500 AE, Enschede, the Netherlands
m.dadvar@utwente.nl f.m.g.dejong@utwente.nl

ABSTRACT

As a result of the invention of social networks friendships, relationships and social communications have all gone to a new level with new definitions. One may have hundreds of friends without even seeing their faces. 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 propose that incorporation of the users' information, 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 their reactions in different online environments - can facilitate this process and provide information that could lead to more accurate detection of cyberbullying.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing – *Linguistic processing.*

General Terms

Algorithms, Experimentation, Security, Human Factors, Languages.

Keywords

Bullying, Cross-system analysis, Cyberharassment, Sentiment analysis, Social network, User profile.

1. PROMBLEM

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 same video or picture. 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 Facebook. Meanwhile, alongside this vast transition of information, ideas, friendships, comments and opinions, an old troubling problem arises with a new appearance in new circumstances: cyberbullying. We focus in particular on cyberbullying among children and teenagers.

Traditionally bullying was considered to be a face-to-face encounter between children and adolescents in school yards, but now it has also found its way into the cyberspace. There is increasing evidence that online social applications 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 herself [2].

Cyberbullying can have deeper and longer-lasting effects compared to physical bullying [1]. Online materials spread fast and they have a wider range of audience. There is also the persistency and durability of online materials and the power of the written word [1]. In the case of cyberbullying through text the targeted 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 a 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 (both the victim and the bully) rather than the users' information and characteristics. For instance, as we will explain in more detail later on, there are differences in the way that boys and girls bully each other. The state-of-the-art studies have investigated the detection of bullying in a particular environment 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, her Twitter postings can be an indication of her feelings and her state of the mind.

Looking at the linguistic features, focusing on the text by itself and finding harassing sentences is not enough to conclude that the act of bullying has taken place. The 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. Therefore, profanity in a sentence cannot be a definite proof for bullying. Moreover, being bullied and becoming a victim of cyberbullying is also dependent on the personality of each person. Someone 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

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

does not necessarily mean that the other party was offended or felt bullied. These subtle points complicate the differentiation of "bullying" detection from "harassment" detection.

2. USE CASE AND APPLICATIONS

The main application 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. So a system that gives warnings if something suspicious is detected would greatly help the moderator to only focus 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 monitoring and tracking the bully. One incident cannot be a certain indication that users involved are a victim or a bully. But following their behaviours after the incident and across different social networks in a time frame can bring us to a more established conclusion as to whether either of them is a victim or a bully. Having sufficient and precise information about the case, allows the people in charge (for instance, teachers) to provide the required help and guidance for the victim or the bully.

3. 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], vandalism detection [8], spam detection [9] and detection of internet abuse and cyberterrorism [10]. However, very little research has been conducted on technical solutions for cyberbullying detection. 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. Spams are mostly commercial advertisements about a product or a service.

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 individual topic-sensitive classifiers can outperform the detection of textual cyberbullying compared to merged dataset or multiclass classifiers. They have illustrated the application of commonsense 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 into account such features will be more useful on social networking websites and can lead to a 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 is 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. In their study N-grams, TFIDF weighting and foul words frequency were used as the baselines. The 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 [11].

A new emerging field of work, that we believe we will integrate it to our study at some point, is the issue of identifying users via interaction over the web and it has been addressed in several applications such as personalization [12]. 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 [13]. In contrast to personalization in social tagging systems that targets single systems [14, 15], cross-system personalization is transferable to other systems.

4. 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 one environment, for instance, the user's comments on a video on YouTube. We envision an algorithm that would go through the comments' text and would classify them as either bullying or non-bullying. At this phase of the 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. Females tend to use relational styles of aggression, such as excluding someone from a group and ganging up against them whereas males use more threatening expressions and profane words [11]. Argamon et al. [16] found that females use more pronouns (e.g. "I", "you", "she") and males use more noun specifiers (e.g. "a", "the", "that").

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 the person feels, is the way that she responds and reacts 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 status on the same user's Facebook 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 user information into automatic cyberbullying detection both within a particular as the baselines (see Figure 1).

