WISP 2019 Proceedings

Pre-ICIS Workshop on Information Security and Privacy (SIGSEC)

12-15-2019

Charged language on Twitter: A predictive model of cyberbullying to prevent victimization

Shuyuan Mary Ho

Dayu Kao

Ming-Jung Chiu-Huang

Wenyi Li

Chung-Jui Lai

See next page for additional authors

Authors
Shuyuan Mary Ho, Dayu Kao, Ming-Jung Chiu-Huang, Wenyi Li, Chung-Jui Lai, and Bismark Ankamah

Charged Language on Twitter: A Predictive Model of Cyberbullying to Prevent Victimization

Shuyuan Mary Ho[1]
School of Information,
Florida State University,
Tallahassee, FL, U.S.A.

Dayu Kao
Information Management,
Central Police University,
Taoyuan, Taiwan

Ming-Jung Chiu-Huang
Information Management,
Central Police University,
Taoyuan, Taiwan

Wenyi Li
College of Education,
Florida State University,
Tallahassee, FL, U.S.A.

Chung-Jui Lai
Information Management,
Central Police University,
Taoyuan, Taiwan

Bismark Ankamah
School of Information,
Florida State University,
Tallahassee, FL, U.S.A.

ABSTRACT

Cyberbullying is not a crime, but has significant potential to harm victims' mental health in the online world as enabled by information and communication technology (ICT). This research in progress aims to derive a predictive mechanism that can protect potential victims from abuse and harm by cyberbullying. The study is based on the collection and processing of 140,000 tweets, and uses a logistic regression model to predict a tendency for cyberbullying based on the manifestation of emotionally charged language on Twitter. Our findings show high potency and statistical significance in the identification of charged language that has the potential to victimize others. The study contributes to a preventative confirmation of cyberbullying, in an effort to provide early warnings for parental mediation and/or mitigation agencies, including school counselors, online bystanders, and law enforcement agencies.

Keywords: Cyberbullying, charged language, language-action cues, routine activity theory, logistic regression analysis, predictive analytics, text mining, social media, Twitter.

[1] Corresponding author. smho@fsu.edu. +1 850 645 0406

INTRODUCTION

Information and communication technology (ICT) has emerged as a force for building communities, and has transformed the fabric of our social interaction. Twitter, a prominent social networking platform, enables people to express their thought and opinions by "tweets" without geographical borders. It creates global online communities where people share thoughts and ideas. Twitter, on average, generates 500 million tweets each day (Al-garadi, Varathan, et al. 2016; Kavanaugh, Fox, et al. 2012), and is estimated to have over 275 million users in 2020 (Clement 2019). This massive development of Twitter poses risks to its users, and sometimes turns Twitter into a "cyberbullying playground" (Xu, Kwang-Sung, et al. 2012). Malicious comments can inflict social and mental wounds on victims with low self-esteem and low self-confidence. Victims' privacy and personal relationships can also be threatened. Although cyberbullying is not a crime, the humiliation of online attacks wreaks havoc on many that are unprepared for online attack, and has caused some victims to attempt suicide (Hinduja and Patchin 2010). This highlights the ability of cyberbullying to negatively impact our youth and those with mental issues. The Cyberbullying Research Center[2] has pointed out that based on a survey of nationally representative samples of 4,972 middle and high school students in the United States, cyberbully victimization rates have drastically increased from 18.8 percent in 2007 to 36.5 percent in 2019 (Patchin and Hinduja 2019).

To make matters worse, the anonymity of ICT-enabled communication creates an opportunity for the phenomenon of cyberbullying to run rampant (Ong 2015). Bullies are now able to create anonymous accounts on Twitter. Such anonymous communication enables bullies to threaten victims without exposing their real identities. The anonymity enabled by Twitter

[2] https://cyberbullying.org/summary-of-our-cyberbullying-research

creates a serious power-imbalance situation that facilitates widespread cyberbullying (Ong 2015).

