Improved Cyberbullying Detection Using Gender Information

Maral Dadvar Franciska de Jong Roeland Ordelman Dolf Trieschnigg

Human Media Interaction Group, University of Twente
PO Box 217, 7500AE, Enschede, the Netherlands
{ m.dadvar , f.m.g.dejong , r.j.f.ordelman , r.b.trieschnigg } @utwente.nl

ABSTRACT

As a result of the invention of social networks, friendships, relationships and social communication are all undergoing changes and new definitions seem to be applicable. One may have hundreds of 'friends' without even seeing their faces. Meanwhile, alongside this transition there is increasing evidence that online social applications are 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 characteristics of the actors involved in cyberbullying. Social studies on cyberbullying reveal that the written language used by a harasser varies with the author's features including gender. In this study we used a support vector machine model to train a gender-specific text classifier. We demonstrated that taking gender-specific language features into account improves the discrimination capacity of a classifier to detect 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

Cyberharassment, Gender distinction, Social networks, Support vector machine, Text mining.

1. INTRODUCTION

Young people have fully embraced the internet for socializing and communicating. The rise of social networks in the digital domain has led to a new definition of friendships, relationships and social communications. One may have hundreds of friends without even seeing their faces. Meanwhile, alongside this vast transition, 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 are being 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 her-self [2].

Cyberbullying can have deeper and longer-lasting effects compared to physical bullying. Online materials spread fast and they have a wider 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 actors (both the victim and the bully) rather than the features and characteristics of those involved. For instance, as we will explain in more detail later on, there are differences in the ways boys and girls bully each other.

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. Having a moderator specially in the fora 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.

1.1 Overview of the 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 cyber-bullying 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 multiclass classifiers. They have illustrated the application of common sense

Copyright is held by the author/owner(s)

DIR'2012, February 23-24, 2012, Ghent, Belgium.

Copyright © 2012 Maral Dadvar

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 whether 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]. Our approach is the first attempt to incorporate gender information into automatic cyberbullying detection.

1.2 Two genders, two vocabularies

Social studies show that there are differences between males and females in the way 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. found that females use more pronouns (e.g. “I”, “you”, “she”) and males use more noun specifiers (e.g. “a”, “the”, “that”) [12]. These findings motivated our study of 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.

2. GENDER-BASED APPROACH

We used a supervised learning approach to detect cyberbullying. We constructed a Support Vector Machine classifier using WEKA [13]. We used MySpace posts as our dataset which was provided by Fundacion Barcelona Media[1]. The language of this dataset is English. In this experiment we assume that the gender of the posts authors’ is known, which is the case in this dataset. 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 harassing (positive) or non-harassing (negative).

To support our hypothesis that developing gender-specific features would lead to more accurate classification of harassing contents, we analysed the use of foul words in 100,000 randomly selected posts from the dataset. We compared the most frequently foul words[2] used 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, as shown in Figure 1.

For our baseline, we used four types of features which are more frequently used for harassment classification [5]; profane words, second person pronouns, other personal pronouns, and the weight of the words in each sentence.

Profane words including their abbreviations and acronyms[2]. This feature is obtained by treating 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 is 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 treat all the pronouns of each post as one single word and then we calculate the ratio of each pronoun in each post. Since the second person pronoun has a more important role in detecting online harassment, we separate 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 are based on gender-specific language features.

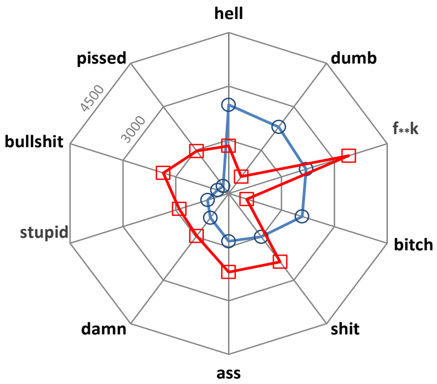

Figure. 1. Top ten frequently used foul words by female (circle) versus male (square) authors

3. EXPERIMENT AND 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.

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. The gender-specific approach improved the baseline by 39% (0.31 to 0.43) in precision, 6% (0.15 to 0.16) in recall, and 15% (0.20 to 0.23) in F-measure. Both precision and recall are important, but considering the usage of this algorithm and to stay of the safe

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

[2] Obtained from http://www.noswearing.com/dictionary

side, it is better to make sure that all the harassing comments are detected. Therefore, having high recall might be of more importance.

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

4. CONCLUSIONS 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 this work we treated each post individually, regardless of its interactions with other posts in a discussion.

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. A second line of future research will address the various use scenarios for the detection of bullying scenarios and the variation in detection approaches that may be required in order to deal adequately with different types of cyber contexts (e.g., MSN, chat rooms, e-mail, social networking services).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.

ACKNOWLEDGMENT

We thank Claudia Hauff for her valuable insights, and Brian Davison for sharing the labelled dataset employed in this paper. The language editing was kindly done by Lynn Packwood. We are grateful to three anonymous reviewers for their valuable comments and suggestions. This research is funded by the EU project PuppyIR http://www.puppyir.eu (EU FP7 231507).

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., 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., 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., Agatston, P.W. 2008. *Cyber bullying: Bullying in the digital age.* Blackwell Publishing.

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

[6] Dinakar, K., Reichart, R., Lieberman, H. 2011. *Modeling the Detection of Textual Cyberbullying.* Social Mobile Web Workshop at International Conference on Weblog and Social Media, Barcelona, Spain.

[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 (Sparks, NV, 2009).

[8] Smets, K., Goethals, B., 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., 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., Ipung, H.P. 2010. *Text Classification Techniques Used to Faciliate Cyber Terrorism Investigation.* pp. 198-200. IEEE.

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

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

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