BullyBlocker

An Interdisciplinary Investigation of Temporal Aspects of Cyberbullying

Linle Jiang, Ashley Trow, Victoria Delgadillo, Carmen Sanchez, Lu Cheng, Yasin Silva, Deborah Hall;

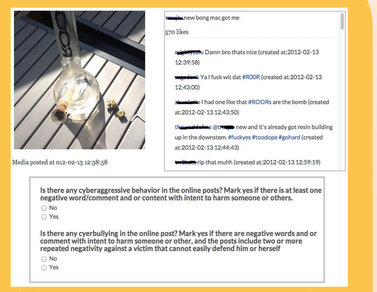
Arizona State University

Motivation

Cyberbullying has become a widespread phenomenon among adolescents in part due to rapid increases in their social media usage. Studies on the negative consequences of cyberbullying underscore the importance of tools for detecting cyberbullying instances. However, relatively little is known about the temporal nature of cyberbullying messages on social media and how the frequency and timing of these messages relate to the identification and perceptions of the severity of cyberbullying. In an exploratory study, we examined temporal aspects of cyberbullying messages for a set of Instagram users.

Data & Analytic Strategy

Our initial dataset, obtained from Hosseinmardi and colleagues (University of Colorado), consisted of 2,218 Instagram posts from users with public profiles. Each post contained at least 15 comments, resulting in a total of 157,147 comments in the full dataset. In previous research (Hosseinmardi et al., 2015), five human coders independently assessed whether each post—when examined holistically with all subsequent comments—constituted cyberbullying or not. Approximately 20% of posts were identified as containing cyberbullying information with over 80% agreement among the five coders.



Whether each individual comment contained cyberbullying information, however, was not classified in the original study. To address this, we employed a bullying trace classifier (Xu et al., 2012) to identify cyberbullying at the level of each individual comment. That is, whereas Hosseinmardi et al. (2015) generated their dataset by manual (human) coding at the 'post level', we added an automated method to identify cyberbullying at the 'comment level'. Only comments classified as cyberbullying with a value larger than .671 were identified as a cyberbullying message in the present study. This criterion identified 31,023 cyberbullying comments, which represented approximately 20% of all available comments.

Finally, we identified 65 posts with less than two cyberbullying comments, which we excluded from the present study given our interest in temporal aspects of cyberbullying. Moreover, due to recording error in the date of some of the posts (i.e., the date of the first cyberbullying comment precedes the first recorded comment in the dataset), an additional 37 posts were excluded.

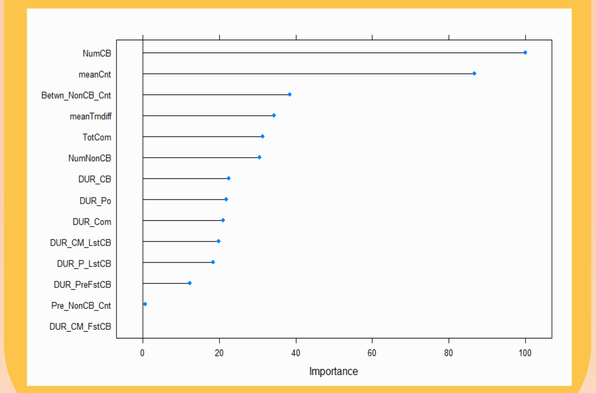
Cyberbullying (CB) Prediction Factors

Variables	Mean	*SD*	Min	Max
NumCB: # of CB Comments	14.42	14.63	1.00	101.00
NumNonCB: # of Non-CB Comments	57.61	41.40	2.00	147.00
TotCom: # of Total Comments	72.03	49.10	8.00	156.00
Pre_NonCB_Cnt: # of Non-CB Comments before the 1st CB Comment	7.25	11.04	0.00	127.00
Betwn_NonCB_Cnt: # of Non-CB Comments between the 1st & last CB Comments	64.79	48.54	1.00	153.00
DUR_Po: Post Duration	271601.67	303283.90	3.00	1719043.00
DUR_Com: Comment Duration	262413.24	294851.90	2.00	1719042.00
DUR_P_LstCB: Post to the Last CB Comment Duration	131449.97	228365.80	0.00	1498118.00
DUR_PreFstCB: Post to the 1st CB Comment Duration	13286.62	66501.72	0.00	995710.00
DUR_CM_FstCB: the 1st Comment to the 1st CB Comment Duration	4098.18	33559.89	0.00	738776.00
DUR_CB: CB Comments Duration	118163.38	208825.80	0.00	1462426.00
DUR_CM_LstCB: the 1st Comment to the Last CB Comment Duration	122261.55	214128.80	0.00	1498111.00
meanTmdiff: Average Duration between Individual CB Comment	15350.98	47440.19	0.00	930392.00
meanCnt: Average Non-CB Comments between Individual CB Comments	5.67	8.50	0.00	129.00