By definition, cyberbullies are those who abuse victims in public and private forums. Thus, we study the charged language of cyberbullies in an effort to create a predictive model for identifying cyberbullying based on the language-action cues found in tweets (Ho and Hancock 2018, 2019; Ho, Hancock, et al. 2016a; Ho, Hancock, et al. 2016b; Ho, Hancock, et al. 2016c; Ho, Hancock, et al. 2015; Ho, Liu, et al. 2016). This paper describes this research in progress, which is outlined as follows. In the second section, we define cyberbullying, and discuss relevant literature to frame the cyberbullying phenomenon. In the third section, we introduce our study framework, and the research methods used to collect and classify data. In the fourth section, we discuss the data analysis process, as well as the results and findings in detail. Last but not least, we discuss study limitations, and we conclude with contributions and future work.

CONCEPTUALIZATION

To conceptualize cyberbullying, we first establish the terminology of a cyberbully, and then explore the interacting factors of influence that define a cyberbully by reviewing routine activity theory.

Definition

Cyberbullying can be seen as the online equivalent to face-to-face bullying. Belsey (2007) built a cyberbully website[3], and defined the term “cyberbullying” as “ICT users who utilize instant messages, personal websites, etc. to support repeated, hostile and mean behavior initiated by individuals or groups with ill-intention to harm others.” Smith, Mahdavi, et al. (2008) defined “cyberbullying’’ as an aggressive act or behavior carried out using electronic methods by a group or an individual repeatedly and overtime against a victim who cannot easily

[3] Cyberbully website: http://www.billbelsey.com/?cat=13

defend him or herself. Hinduja and Patchin (2010) depicted “cyberbullying” as “willful and repeated harm inflicted using computers, cell phones, and other electronic devices.’’ Langos (2012) also pointed out that the definition of cyberbullying was constructed based on four elements: repetition, power imbalance, aggression, and occurrence on the cyberspace. They also pointed out that “repetition” is a critical element that differentiates jokes and teasing from bullying. Purposeful cyberbullying is not a short-term process. It is a long-term process that results in real pain for the victims.

Li (2007) conducted a survey on cyberbullying among 177 middle- school students, and found out that generally 54% of students had experienced bullying in the physical space, and over 25% of them had been victims in cyberspace. Li (2007) also discovered that both victims and bystanders tend to be quiet in response. One-third of the victims falsely believed that an adult would not stop cyberbullying even though they knew it was causing harm. Li (2007) actively promoted the establishment of a greater trust relationship between students and adults (i.e., school staff) as a way of mitigating the impact of cyberbullying. As social media becomes popular in students’ communication, the problem of cyberbullying also has increased drastically in these online venues.

Routine Activity Theory

Although the laws around using threatening and charged language are limited, Cohen and Felson (1979) proposed a routine activity approach—with an understanding of the social structure and social change—to explain the reasons and the trends of these potentially illegal activities. Three critical elements—motivated offenders, suitable targets, and the absence of capable guardians against a violation—were identified as being essential factors in explaining these activities. This theoretical lens describes the triangulated interactions between motivated offenders, suitable targets, and the absence of capable guardians. Mainly, cyberbully motivation and behavior can be expressed through language, the potential offenders employ online. Through understanding the charged language used by potential offenders, the underlying reasons for bullies to behave in certain ways can be understood. Mesch (2009) conducted a survey study and proposed two logistic regression prediction models—restrictive vs. evaluative parental mediation—to understand the influential factors around cyberbullying, so as to predict the likelihood of certain individuals being bullied. The study identified that restrictions and rules about adolescents' website usage would likely reduce the risks and exposure of victims to online bullying. Moreover, adolescents who have not developed a mature conscience tend to express their opinion without discrimination, and this increases the risk of cyberbullying. The importance of parental mediation is emphasized. While parental mediation matters, the victims' personality and their social network behaviors also matter. Peluchette, Karl, et al. (2015) pointed out that individuals who are more extroverted and open are more likely to become a target of cyberbullying because of their tendency for self-disclosure. Ang (2015) also adopted routine activity theory to raise awareness on the importance of the parent-adolescent relationships in addressing this issue. The closer parents and adolescents are, the less chance that adolescents will be bullied. The study further suggested prevention and intervention strategies to thwart adolescents from being victimized by cyberbullying.

STUDY FRAMEWORK

Being able to identify the probability of charged language, as reflected by language-actions cues in tweets, will allow us to gauge the tendency of a person to become a cyberbully. Our study framework is thus designed and illustrated in Figure 1. Based on routine activity theory, this framework describes two phases of a study that seeks to derive preventative confirmation of cyberbullying tendencies so as to protect potential victims. Phase I illustrates the logistic regression predictive modeling based on language-action cues. Phase II illustrates future work that aims to identify motivated cyberbullies based on the analysis of victims' social networks. This paper describes the Phase I study of the framework.

