Improving Cyberbullying Detection with User Context

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

[1] Human Media Interaction Group, University of Twente, Netherlands
{m.dadvar,r.j.f.ordelman,f.m.g.dejong}@utwente.nl
[2] Database Group, University of Twente, Netherlands
d.trieschnigg@utwente.nl

Abstract. The negative consequences of cyberbullying are becoming more alarming every day and technical solutions that allow for taking appropriate action by means of automated detection are still very limited. Up until now, studies on cyberbullying detection have focused on individual comments only, disregarding context such as users' characteristics and profile information. In this paper we show that taking user context into account improves the detection of cyberbullying.

1 Introduction

More and more teenagers in online communities are exposed to and harmed by cyberbullying. Studies [1] show that in Europe about 18% of the children have been involved in cyberbullying, leading to severe depressions and even suicide attempts. Cyberbullying is defined as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact repeatedly or over time, against a victim who cannot easily defend him- or herself [1]. Besides social measures, technical solutions have to be found to deal with this social problem. At present social network platforms rely on users alerting network moderators who in turn may remove bullying comments. The potential for alerting moderators can be improved by automatically detecting such comments allowing a moderator to act faster. Studies on automatic cyberbullying detection are few and typically limited to the individual comments and do not take context into account [2-3]. In this study we show that taking user context, such as a user's comments history and user characteristics [4], into account can improve the performance of detection tools for cyberbullying incidents considerably. We approach cyberbullying detection as a supervised classification task for which we investigated three incremental feature sets. In the next sections the experimental setup and results will be described, followed by a discussion of related work and conclusions.

[1] EU COST Action IS0801on Cyberbullying (https://sites.google.com/site/costis0801/).

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 693–696, 2013.

2 Experiment

2.1 Corpus

YouTube is the world's largest user-generated content site and its broad scope in terms of audience, videos, and users' comments make it a platform that is eligible for bullying and therefore an appropriate platform for collecting datasets for cyberbullying studies. As no cyberbullying dataset was publicly available, we collected a dataset of comments on YouTube movies. To cover a variety of topics, we collected the comments from the top 3 videos in the different categories found in YouTube. For each comment the user id, its date and time were also stored. Only the users with public profiles (78%) were kept. The final dataset consists of 4626 comments from 3858 distinct users. The comments were manually labelled as bullying (9.7%) and non-bullying based on the definition of cyberbullying in this study (inter-annotator agreement 93%). For each user we collected the comment history, consisting of up to 6 months of comments, on average 54 comments per user.

2.2 Feature Space Design

The following three feature sets were used to train cyberbullying classifier.

Content-Based Features. These features are based on the contents of the comments itself and are frequently used for sentiment analysis. The following features are included: 1) The *number of profane words* in the comment, based on a dictionary [2], normalized by the total number of words in the comment. The dictionary consists of 414 profane words including acronyms and abbreviation of the words. The majority of the words are adjectives and nouns. 2) To detect the comments which are personal and targeting a specific person, we included the normalized *number of first and second person pronouns* in the comment, based on a list of pronouns. 3) *Profanity windows* of different sizes (2 to 5 words) were chosen. These are Boolean features which indicate whether a second person pronoun is followed by a profane word within the size of the window. 4) To capture explicit emotions, *the number of emoticons* was counted and normalized by the number of words. And finally 5) to capture shouting in comments, the *ratio of capital letters* in a comment was computed.

Cyberbullying Features. The second set of features aims at identifying frequent bullying topics such as minority races, religions and physical characteristics. It consists of: 1) the (normalized) *number of cyberbullying words*, based on a manually compiled dictionary, and 2) in order to detect typically short bullying comments, *the length of the comment*.

User-Based Features. To be able to exploit information about the background of the users in the detection process, we looked at the *history of user's activities* in our dataset and used the averaged content-based features on the users' history to see whether there was a pattern of offensive language use. We checked the frequency of profanity in their previous comments. Also, other linguistic characteristics such as number of pronouns, average length of the comments and usage of capital letters and the use of

[2] `http://www.noswearing.com/dictionary` [September 2012].

emoticons were taken into account. As type of words and language structures may vary in different ages, we also considered the *age of the users* as a feature.

2.3 Experimental Setup

We used the three incremental feature sets for training a Support Vector Machine to classify comments as bullying or non-bullying. As a baseline we only used content-based features (further referred to as Set 1). For Set 2 we included the cyberbullying features and for Set 3 also the user-based features (Set 3) were used. As a pre-processing step, stop-word removal and stemming were applied. We used 10-fold cross validation evaluated with precision, recall and *F*-measure.