5. METHODOLOGY

One important shortage in this field of study is the lack of an appropriate standard labelled dataset. We need a text dataset, such as comments or discussion, and it is required to label the bullying ones. Since the ratio of harassment to non-harassment comments/posts is small, collecting training data to evaluate our approach is a challenging task. Since it is not always obvious when someone is bullied, labelling is a difficult and time consuming procedure. The current available datasets are not appropriate for our study. We have to first develop a suitable dataset. It should contain a sufficient number of harassing posts.

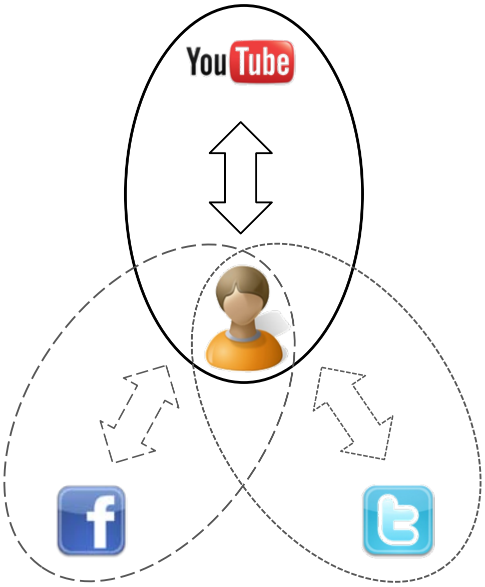

Figure 1. The conceptual model of the proposed approach. Incorporation of user information into automatic cyberbullying detection both within a particular system and multiple systems.

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 amount of data.

5.1 Auxiliary information

We will use a supervised learning approach to train a classifier for detecting online harassment. We will employ Support Vector Machines (SVM) model in WEKA [17] as our classification tool. 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 values of personal pronouns used in each post, grouped to second person pronouns and other pronouns. We will also make use of auxiliary information of the users, such as age and gender. For this purpose, we will split our dataset into male and female authored posts or based on the age groups, into adult and teenagers authored posts. The classifiers will be trained separately for each group.

5.2 Cross-system analysis

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 about them. *Mypes* is a service that allows the aggregation of users' profiles [18]. Mypes supports the task of gathering information about users aims to provide a uniform interface to public profile data distributed on the Social Web [19]. To employ this subject in our research we still need more thorough studies and investigations.

6. PRILIMINARY RESULTS

Here our preliminary experiment, cyberbullying detection using user's information in a single system [20], is presented.

6.1 Dataset

For the first part of this study we used MySpace posts as our dataset which was provided by Fundacion Barcelona Media[1]. MySpace is a social networking site on which users can participate in forum discussions about predefined topics. This dataset consists of more than 381,000 posts in about 16,000 threads. Overall, 34% of posts are written by female and 64% by male authors. The ground truth dataset has 2,200 posts and has been manually labelled by three students as bullying or non-bullying. Any post that was marked bullying by at least two students, was labelled as bullying.

6.2 Approach

We studied the effect of gender-specific language features on the detection of cyberbullying in social networks. We hypothesized that the inclusion of gender-specific language features could improve the overall detection accuracy. The focus of this study was only on vocabulary and pronoun usage differences. To support our hypothesis that developing gender-specific features would lead to more accurate classification of harassing contents, we analyzed the use of foul words in 100,000 randomly selected posts from the dataset. We compared the foul words used most frequently by each gender and, based on a Wilcoxon signed rank test, determined that male and female authors used significantly ($p < 0.05$) different frequencies of foul words in their posts (see Figure 2).

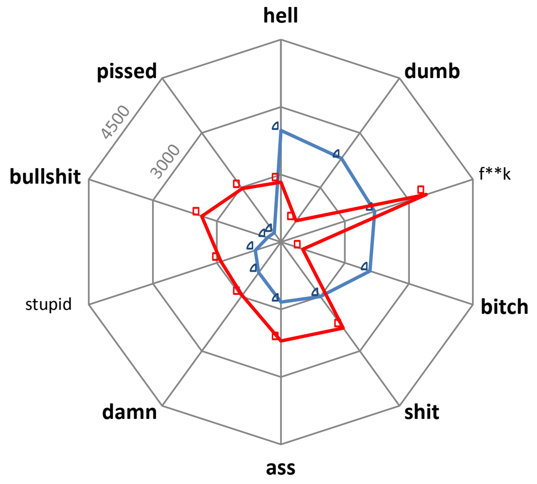

Figure 2. Top ten frequently used foul words by female (circle) versus male (square) authors.

