Adopting MaxEnt to Identification of Bullying Incidents in Social Networks

Maral Dadvar
Web-Based Information Systems and Services
Stuttgart Media University
Stuttgart, Germany
dadvar@hdm-stuttgart.de

Aidin Niamir
Data and Modelling Centre
Senckenberg Biodiversity and Climate Research Centre
Frankfurt am Main, Germany
niamir@gmail.com

***Abstract*— Bullying is a widespread problem in cyberspace and social networks. Therefore, in the recent years many studies have been dedicated to cyberbullying. Lack of appropriate dataset, due to variety of reasons, is one of the major obstacles faced in most studies. In this work we suggest that to overcome some of these barriers a model should be employed which is minimally affected by prevalence and small sample size. To this end we adopted the use of the Maximum Entropy method (MaxEnt) to identify the bully users in YouTube. The final results were compared with the commonly used methods. All models provided reasonable prediction of the bullying incidents. MaxEnt models had the highest discrimination capacity of bullying posts and the lowest sensitivity towards prevalence. We demonstrate that MaxEnt can be successfully adopted to cyberbullying studies with imbalanced datasets.**

Keywords*— *Cyberbullying, Maximum Entropy, Prevalence, Sample Size, Sentiment Analsysis, Social Networks, Text Retreival, YouTube

I. INTRODUCTION

Bullying is a widespread problem in cyberspace and social networks. Cyberbullying is 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 of bullying is the posting of hateful comments about someone in social networks. Identification of bullying incidents is one of the main courses of actions to combat such misbehaviour in social networks.

To this end there are several studies which have concentrated on detection of the bullying comments and harassing contents [2]–[5] , as well as identification of the bully users [6], [7]. As we have extensively explained in our previous studies [8], one essential obstacle that is commonly faced in almost all of the cyberbullying studies is lack of a suitable dataset representing cyberbullying in social networks. Imbalance of bullying and non-bullying incidents in the online materials as well as the cumbersome process of labelling the dataset make it even harder to develop the appropriate dataset for these studies. Advances in artificial intelligence along with powerful computational facilities have fuelled a rapid increase in predictive modelling of bullying incidents from massive social network's data. However, low prevalence of these incidents made the labelling process costly and laborious. Commonly used methods for identification of bullying incidents have been criticized for being inherently dependent on prevalence, and have been argued that the low number of bullying incidents introduces statistical artefacts.

In this work we suggest that to overcome the stated barriers a model should be employed which is minimally affected by prevalence and small sample size. For this purpose, we adopted the use of the Maximum Entropy (MaxEnt) method for modelling these incidents in social networks. MaxEnt is a general-purpose machine learning method with a simple and precise mathematical formulation, and it has number of aspects that make it well-suited for studies such as cyberbullying detection in which the target incidents are scarce. In order to evaluate the proposed method, we performed a case study using a manually labelled YouTube dataset. We compiled a set of features to identify bully users representing the personal characteristics of the users, content of their online activities and behaviour of the users, respectively. MaxEnt predictions, solely based on bullying incidents, were compared with those of commonly used modelling methods; Generalized Linear Models, Random Forests, and Support Vector Machine.

II. MAXIMUM ENTROPY (MAXENT)

Maximum Entropy is a statistical learning method. It has been developed and used in other fields, and has been extensively used in modelling the geographical distribution of species, where similar to our case of study, datasets with both observed and not-observed classes are scarce. In this method, the multivariate distribution of incidents (here the bully users) in feature-space is estimated according to the principle of maximum entropy. It states that the best approximation of an unknown distribution is the one with maximum entropy (the most spread out) subject to known constraints. The constraints are defined by the expected value of the distribution, which is estimated from a set of incidents.

Here we used Maxent software package (version 3.3.3; [9]) which is particularly popular in species distribution and environmental niche modelling, with over 2000 applications published since 2006. [9] outlined some advantages and disadvantages of MaxEnt compare to other methods; Maxent only requires incident data, often called incident-only data plus features for the whole datasets. The results are amenable to interpretation of the form of the feature response functions. MaxEnt has properties that make it very robust to limited amount of training data (i.e. small sample size), and is well-regularized [10]. Because it uses an exponential model for probabilities, it can give very large predicted values for conditions that are outside the rage of those found in the data used to develop the model. Nevertheless, extrapolation outside

of the range of values used to develop a model should be done very cautiously no matter what modelling method is used.