3 Results and Discussion

The results of our experiments are listed in Table 1. It shows that detection performance improves when we add more bullying-specific features and that it improves further when context information is added. For Set 1, bag of profane words, pronoun-profanity windows, and second person pronouns' frequency were the main contributing features. Capital letters and emoticons however, did not add a significant contribution. This observation indicates that in the YouTube dataset, capital letters are not more frequently used in bullying comments and emoticons are not necessarily more frequent in non-bullying comments. The low recall of the first feature set can be explained by the occurrence of bullying comments without explicit profanities and by implicit bullying through sarcasm, or comments addressing sensitive topics using other words than profanities. Adding cyberbullying features (Set 2) significantly ($p < 0.05$) improved both precision and recall. In Set 2 the length feature did not have any significant contribution, while updated bag of profane words contributed the most. With further analyses we observed that the most effective words for classification were vulgar words that refer to race and to sexuality. As we hypothesized, incorporation of users' profile information further improved the precision and the recall to 77% and 55% respectively. As the classification was not just based on one comment and one instance of profanity use, the non-bullying cases were identified more accurately which lead to higher precision. Moreover, the recall was also improved as bullying comments without explicit profanities and appeared to convey neutral emotions now were correctly identified as bullying by considering the background of their authors. The number of profanities in the history of each user had the highest contribution, and the age feature had contributed but not as much as expected in the classification of bullying comments. The latter might be due to the fact that many users do not indicate their real personal information.

Table 1. Summary of the experiment results

Feature sets	Precision	Recall	*F*-measure
Set 1 (Content-based)	0.72	0.45	0.55
Set 2 (Set 1 + Cyberbullying)	0.75	0.51	0.60
Set 3 (Set 2 + User-based)	0.77	0.55	0.64
Set 3 – [number of profanities in user's history]	0.76	0.52	0.62
Set 3 – [number of profanities]	0.78	0.54	0.63
Set 3 – [pronoun-profanity window]	0.76	0.55	0.63

4 Related Works

Due to space limitations, we provide references to studies on profanity and offensiveness detection [2, 5-6] and only address recent studies on cyberbullying detection based on YouTube comments. Because of privacy issues the datasets used in these studies were not accessible. Dinakar et al. [3] applied a set of features similar to our baseline, along with some other features which were specific to the topic of the videos. They showed that using topic-based features improves classification. Chen et al. [7] proposed the use of a lexical syntactic feature approach to detect the level of offensiveness in the comments and potentially offensive users. They also considered the writing style of the users, but for identification of the potential offensive users rather than for detecting bullying comments. As the data sets are different, it is not possible to come up with a clear comparison of our results and those from the other studies.

5 Conclusion and Future Work

In this paper, we presented the results of a study on the detection of cyberbullying in YouTube comments. We used a combination of content-based, cyberbullying-specific and user-based features. Our results showed that incorporation of context in the form of users' activity histories improves cyberbullying detection accuracy. This work can be extended to develop models that detect expressions involving sarcasm or implicit harassment. In future studies, other user features such as gender and the channels subscribed to could also be taken into account. Furthermore, since users' profile information is not always stated correctly, it might be beneficial to employ predicting algorithms such as age prediction, prior to using the profile information for improving detection accuracy.

References

1. Espelage, D.L., Swearer, S.M.: Research on school bullying and victimization: What have we learned and where do we go from here? School Psychology Review 32(3), 365–383 (2003)
2. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of Harassment on Web 2.0. In: Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW 2009, Madrid, Spain (2009)
3. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: International Conference on Weblog and Social Media - Social Mobile Web Workshop, Barcelona, Spain (2011)
4. Dadvar, M., de Jong, F.M.G., Ordelman, R.J.F., Trieschnigg, D.: Improved Cyberbullying Detection Using Gender Information. In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium, pp. 23–26 (2012)
5. Sood, S., Antin, J., Churchill, E.: Using Crowdsourcing to Improve Profanity Detection. In: AAAI Spring Symposium Series, pp. 69–74 (2012)
6. Kontostathis, A., Leatherman, L.E.A.: ChatCoder: Toward the tracking and categorization of internet predators. In: Proceedings of Text Mining Workshop 2009 held in Conjunction with the Ninth SIAM International Conference on Data Mining, Nevada, USA (2009)
7. Chen, Y., Zhu, S., Zhou, Y., Xu, H.: Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In: Symposium on Usable Privacy and Security, Pittsburgh, USA (2011)