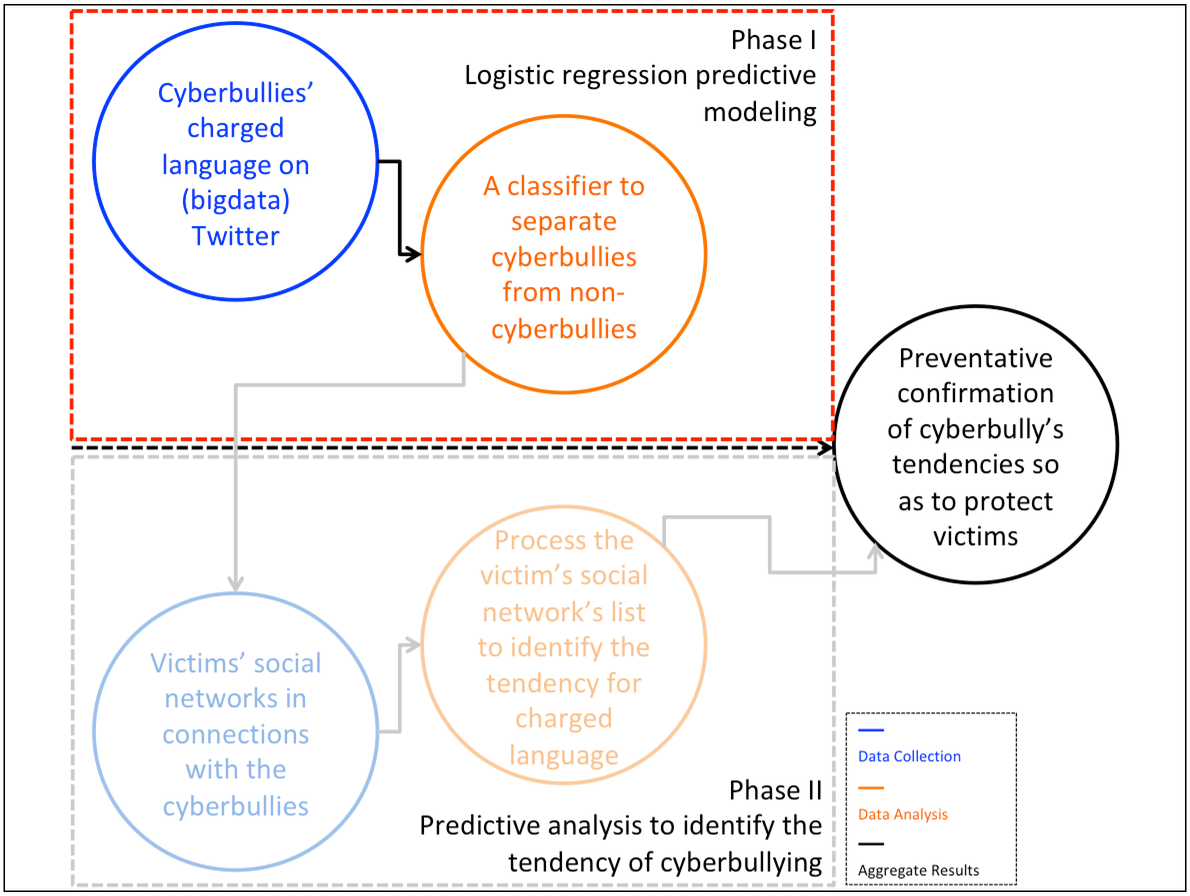
Phase I
Logistic regression predictive modeling
Cyberbullies' charged language on (bigdata) Twitter
A classifier to separate cyberbullies from non-cyberbullies
Preventative confirmation of cyberbully's tendencies so as to protect victims
Process the victim's social network's list to identify the tendency for charged language
Victims' social networks in connections with the cyberbullies
Phase II
Predictive analysis to identify the tendency of cyberbullying
Data Collection
Data Analysis
Aggregate Results

Figure 1. Research design inspired by the routine activity theory

This framework first collects a large dataset that contains the manifestation of charged language from Twitter networks. Then, a classifier is built to differentiate and separate the tendency of cyberbullying from non-cyberbullying. Third, a predictive model based on logistic regression analysis will compare the collected dataset with the existing dataset in Kasture's (2015) study.

