# Experts and Machines Against Bullies:<br>A Hybrid Approach to Detect Cyberbullies

Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong

Human Media Interaction Group, University of Twente
POBox 217, 7500 AE, Enschede, the Netherlands
{m.dadvar, d.trieschnigg, f.m.g.dejong}@utwente.nl

**Abstract.** Cyberbullying is becoming a major concern in online environments with troubling consequences. However, most of the technical studies have focused on the detection of cyberbullying through identifying harassing comments rather than preventing the incidents by detecting the bullies. In this work we study the automatic detection of bully users on YouTube. We compare three types of automatic detection: an expert system, supervised machine learning models, and a hybrid type combining the two. All these systems assign a score indicating the level of "bulliness" of online bullies. We demonstrate that the expert system outperforms the machine learning models. The hybrid classifier shows an even better performance.

## 1 Introduction

With the growth of the use of Internet as a social medium, a new form of bullying has emerged, called cyberbullying. Cyberbullying 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 [1]. One of the most common forms is the posting of hateful comments about someone in social networks. Many social studies have been conducted to provide support and training for adults and teenagers [2, 3]. The majority of the existing technical studies on cyberbullying have concentrated on the detection of bullying or harassing comments [4-6], while there is hardly work on the more challenging task of detecting cyberbullies and studies for this area of research are largely missing. There are few exceptions however, that point out an interesting direction for the incorporation of user information in detecting offensive contents, but more advanced user information or personal characteristics such as writing style or possible network activities has not been included in these studies [7, 8]. Cyberbullying prevention based on user profiles was addressed for the first time in our latest study in which an expert system was developed that assigns scores to social network users to indicate their level of '*bulliness*' and their potential for future misbehaviour based on the history of their activities [9]. In the previous work we did not investigate machine learning models. In this study we focus again on the detection of bully users in online social networks but now we look into the efficiency of *both* expert systems and machine learning models for identifying the potential bully users. We compare the performance of both systems for the task of

adfa, p. 1, 2011.

assigning a score to social network users that indicates their level of bulliness. We demonstrate that the expert system outperforms the machine learner and can be effectively combined in a hybrid classifier. The approach we propose can be used for building monitoring tools to stop potential bullies from conducting further harm.

## 2 Data Collection and Feature Selection

In this section we will explain the characteristics of the corpus used in this study. We also describe the feature space and the three feature categories that have been used for the design of the expert system and for the machine learning models.

### 2.1 Corpus

YouTube is a popular user-generated content video platform. Its audience demographics match those from the general internet population [10]. Unfortunately, bullies misuse the platform to victimize their targets through harassing comments and other misbehaviours. To our knowledge no dataset for cyberbully detection is publicly available and therefore we decided to collect our own. To have a reasonable number and variety of cyberbullies we searched YouTube for topics sensitive to cyberbullying. We determined the users who commented on the top three YouTube videos in each topic. A total of 3,825 users were extracted and we collected a log of the users' activities for a period of 4 months (April – June 2012). We also captured profile information of the users, such as their age and the date they signed up. In total there are 54,050 comments in our dataset. On average there are 15 comments per user (StDev = 10.7, Median = 14). The average age of the users is 24 with 1.5 years of membership duration. Two graduate students were employed to independently annotate the users as bullies or non-bullies (inter-annotator agreement = 93%, Kappa = 0.78) based on their comments and the definition of cyberbullying provided earlier. The users that both annotators had labelled as bullying (n=419) were marked as bullies. Disagreements were resolved by the decision of a third annotator. In total, 12% of the users were labelled as bullies.

### 2.2 Feature space

We compiled a set of 11 features in three categories to be used in our models. The selection of the features was limited by what is technically possible to extract from YouTube. The three categories representing the actions, behaviour, and characteristics of the users are presented in Table 1.

## 3 Methods

In this section we introduce the three types of models used for calculating and assigning the bulliness score to the social network users: a multi-criteria evaluation system, a set of machine learning models and two hybrid models that combine the two.

### 3.1 Multi-Criteria Evaluation System

Multi-Criteria Evaluation Systems (MCES) are a commonly used technique for decision-making in business, industry, and finance [11], but also in fields such as information retrieval [12]. By assigning weights and importance levels to features or criteria, MCES can combine different sources of knowledge to make decisions. In our scenario, a panel of 12 experts in the area of cyberbullying was asked to answer questions about the features presented in Table 1. For each feature they indicated 1) the likelihood that a bully user belongs to a certain category relevant for that feature and 2) the importance of that feature. The likelihood was indicated on four-point scale ‘Unlikely’, ‘Less likely’, ‘Likely’ and ‘Very likely’ corresponding to values 0.125, 0.375, 0.625 and 0.875 respectively [13]. The 'I don't know' option was also available. The importance was indicated on a four-point scale of 1: not informative, 2: partially informative, 3: informative and 4: very informative. Based on the averaged likelihood and importance weights indicated by the experts a “bulliness score” can be calculated for a user by taking the weighted average of the criteria. Multiple experts indicated combined features to be important for determining bullies. Therefore we added two combined criteria to those explained in section 2.2 which were based on age and profanity (C1), and age and misspellings (C2).