We used a supervised learning approach, Support Vector Machine classifier using WEKA to detect cyberbullying. For our baseline, we used four types of features which are more frequently used for harassment classification [5]; First, profane words including their abbreviations and acronyms. This feature was obtained by treating

[1] Available at http://caw2.barcelonamedia.org

all the profane words of each post as a single term and then calculating the ratio of the foul words in the post. The number of foul words in a post was normalized by dividing by the post length. Personal pronouns are frequently used in harassing posts, which can be another sign for the occurrence of harassment. The second feature is the second person pronouns and the third feature is all the other pronouns. For both of these features, we treated all the pronouns of each post as one single word and then we calculated the ratio of each pronoun in each post. Since the second person pronoun has a more important role in detecting online harassment, we separated them from the other pronouns. The fourth type is the TFIDF value of all the words in each post. In this study we split our dataset into male and female authored posts and trained two classifiers separately for each group. The ratios of foul words and pronouns were based on gender-specific language features.

6.3 Results

We employed the features mentioned earlier to train the classifier. We first used a corpus with posts written by both male and female users as our dataset. In the next step we trained the classifier separately for each gender group. We then calculated the final result, based on the proportion of each group in the whole corpus (34% female, and 66% male). To evaluate the classification accuracy we used 10-fold cross validation and calculated corresponding precision, recall and F-measure. The evaluation measures are given in Table 1.

Table 1. The accuracy measures for basic and gender-based approaches for cyberbullying detection in a MySpace corpus

Feature used in classifier	Precision	Recall	F-measure
Baseline	0.31	0.15	0.20
Gender-specific	0.43	0.16	0.23
Female-specific (34% corpus)	0.40	0.05	0.08
Male-specific (66% corpus)	0.44	0.21	0.28

Incorporation of gender-specific features improved the overall accuracy measures. This algorithm gave better detection results in male-specific posts in comparison to female-specific. This can be due to the small size of the training dataset for female harassing posts. Another reason can be the usage of foul words by girls and boys - girls tend to use less explicit profanities, and express more indirect negative and excluding attitude in their sentences.

7. CONCLUSION AND FUTURE WORK

Cyberbullying is a growing problem in the social web and is becoming a major threat to teenagers and adolescents. The main focus of the technical studies which have been conducted so far on cyberbullying detection is mainly on the content of the text written by the users but not the users' information.

We hypothesized that incorporation of the users' information, such as age and gender, will improve the accuracy of cyberbullying detection. In this study we have investigated the gender-based approach for cyberbullying detection in MySpace, in which we observed improvements in classification. Our analysis showed that author information can be leveraged to improve the detection of misbehaviour in online social networks.

In future stages this work will be extended by considering contextual features of the text as well as the word level features. In the dataset that was used in this study the gender of the authors was known, while this might not always be the case. Using a gender detector beforehand might be a way to cope with this limitation. It would also be interesting to consider the pragmatics of conversations between authors of same gender versus opposite gender. Moreover, it is worthwhile to compare different classification approaches and analyse their performances.

One limitation for the experiment conducted was the limited size of the dataset. A larger and more diverse dataset will be developed for future work in automatic cyberbullying detection. The ground truth annotation can be done through crowdsourcing. We are also going to investigate other features which may differentiate the writing styles of the users such as age, profession, and educational level. For this purpose we need a dataset which contains sufficient number of harassing posts authored by each group. This will be based on collaboration with potential users to take into account the requirements inherent to real use scenarios. Also, a social scientist will be consulted for the definition of an enlarged feature set.