Data Collection

Twitter sets limitations for search API, a Python program, which allows for searches only over the past seven (7) days at random. We set the API to collect 10,000 tweets for each of the 14 emotionally charged words—with a total of 140,000 tweets collected between August 31, 2019 and September 6, 2019. Tweets were collected based on swear words suggested in Nand, Perera, & Kasture's (2016) study. The finalized dataset contains the following 14 emotionally charged words: die, faggot, fat, fuck, kill, loser, shit, slut, suck, whore, bitch, cunt, dick, and pussy.

The API returns each tweet in JSOM format, which contains metadata of the tweets (e.g., created time of tweets, username, followers' number, the text of tweets, etc.). The JSOM format of tweets was then converted into Excel spreadsheets (.xlsx) for linguistic analysis using the following steps.

Data Cleaning and Filtering

The original collected dataset contained duplicates and unformatted tweets and retweets. We eliminated these duplicates, resulting in the final dataset of 56,607 tweets in total.

SENTIMENT ANALYSIS

Kasture (2015) conducted a study on predictive modeling to detect cyberbullying on Twitter. Kasture (2015) collected 1,313 tweets and used human judgment to tag and separate these tweets (Table 1). A total of 427 tweets were classified as cyberbullying tweets in the first round, and the inter-rater reliability between two human annotators is 0.833 (Nand, Perera, et al. 2016, p. 700). In addition, these two human-annotators agreed on 367 tweets that were true positive (i.e., cyberbully). Every conversational tweet was classified by LIWC2007 to convert language into numerical values. Then, the Random Forest classifier was adopted to classify the resulting dataset with 97% accuracy for 367 true positive tweets by splitting the 1,313 tweets into a trained and tested dataset. We thus use the dataset classified by Kasture (2015) as the baseline dataset for our data in order to differentiate cyberbullying from non-cyberbullying.

Table 1. Baseline Data from Kasture (2015)

Cyberbullying/ Non-cyberbullying	**Dataset (the number of tweets)**	**Percentage**
cyberbullying	376	28.64%
Non-cyberbullying	937	71.36%

Our dataset was processed by Linguistic Inquiry and Word Count (LIWC2015) to extract linguistic features from texts, and to compute word frequencies into 90 different types of psychology items using its default dictionary. We also reprocessed Kasture (2015) dataset using LIWC2015 so that it can be compared with our dataset.

LOGISTIC REGRESSION ANALYSIS

We used R-studio to run a logistic regression analysis on Nand, Perera, and Kasture's (2016) dataset using ten (10) linguistic features in LIWC suggested in their study. Kasture (2015) had already categorized the dataset into positive (cyberbullying) and negative (non-cyberbullying) using human judges. We thus used these positive (coded as 1) and negative (coded as 2) as our dependent variable.

Table 2. Logistic regression predictive model using 10 LIWC linguistic features for detecting cyberbullying on Twitter

Variables	Coef. Estimate	Std. Error	Z-value
Intercept	-4.270	.321	-13.308***
You	.007	.025	.283
Negative emotion	.127	.030	4.287***
Anger	.006	.041	.150
Biology	-.235	.065	-3.633***
Body	.281	.054	5.166***
Health	.296	.071	4.179***
Sexual	.428	.068	6.308***
Ingestion	.248	.077	3.215***
Death	.496	.060	8.328***
Swear	.169	.034	4.943***

Note1: ***: p<.001, **: p<0.01, *: p<0.05

Table 2 describes the influential predictor variables of detecting the likelihood of cyberbullying based on the Nand, Perera, and Kasture's (2016) dataset. Most of the variables were found to be statistically significant except for the linguistic features 'You' and 'Anger' as derived by the LIWC toolkit.

Table 3 and Figure 2 illustrate the frequencies and percentages of Cyberbully activity being predicted based on the logistic regression predictive model (in Table 2). The frequencies and percentages table (Table 3) shows the results after the prediction of all selected tweets based on emotionally charged words. The majority of emotionally charged words detected by our framework ranged from 30%-50%. The word 'Loser' has the lowest detection rate, which was below 10%. The words 'Dick' (61%) and 'Pussy' (75%) occur in a higher percentage than the rest of the studied words for detecting cyberbullying.