**Table 1.** The Feature Space

| | | |
|---|---|---|
| **Content features** model the content of the user comments | | |
| 1 | *Number of profane words in the comments* | Based on a dictionary of 414 profane words including acronyms and abbreviations [6] |
| 2 | *Length of the comments* | Bullying comments are typically short [5] |
| 3 | *First person pronouns* | To detect comments targeted at a specific person |
| 4 | *Second person pronouns* | To detect comments targeted at a specific person |
| 5 | *Usernames containing profanities* | Users with bad intentions might hide their identity |
| 6 | *Non-standard spelling of the words* | Includes misspellings (e.g. 'funy'), or informal short forms of the words (e.g. 'brb') |
| **Activity features** model the activity of the user is in the online environment | | |
| 7 | *Number of uploads* | YouTube users have a public channel, in which their activities such as posted comments and uploaded videos can be viewed. Users can also subscribe to others channels. |
| 8 | *Number of subscriptions* | |
| 9 | *Number of comments* | |
| **User features** model personal and demographic profile information | | |
| 10 | *Age of the user* | Frequency of bullying incidents as well as choice of words and language structures change in different age groups. Divided into 5 age categories: 13-16, 17-19, 20-25, 26-30, and above 30 years old. |
| 11 | *Membership duration of the user* | Divided into 3 groups: less than 1 year, 1- 3 years and above 3 years. |

### 3.2 Machine Learning Approaches

We used three well-known machine learning methods, which use pre-labelled training data for automatic learning: a Naive Bayes classifier, a classifier based on decision trees (C4.5) and Support Vector Machines (SVM) with a linear kernel [14]. The implementation available in WEKA 3 was used [15]. The machine learning models uses the features used by the expert system and, additionally, a number of features that are only interpretable by the machine. These are: (M1) The ratio of capital letters in a comment, to capture shouting in comments; (M2) The number of emoticons used in the comments, to capture explicit emotions; (M3) The occurrence of a second person pronoun followed by a profane word in profanity windows of different sizes (2 to 5 words), to capture targeted harassment; (M4) The term frequency–inverse document frequency (*tfidf*) of frequently repeated words, to capture emphasized content. As the baseline, we trained an SVM classifier ($SVM_B$) using only the content features listed in Table 2.

### 3.3 Hybrid Approach

Two hybrid approaches were tested to combine the advantages of the expert system with the machine learning models. The expert system is not affected by biased training data and may come up with rules that generalize better to unseen data. Machine learning models on the other hand can detect and analyse complex patterns that experts are not capable of observing. We construct two types of hybrid systems:

**H1**: Using the outcome of the expert system as an extra feature for training the machine learning models. The hybrid system is formed by adding the following features to the machine learning classifier: 1) the results of the MCES, 2) the features' categories that were used in the expert system as new set of features, and 3) the combined features (C1 and C2).

**H2**: Using the results of the machine learning model as a new criterion for the expert system. As previously done in the MCES, we assigned equal weights to all the criteria used in the system, including the machine learner criterion.

### 3.4 Evaluation

In this study the output of the models (i.e. bulliness scores) are probability values ranging from 0 to 1, and not a binary class. Therefore we used a threshold independent measure to evaluate and compare the performance of approaches. We evaluated the discrimination capacity by analysing their receiver operation characteristic (ROC) curves. A ROC curve plots "sensitivity" values (true positive fraction) on the y-axis against "1 - specificity" values (false positive fraction) for all thresholds on the x-axis [16]. The area under such a curve (AUC) is a threshold-independent metric and provides a single measure of the performance of the model. AUC scores vary from 0 to 1. AUC values of less than 0.5 indicate discrimination worse than chance; a score of 0.5 implies random predictive discrimination; and score of 1 indicates perfect discrimination. We used 10-fold cross-validation to evaluate the performance of the machine learning classifiers.

## 4 Results and Discussion

Based on weights that were assigned to each feature by the experts, profanities and bullying sensitive topics in the history of a user's comments is the most informative feature (average weight equals 3.6) and the least informative feature is the number of non-standard spellings (average weight equals 1.7). The age range between 13 and 16 years was indicated to most likely to contain bullies. Moreover, it is more likely that bully users have a high ratio of second person pronouns and profane words in their usernames. In the Activity features set, a high number of posting comments has the highest likelihood. The discrimination capacity of the MCES was 0.72.

