Detecting Aggressors and Bullies on Twitter

Despoina Chatzakou†, Nicolas Kourtellis‡, Jeremy Blackburn‡
Emiliano De Cristofaro♯, Gianluca Stringhini♯, Athena Vakali†

†Aristotle University of Thessaloniki ‡Telefonica Research ♯University College London
deppych@csd.auth.gr, nicolas.kourtellis@telefonica.com, jeremyb@tid.es
e.decristofaro@ucl.ac.uk, g.stringhini@ucl.ac.uk, avakali@csd.auth.gr

ABSTRACT

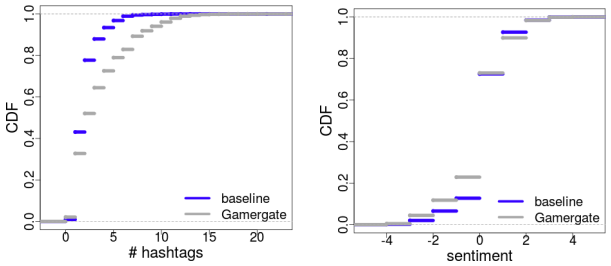
Online social networks constitute an integral part of people's every day social activity and the existence of aggressive and bullying phenomena in such spaces is inevitable. In this work, we analyze user behavior on Twitter in an effort to detect cyberbullies and cyber-aggressors by considering specific attributes of their online activity using machine learning classifiers.

1. INTRODUCTION

Aggression can be a one-time action where someone purposely says or does something to hurt someone. Bullying is a repeated and intentional negative behavior (or aggression) of a group or an individual that can appear in many ways, e.g., threats, rumors, or verbal attacks targeting one or a group of individuals. Cyber -bullying and -aggression are the digital manifestations of bullying and aggression, respectively. Incidents of such behaviors are regularly reported on social media, especially among teenagers whose engagement with online networks is rapidly increasing. In fact, in 2014, over 50% of young people who use social media have reported being cyberbullied[1], and racist and sexist attacks have been also been reported on Twitter.[2] The research community has recently focused on detecting bully and aggressive behavior across various social platforms, e.g., Instagram [7] and Yahoo Finance [4]. All such works build upon either textual, structural or visual attributes to distinguish abusive and/or bullying content. Few works have focused on characterizing the bullying users themselves and not only their abusive content, e.g., [3]. In this work, we explore the characteristics of Twitter users with respect to their content and network embeddedness, and leverage such attributes with a machine learning classifier to automatically detect Twitter aggressors and bullies. The results indicate that we can distinguish between aggressive, bully and normal users with 87.8% precision and 90.1% recall.

2. METHODOLOGY

To study the problem of cyberbullying and aggression on Twitter, we collected a set of tweets and created a ground truth of labeled users from their posts. Furthermore, we extracted various features characterizing these users. Finally, we trained machine learning classifiers to automatically detect bullying and aggressive behavior of Twitter users. Next, we explain briefly each of these steps.



(a) Hashtags distribution. **(b)** Sentiment distribution.

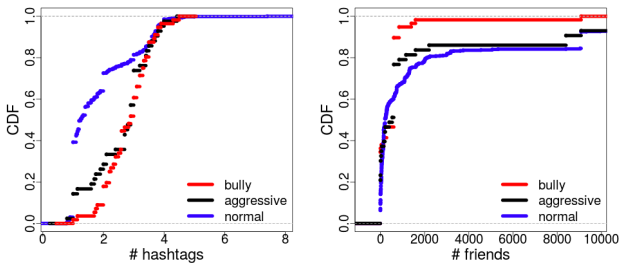
Figure 1: CDF of number of users' hashtags and expressed sentiment.

Dataset collection. During June to August 2016 we collected from the Twitter Streaming API two sets of tweets: (i) a *baseline* of 1M random tweets, and (ii) a *hate-related* set of 650k tweets based on 309 hashtags associated with bullying and hateful speech. To create the list of 309 hashtags, at first we parsed all the tweets collected during the aforementioned period to select those containing #GamerGate, as the Gamergate controversy [9] is one of the most well documented large-scale instances of bullying/aggressive behavior. Then, the list was completed by considering hashtags that coexisted within the tweets with the #GamerGate. After a manual inspection we saw that multiple hashtags contained hateful words, e.g., #IStandWithHateSpeech, #KillAllNiggers, and #InternationalOffendAFeministDay. The random set of tweets served as a baseline, as it is less prone to contain any kind of offensive behaviors. Figures 1a and 1b show that there are substantial differences in the tweeting activity among the two groups of users, when comparing the number of the used hashtags and the expressed sentiment. Overall, hate-related users tend to use more hashtags than baseline users, which could be because they use Twitter as a rebroadcasting mechanism aiming at attracting attention on the topic. They also express with greater negativity which aligns with the fact that the #GamerGate tweets contain a large proportion of offensive posts.

Ground truth (GT). To build a ground truth dataset, i.e., a dataset where the users are characterized either as bullies, aggressors or normal, we proceeded with the crowdflower.com platform to recruit human workers to complete the labeling task. The recruiters were redirected to an online survey tool we developed and were asked to label 10 sets of tweets each. Each set contained 5-10 tweets of the same user (preserving the chronological ordering of their posted time) so that we detect bullying behaviors which involve repetitive actions over time, in addition to the aggressive and normal ones.