Table 3. Frequency word count for detecting cyberbully based on manifestations of charged language

Charged Language (# tweets)	Cyberbully	Non-Cyberbully	% of Cyberbully
Die	1572	2756	36.32
Bitch	1523	2996	33.70
Cunt	1983	3567	35.73
Dick	2298	1466	61.05
Faggot	3613	972	21.20
Slut	1772	1880	48.52
Shit	1474	3702	28.48
Pussy	2964	989	74.89
Loser	132	1371	8.78
Kill	1548	2315	40.07
Fuck	2629	2470	51.56
Fat	2195	2494	46.81
Suck	1492	2157	40.89
Whore	1182	1095	51.91

Moreover, we combined all the cyberbullying tweets with non-cyberbullying tweets to compare the differences in ten (10) linguistic features from LIWC between the two groups (cyberbullying vs. non-cyberbullying) using Cohen's (1988) D effect size. For example, Cohen's D effect size in category "you" is necessary, and refers to the difference between a cyberbully and non-cyberbully divided by pooled standard deviations of the cyberbully and non-cyberbully. The results of Cohen's D effect size (Table 4) shows that all the categories have at least a small to large effect size except "you," which has the negligible effect size (cyberbullying vs. non-cyberbullying). These results show a significant magnitude in these nine (9) linguistic features between cyberbully and non-cyberbully.

Table 4. The Cohen's D effect size for ten (10) categories from LIWC in our prediction

Category	Cohen's D Effect Size
You	.088(negligible)
Negative emotion	.823(large)
Anger	.734(large)
Biology	1.118(large)
Body	.830(large)
Health	.300(small)
Sexual	1.071(large)
Ingestion	.282(small)
Death	.341(small)
Swear	1.195(large)