Among the machine learning classifiers the decision tree classifier performed the worst, followed by the SVM classifier. Naive Bayes with discrimination capacity of 0.66 outperformed the other two algorithms. The contribution of each feature in classification task was determined by excluding it from the feature sets. The number of profane words, second person pronouns and pronoun-profanity windows, were the strongest contributing features. The discrimination capacity also improved by adding users' profile information and history of activities to the training features. The MCES outperformed all machine learning models in terms of discrimination capacity. A possible explanation for these differences is the sensitivity of the machine learning methods for class skew which was quite high in our dataset (10% bullying, 90% non-bullying). On the other hand a method based on human reasoning might not be as sensitive to the training data as the machine learning models and therefore give a better performance. For the first scenario of our hybrid approach (H1), the discrimination capacity of all machine learning methods improved. The improvement for the decision tree was smallest. Although the SVM gained the highest improvement in the discrimination capacity, the Naive Bayes classifier still outperformed the others. All measured improvements for H1 were significant (t-test, P<0.05). Also the hybrid model in the second scenario (H2) showed improvement of the discrimination capacity.

**Table** 2**.** The discrimination capacity of the tested cyberbully detection methods

| | **Approach** | **AUC** | **Hybrid approach** | **AUC** |
|---|---|---|---|---|
| *Baseline* | $\mathrm{SVM}_B$ | 0.57 | | |
| *Machine Learning* | Naive Bayes | 0.66 | + MCES (H1) | **0.72** |
| | Decision Tree | 0.52 | + MCES (H1) | **0.54** |
| | SVM | 0.59 | + MCES (H1) | **0.68** |
| *Expert System* | MCES | 0.72 | + Naive Bayes (H2) | **0.76** |

## 5 Conclusion

In this experiment we developed a multi-criteria evaluation system for identifying potential bully users. We compared this expert system with a variety of machine learning models to assign a 'bulliness' score to YouTube and combined them in two hybrid models. We demonstrated that the expert system outperforms the machine learning models and the hybrid approach results in a further but marginal improvement in prediction performance. The proposed approach is in principle language independent can be adapted to other social networks as well. Spatial features such as location of the users as well as temporal features such as the time of their activities might be useful features to look into. They can provide further information about the pattern of bullying incidents repetitions and common locations and times in which bullying incidents happen.

**Acknowledgments.** The authors gratefully acknowledge the help of the expert panel for sharing their valuable knowledge and experience with responding to the questionnaire. We are thankful to Professor Dr J. Connolly (York University, Canada), Professor Dr F. Mishna (University of Toronto, Canada), Dr S. Drossaert (University of Twente, the Netherlands), J. van der Zwaan (University of Delft, the Netherlands), and seven other experts who prefer anonymity. The foundation of this study was laid during the visit of M. Dadvar to the York University, Canada, hosted by Professor Dr J. Huang, in January 2013.

## References

1. Smith, P.K., et al., *Cyberbullying: its nature and impact in secondary school pupils.* Journal of Child Psychology and Psychiatry, 2008. **49**(4): p. 376-385.
2. Perren, S., et al., *Coping with Cyberbullying:ASystematic Literature Review.* Final Report of the 'COST'IS 0801', 2012.
3. Campbell, M.A., *Cyber bullying: An old problem in a new guise?* Australian Journal of Guidance and Counselling, 2005. **15**(1): p. 68-76.
4. Dinakar, K., R. Reichart, and H. Lieberman, *Modeling the Detection of Textual Cyberbullying.* Social Mobile Web Workshop at International Conference on Weblog and Social Media 2011.
5. Yin, D., et al., *Detection of harassment on Web 2.0.* Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009, Madrid, Spain 2009.
6. Dadvar, M., et al., *Improving cyberbullying detection with user context*, in *Advances in Information Retrieval*. 2013, Springer. p. 693-696.
7. Argamon, S., et al., *Mining the Blogosphere: Age, gender and the varieties of self-expression.* First Monday, 2007. **12**(9).
8. Pazienza, M.T. and A.G. Tudorache, *Interdisciplinary contributions to flame modeling*, in *AI* IA 2011: Artificial Intelligence Around Man and Beyond*. 2011, Springer. p. 213-224.
9. Dadvar, M., F. De Jong, and D. Trieschnigg. *Expert knowledge for automatic detection of bullies in social networks*. in *The 25th Benelux Conference on Artificial Intelligence (BNAIC 2013)*. 2013. Delft.
10. Cha, M., et al. *I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system*. in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 2007, ACM.
11. Figueira, J., S. Greco, and M. Ehrgott, *Multiple criteria decision analysis: state of the art surveys*. Vol. 78. 2005: Springer.
12. Farah, M. and D. Vanderpooten. *A multiple criteria approach for information retrieval*. in *String Processing and Information Retrieval*. 2006: Springer. p. 242-254.
13. Xu, Z., T.M. Khoshgoftaar, and E.B. Allen, *Application of fuzzy expert systems in assessing operational risk of software.* Information and Software Technology, 2003. **45**(7): p. 373-388.
14. Witten, I.H., E. Frank, and M.A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. 2011: Elsevier.
15. Hall, M., et al., *The WEKA data mining software: an update.* ACM SIGKDD Explorations Newsletter, 2009. **11**(1): p. 10-18.
16. Fielding, A.H. and J.F. Bell, *A review of methods for the assessment of prediction errors in conservation presence/absence models.* Environmental conservation, 1997. **24**(1): p. 38-49.