III. EXPERIMENTAL SETTINGS

A. Corpus

We used the labelled YouTube dataset provided by [11]. To our knowledge no other comprehensive dataset for cyberbully detection is publicly available. The dataset consist of the activity logs of 3,825 users in the period of 4 months (April – June 2012), along with their profile information, such as their age and the date they signed up. In total there are 54,050 comments in the 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. The dataset has been labelled manually as bullies or non-bullies. In total, 765 users (12% of the users) are labelled as bullies.

B. Feature Space

We compiled a set of fourteen features in three categories to be used in our models (Table 1).

The *activity features* are the activities that users can undertake in the social network. These features help to determine how active the user is in the online environment; for instance uploading videos, posting comments on uploaded videos, or responding to other user's comments. The *user features* are the demographic and personal information of the users, which were publicly available in their profile, such as user's age, or the membership duration of the users. The *content features* are the ones which are extracted from the comments posted by the users and pertain to the writing structure and usage of specific words which represent their writing style and structure. For more details please see [6].

Since correlation among features [12] violates the assumption of independence of most standard statistical procedures [13][14], the compiled features was investigated using the variance inflation factor (VIF) as a measure of multicollinearity.

C. Classification Techniques

We employed three well-known classification methods, namely the generalised linear model [15], random forests [16], and support vector machine [17], along with MaxEnt to identify bully users.

The generalized linear model (GLM) uses a parametric function to link the response variable to a linear, quadratic or cubic combination of explanatory variables. We used an ordinary polynomial GLM with an automatic stepwise model selection based on the Akaike Information Criterion. The random forests (RF) algorithm selects many bootstrap samples from the data and fits a large number of regression trees to each of these subsamples. Each tree is then used to predict those subsamples that were not selected as bootstrap samples. The classification is provided by considering each tree as a 'vote', and the predicted class of an observation is determined by the majority vote among all trees. The models presented here used 1000 trees. The support vector machine (SVM) is a machine-learning generalised linear classifier that estimates the potential bully users that is subject to the feature values by separating the feature space by hyper-planes into bullying and non-bullying feature values. The optimality criterion used to find the separating hyper-plane is the maximised distance to the training data points.

We randomly split the data, 75% of which was used to train the models and the remaining 25% of which was used to evaluate the model performance. All models except MaxEnt were trained using both bully and non-bully labelled data, whereas MaxEnt models were trained using bully-only labelled data. We iterated this step 25 times and calculated the variation and therefore robustness of the models.

TABLE I. THE FEATURE SPACE

Activity features		***VIF****
1	Number of comments	1.30
2	Number of subscriptions	1.02
3	Number of uploads	1.04
User features		
4	Age of the user	1.05
5	Membership duration of the user	1.06
Content features		
6	Number of profane words in the comments	1.02
7	Usernames containing profanities	1.00
8	Length of the comments	2.23
9	First person pronouns	1.62
10	Second person pronouns	1.61
11	Non-standard spelling of the words	1.38
12	Number of smilies in the comments	1.03
13	Number of capital letters in the comments	1.28
14	Second person pronouns followed by profanities	1.07

*. Variance Inflation Factor

D. Evalaution

The outputs of the models (i.e. probability of a user being bully) are values ranging from 0 to 1. We used a threshold independent measure to evaluate and compare the performance of models. 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 [18]. 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 also assessed the goodnees-of-fit [19] of the models using Miller's calibration statistic [20], [21]. Miller's calibration statistic evaluates the ability of a prediction model to correctly predict the proportion of bully users with a given feature profile. It is based on the hypothesis that the calibration

line – perfect calibration – has an intercept of zero and a slope of one. The calibration plot shows the model's estimated probability (x-axis) against the mean observed proportion of positive cases (y-axis) for equally sized probability intervals (number of intervals = 10).

IV. Results and Discussions

All models provided reasonable prediction of the bullying incidents and were significantly ($P < 0.001$ in all four models) better than random in both binomial tests of omission and receiver operating characteristic (ROC) analyses (Table 2). The area under the ROC curve was always higher for MaxEnt, indicating stronger discrimination power of bullying users. Variation in the performance of MaxEnt was as small as the other models (Figure 1).