8. ACKNOWLEDGMENTS

We thank R. Ordelman, D. Trieschnigg, and C. Hauff for their valuable insights. The language editing was kindly done by Lynn Packwood. We are grateful to reviewers for their valuable comments and suggestions. This research is funded by the EU project PuppyIR http://www.puppyir.eu (EU FP7 231507).

9. REFERENCES

[1] Campbell, M. A. 2005. Cyber bullying: An old problem in a new guise? *Australian Journal of Guidance and Counselling,* 15, 68-76.

[2] Espelage, D. L. and Swearer, S. M. 2003. Research on school bullying and victimization: What have we learned and where do we go from here? *School Psychology Review,* 32, 365-383.

[3] Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., and Tippett, N. 2008. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry,* 49, 376-385.

[4] Kowalski, R. M., Limber, S. P., and Agatston, P. W. 2008. Cyber bullying: Bullying in the digital age. *Blackwell Publishing,* 224 pp.

[5] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. 2009. Detection of harassment on Web 2.0. In *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009* (Madrid, Spain, April 20-24, 2009).

[6] Dinakar, K., Reichart, R., and Lieberman, H. 2011. Modelling the Detection of Textual Cyberbullying. *Social Mobile Web Workshop at International Conference on Weblog and Social Media* (Barcelona, Spain, July 17-21, 2011).

[7] Kontostathis, A. 2009. ChatCoder: Toward the tracking and categorization of internet predators. In *Proceedings of Text Mining Workshop Held in Conjunction with the Ninth SIAM International Conference on Data Mining (SDM 2009)* (Sparks, NV, May 2, 2009).

[8] Smets, K., Goethals, B., and Verdonk, B. 2008. Automatic vandalism detection in Wikipedia: Towards a machine learning approach, *Wikipedia and Artificial Intelligence: an Evolving Synergy (WikiAi08) Workshop by Association for the Advancement of Artificial Intelligence,* pp. 43–48.

[9] Tan, P. N., Chen, F., and Jain, A. 2010. Information assurance: Detection of web spam attacks in social media. In *Proceedings of the 27th Army Science Conference* (Orland, Florida, 2010).

[10] Simanjuntak, D. A., and Ipung, H. P. 2010. Text Classification Techniques Used to Facilitate Cyber Terrorism Investigation. *Second International Conference on Advances in Computing, Control and Telecommunication Technologies (ACT)*, 198-200.

[11] Chisholm, J. F. 2006. Cyberspace violence against girls and adolescent females. *Annals of the New York Academy of Sciences*, 1087, 74-89.

[12] Carmagnola, F., and Cena, F. 2009. User identification for cross-system personalisation. *Information Sciences,* 179, 16-32.

[13] Abel, F., Araújo, S., Gao, Q., and Houben, G. J. 2011. Analyzing Cross-System User Modelling on the Social Web. In *Proceedings of Eleventh International Conference on Web Engineering (ICWE)* (Paphos, Cyprus, June 2011), 28-43.

[14] Sen, S., Vig, J., and Riedl, J. 2009. Tagommenders: connecting users to items through tags. In *Proceedings of International World Wide Web Conference* (Madrid, Spain, 2009). ACM, 671-680.

[15] Sigurbjörnsson, B., and Van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web* (Beijing, China, April 21-25, 2008). ACM, 327-336.

[16] Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. 2003. Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse,* 23,321-346.

[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter,* 11, 10-18.

[18] Abel, F., Henze, N., Herder, E., and Krause, D. 2010. Linkage, aggregation, alignment and enrichment of public user profiles with Mypes. In *Proceedings of 6th International Conference on Semantic Systems (I-SEMANTICS)* (Graz, Austria, September 2010). ACM, 1-8.

[19] Abel, F., Herder, E., Houben, G.J., Henze, N., and Krause, D. 2011. Cross-system user modelling and personalization on the social web. *User Modelling and User-Adapted Interaction (UMUAI),* Special Issue on Personalization in Social Web Systems 22, 1-42.

[20] Dadvar, M., F. de Jong, Ordelman, R. and Trieschnigg, D. 2012. Improved Cyberbullying Detection Using Gender Information, In *Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop (DIR2012)* (Ghent, Belgium 2012).