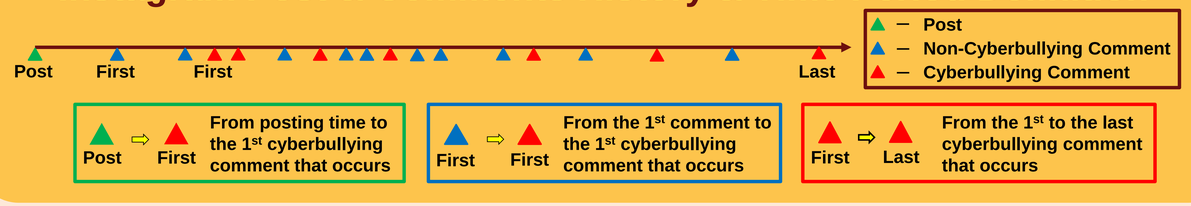
Relative Importance of CB Prediction Factors

We used a *Random Forest* analysis to evaluate the relative importance of a range of temporal factors in predicting cyberbullying identification.

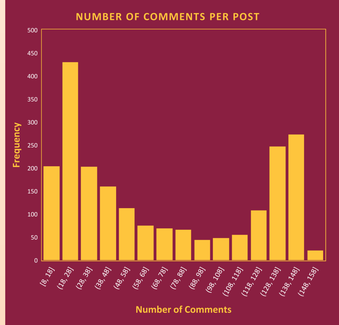
The model was trained and tested using 10-fold cross-validation method for 5 times. The optimal prediction model was achieved when *mtry* (i.e., number of variables randomly sampled at each split) was 1 and *ntree* (i.e., number of trees to be grown) was 2,000, with an accuracy of identifying cyberbullying and non-cyberbullying posts at 84% and *kappa* = .39.



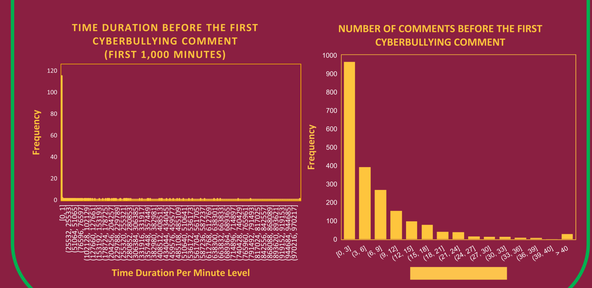
Instagram Post & Comments History & Time Period Definition



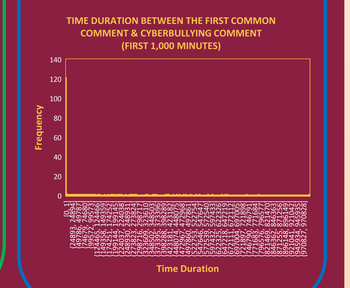
Descriptive Statistics



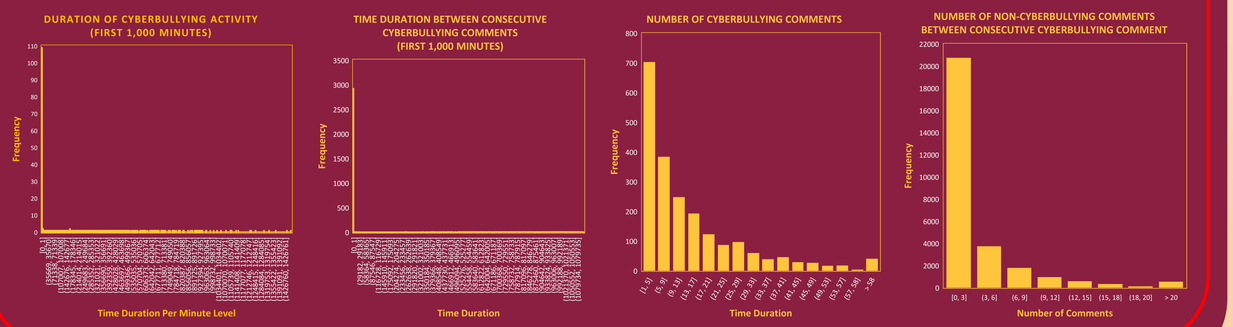
Post ⇨ First



First ⇨ First



First ⇨ Last



References

- Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015, December). *Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network*. In International Conference on Social Informatics (pp. 49-66). Springer, Cham.
- Xu, J. M., Jun, K. S., Zhu, X., & Bellmore, A. (2012, June). *Learning from Bullying Traces in Social Media*. In Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 656-666). Association for Computational Linguistics.
- Silva, Y. N., Rich, C., Chon, J., & Tsosie, L. M. (2016, August). *BullyBlocker: An App to Identify Cyberbullying in Facebook*. In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (pp. 1401-1405). IEEE.
- Silva, Y. N., Hall, D. L., & Rich, C. (2018). BullyBlocker: Toward an Interdisciplinary Approach to Identify Cyberbullying. *Social Network Analysis and Mining*, 8(1), 18.

http://bullyblocker.asu.edu

Download on the App Store https://itunes.apple.com/us/app/bullyblocker-app/id1236410370?mt=8