MaxEnt and RF models were better calibrated compared to the GLM and SVM models, meaning that given feature profile, they accurately predict the proportion of bully users to the whole dataset (Figure 2). Better-calibrated models are of greater interest if the objective lies in independent training of the model, and then transferring the model and producing a general conclusion beyond the training extent over which the models are fitted.

Analysis of the feature's contribution to the MaxEnt models revealed that the number of profane words in the comments has the highest contribution (~ 33%), followed by the number of the comments (Figure 3). Although all the features significantly contributed to the models ($P < 0.01$ in all fourteen features), number of subscription had the least contribution to the models (~ 1%).

TABLE II. The Discrimiantion Capacity of the Models

Model	***AUC****	***AUC.sd***
Maximum Entropy (MaxEnt)	0.75	0.032
Generalized Linear Model (GLM)	0.64	0.034
Random Forests (RF)	0.69	0.032
Support Vector Machine (SVM)	0.59	0.032

*. Area under the receiver operating characteristic (ROC) curve

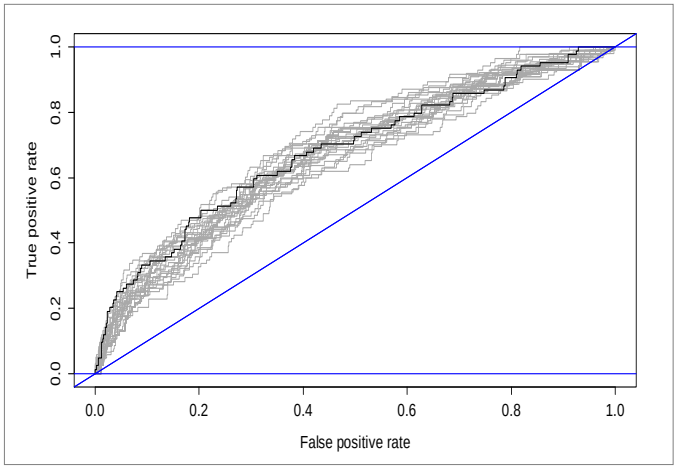

Figure 1. ROC plot of MaxEnt Models (n=25 iterations)

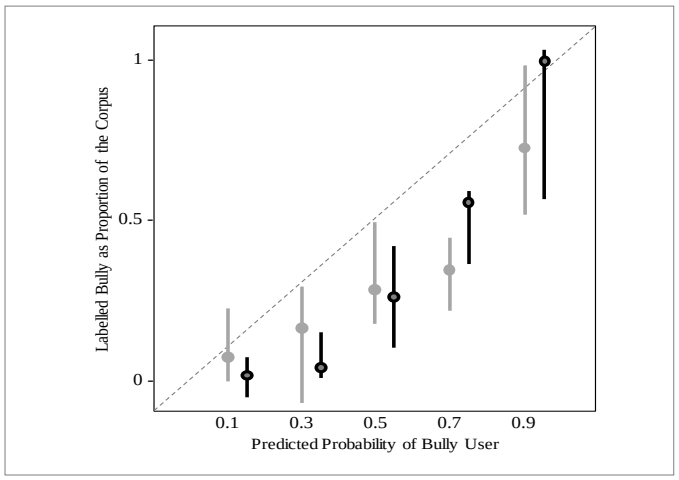

Figure 2. Calibration plot for MaxEnt (light grey) and RF (dark grey)