[1] https://www.stopbullying.gov/media/facts/index.html
[2] https://www.theguardian.com/technology/2016/oct/18/did-trolls-cost-twitter-35bn

©2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License.
WWW 2017 Companion April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054211



(a) Hashtags distribution. **(b)** Friends distribution.

Figure 2: CDF distribution for Hashtags and Friends.

As Twitter contains a non-negligible amount of spammers [2], i.e., users posting unsolicited content, in addition to the previous referred labels, workers could also characterize a user as spammer. In total, 1,500 sets of tweets were used in the annotation process maintaining the same number of sets for both the hate-related and random tweets. We recruited 834 workers, whom we allowed to participate only once (to eliminate behavioral bias across tasks and discourage rushed tasks). Each set was annotated by 5 different workers, while in the end the majority vote was used to create the final annotation labels excluding the sets of tweets where the majority could not be determined. Overall, we concluded to 1,307 sets (containing in total 9,484 tweets) where 4.5% corresponded to bully users, 3.4% to aggressors, 31.8% to spammers and 60.3% to normal ones. Sets labeled as spammers were excluded from the analysis presented next. We assume that such users and their posts can already be filtered out using advanced spam removal processes applied specifically on Twitter, e.g., [2, 6]. Finally, we also studied the inter-rater agreement using the Fleiss' kappa measure [5] which assesses the reliability of agreement between a fixed number of raters. The overall *Fleiss' kappa* value equals to 21.89% which can be characterized as a fair agreement between our workers [8]. We detail this process in our extended paper [1].

Preprocessing. To reduce noise from the dataset and before extracting any features, we followed typical clean up processes on text sources, i.e., removal of stop words, URLs, and punctuations, as well as normalization, i.e., removal of repetitive characters which show feelings with intensity.

Feature Extraction. We extracted various features, either user-based (e.g., #posts, #days since account creation), text-based (e.g., #hashtags, sentiment), or network-based (e.g., #followers, #friends) and tested them to select the best performing. Indicatively, Figures 2a and 2b plot the CDF of the average number of hashtags and number of friends for the 3 different user classes, respectively. From Figure 2a we observe that aggressors and bullies have a propensity to use fewer hashtags within their tweets, while Figure 2b indicates that bullies have fewer friends than the other categories, which is quite useful in distinguishing aggressive users from the other two classes. Based on the information gain computed for each feature, the user-based features, followed by the textual ones, contribute more in the experimental setup.

Machine Learning Classifiers. We experimented with more than 15 machine learning algorithms, such as probabilistic (e.g., simple Naive Bayes or networks), tree-based (e.g., decision tree and random forest), or ensemble classifiers, in an effort to distinguish bully and aggressive users from the normal ones. Considering both the time for training each classifier and the classification performance, here, we present the best results obtained with the Random Forest classifier (constructs an ensemble of decision trees with random subsets of features during the classification process).

Table 1: Classification with Random Forest.

(a) Results

	Prec.	**Rec.**	**ROC**
bully	0.464	0.448	0.918
aggressive	0.286	0.093	0.868
normal	0.941	0.978	0.925
Avg.	0.878	0.901	0.922

(b) Confusion matrix

bully	aggres.	normal	
26	7	25	bully (GT)
16	4	23	aggres. (GT)
14	3	770	normal (GT)

Classification Results. Table 1 shows the results obtained with a 10-fold cross validation process. Overall, the average precision and recall is 87.8% and 90.1%, respectively, while the weighted AUC of 92.2% shows that our features and classification technique can perform quite well at detecting bullies and aggressors and distinguishing them from the typical Twitter users. Based on the confusion matrix (Table 1b), the misclassifications in the bully case mostly fall in the normal class. Concerning the aggressive case, we observe a higher "confusion" which indicates that the boundaries among the three classes are not clear and more work is needed along this line.

3. CONCLUSION

On a daily basis, various cases are documented where the content of (a set of) posts on online social platforms is harsh, mean, or even cruel. In this work, we study Twitter bullies and aggressors, two types of users who require special attention from the research community and tech industry, due to the explosion of such behavior manifesting daily in online social communities. Detecting the warning signs of cyberbullying poses several difficulties, as by definition, bullying is often a covert behavior through superficial comments and criticisms. In this paper, a method which builds upon different types of features was tested to distinguish among bullies, aggressors and typical Twitter users (a more detailed description can be found in [1]). The results show our methodology is promising in detecting aggressive and bully users with high accuracy.

4. ACKNOWLEDGEMENT

This research has been fully funded by the European Commission as part of the ENCASE project (H2020-MSCA-RISE of the European Union under GA number 691025).

5. REFERENCES

[1] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali. Mean Birds: Detecting Aggression and Bullying on Twitter. *goo.gl/MSJIw7*, 2017.

[2] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou. 6 million spam tweets: A large ground truth for timely Twitter spam detection. In *ICC IEEE*, 2015.

[3] M. Dadvar, D. Trieschnigg, and F. de Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *CAIAC*, pages 275–281, 2014.

[4] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate Speech Detection with Comment Embeddings. In *WWW*, 2015.

[5] J. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 1971.

[6] M. Giatsoglou, D. Chatzakou, N. Shah, A. Beutel, C. Faloutsos, and A. Vakali. Nd-sync:detecting synchronized fraud activities. In *PAKDD*, pages 201–214, 2015.

[7] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *SocInfo*, 2015.

[8] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 1977.

[9] A. Massanari. #gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 2015.