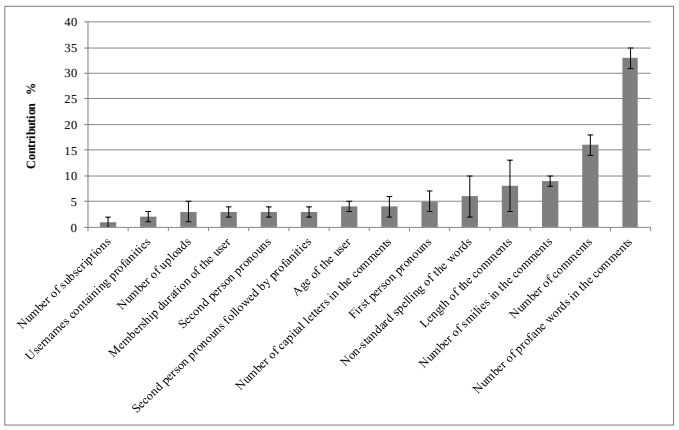

Figure 3. Retative contribution of feature variables to the MaxEnt Model

V. Conclusion

In this experiment we adopted MaxEnt for identification of potential bullying users. We compared the MaxEnt with a variety of common models to calculate the probability of a user being bully, given the features profile. We demonstrated that the MaxEnt outperforms the other models in discrimination capacity and also provide well-calibrated models that are reliably transferable beyond the training extent over which the models are fitted. The proposed approach is in principle language independent and 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. We recommend using MaxEnt as an incident-only approach in cyberbullying studies with imbalanced datasets or rare number of target incidents.

References

[1] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: its nature and impact in secondary school pupils.," *J. Child Psychol. Psychiatry.*, vol. 49, no. 4, pp. 376–85, Apr. 2008.

[2] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the Detection of Textual Cyberbullying,” *Assoc. Adv. Artif. Intell.*, pp. 11–17, 2011.

[3] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, “Detection of Harassment on Web 2.0,” *Proc. Content Anal. WEB 2.0 Work. WWW2009*, ., 2009.

[4] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, “Improving cyberbullying detection with user context,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7814 LNCS, pp. 693–696, 2013.

[5] V. Nahar, X. Li, and C. Pang, “An Effective Approach for Cyberbullying Detection,” *ADC*, pp. 160–171, 2014.

[6] M. Dadvar, D. Trieschnigg, and F. de Jong, “Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies,” Springer International Publishing, 2014, pp. 275–281.

[7] M. Dadvar, D. Trieschnigg, and F. de Jong, “Expert knowledge for automatic detection of bullies in social networks,” pp. 57–64, 2013.

[8] M. Dadvar, “Experts and machines united against cyberbullying,” University of Twente, Enschede, The Netherlands, 2014.

[9] S. J. Phillips, R. P. Anderson, and R. E. Schapire, “Maximum entropy modeling of species geographic distributions,” *Ecol. Modell.*, vol. 190, no. 3–4, pp. 231–259, Jan. 2006.

[10] S. J. Phillips and M. Dudík, “Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation,” *Ecography (Cop.).*, vol. 31, no. 2, pp. 161–175, Apr. 2008.

[11] M. Dadvar, D. Trieschnigg, and F. de Jong, “Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies,” in *Advances in Artificial Intelligence*, Springer International Publishing, 2014, pp. 275–281.

[12] D. C. Montgomery and E. A. Peck, *introduction to linear regression analysis*. New York, New York, USA: John Wiley and Sons, 1982.

[13] P. Legendre, “Spatial Autocorrelation: Trouble or New Paradim,” *Ecology*, no. 74, pp. 1659–1673, 1993.

[14] A. Niamir, A. K. Skidmore, A. G. Toxopeus, A. R. Munoz, and R. Real, “Finessing atlas data for species distribution models,” *Divers. Distrib.*, vol. 17, no. 6, pp. 1173–1185, 2011.

[15] P. ; N. McCullagh J. A., *Generalized Linear Models*, vol. 135, no. 3. London: Chapman and Hall, 1989.

[16] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[17] S. P. Maher, C. F. Randin, A. Guisan, and J. M. Drake, “Pattern-recognition ecological niche models fit to presence-only and presence-absence data,” *Methods Ecol. Evol.*, vol. 5, no. 8, pp. 761–770, 2014.

[18] A. H. Fielding and J. F. Bell, “A review of methods for the assessment of prediction errors in conservation presence/absence models,” *Environ. Conserv.*, vol. 24, no. 01, Mar. 1997.

[19] S. Lemeshow and D. W. Hosmer, “A review of goodness of fit statistics for use in the development of logistic regression models.,” *Am. J. Epidemiol.*, vol. 115, no. 1, pp. 92–106, 1982.

[20] M. E. Miller, S. L. Hui, and W. M. Tierney, “Validation techniques for logistic regression models.,” *Stat. Med.*, vol. 10, no. 8, pp. 1213–1226, 1991.

[21] J. Pearce and S. Ferrier, “An evaluation of alternative algorithms for fitting species distribution models using logistic regression,” *Ecol. Modell.*, vol. 128, no. 2–3, pp. 127–147, 2000.