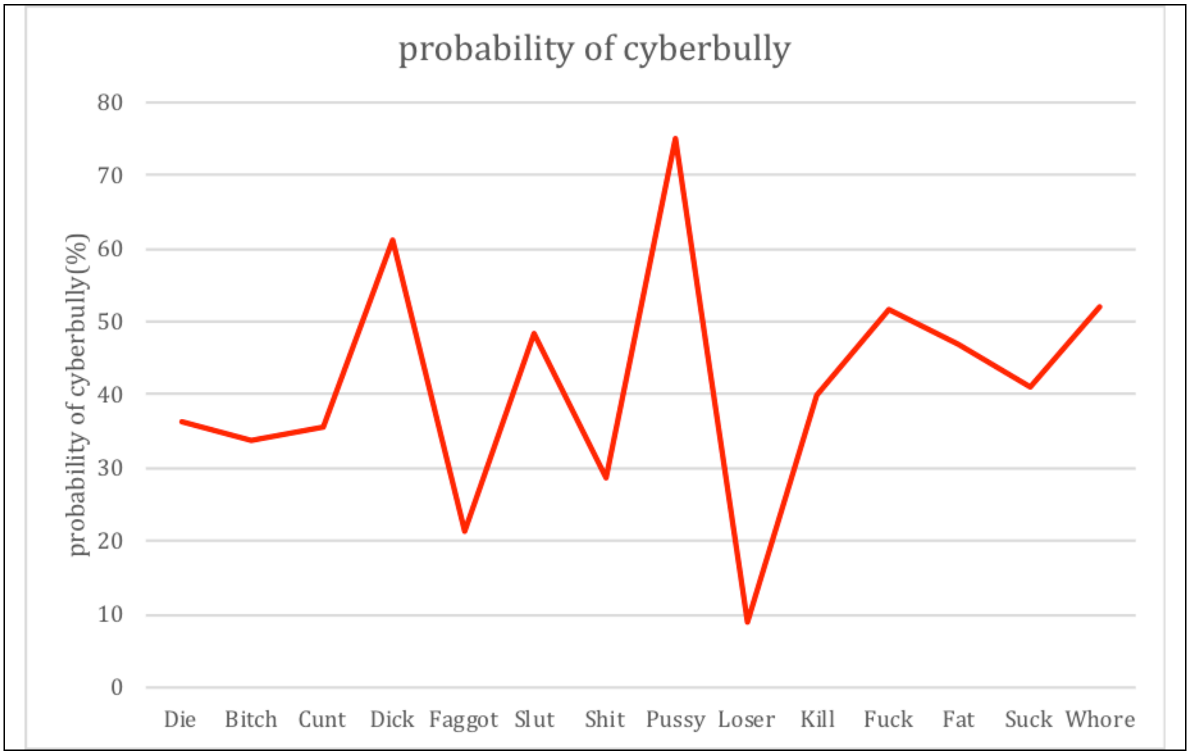

Figure 2. Probability of Cyberbullying for each emotionally-charged word

Figure 2 illustrates the probability of predicting cyberbullying for 14 emotionally charged words. The results of the predictive model show that we have four predictor variables (i.e., dick, pussy, fuck, and whore) that are higher than 50%, and only two predictor variables (i.e., faggot, and loser) that are below 30% of cyberbully (Table 3). We would say that these predictor variables are good at classifying cyberbullying and non-cyberbullying tweets. For example, the percentage of cyberbully language-actions cues in the category "die" is 36.32%, which shows that our model differentiates 1,572 tweets that are cyberbully among a total of 4,328 tweets. The Cyberbully Research Center (CRC) reported cyberbullying victimization rate of 36.5% among 4,972 students in 2019 (Patchin and Hinduja 2019). This data supports the fact that our prediction is close to reality—the CRC's percentage of the lifetime cyberbullying victimization rates in 2019.

DISCUSSION AND CONCLUSION

Predictive analytics modeling can provide preventative confirmation to mitigating entities (e.g., victims' parents, bystanders, school counseling, or law enforcement agencies, etc.) in order to protect against victimization. Our study predicts significant indicators of cyberbullying on Twitter, which confirms that cyberbullying has become a serious problem in social media. We suggest that Twitter could ban the use of highly charged language, which suggests a tendency to participate in cyberbullying. One limitation is the fact that we do not have these human judges to determine the validity of our detection. Our future study includes the conformity on inter-rater reliability and validity of the detection of cyberbullying tendency based on manifestations of emotionally charged language. We also plan to perform this predictive modeling to include more charged language in terms of words that can build a database to illustrate the potential for using charged language to detect cyberbullying on social media.

ACKNOWLEDGEMENTS

The authors wish to thank Florida Center for Cybersecurity (FC2) Capability Building Program for the grant #FC2-3910-1007-00-B, 07/01/2018—12/31/2019.

REFERENCES

Al-garadi, M. A., Varathan, K. D., and Ravana, S. D. 2016. "Cybercrime detection in online communications: The experimental case of cyberbullying detection i the Twitter network," *Computers in Human Behavior* (63), pp. 433-443.

Ang, R. P. 2015. "Adolescent cyberbullying: A review of characteristics, prevention and intervention strategies," *Aggression and Violent Behavior* (25:Part A), pp. 35-42.

Belsey, B. 2007. "Cyberbullying: An emerging threat to the "always on" generation," Belsey, Bill.

Clement, J. 2019. "Number of Twitter users worldwide from 2014 to 2020 (in millions)," Statista.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, (2nd ed.) Lawrence Erlbaum Associates.

Cohen, L. E., and Felson, M. 1979. "Social change and crime rate trends: A routine activity approach," *American Sociological Review* (44:4), pp. 588-608.

Hinduja, S., and Patchin, J. W. 2010. "Bullying, cyberbullying and suicide," *Archives of Suicide Research* (14:3), pp. 206-221.

Ho, S. M., and Hancock, J. T. 2018. "Computer-mediated deception: Collective language-action cues as stigmergic signals for computational intelligence," Proceedings of the 2018 51th Hawaii International Conference on System Sciences (HICSS-51), University of Hawaii, Big Island, Hawaii, 2018, pp. 1671-1680.

Ho, S. M., and Hancock, J. T. 2019. "Context in a bottle: Language-action cues in spontaneous computer-mediated decepion," *Computers in Human Behavior* (91), pp. 33-41.

Ho, S. M., Hancock, J. T., Booth, C., Burmester, M., Liu, X., and Timmarajus, S. S. 2016a. "Demystifying insider threat: Language-action cues in group dynamics," Hawaii International Conference on System Sciences (HICSS-49), IEEE Computer Society, Kauai, Hawaii, 2016a, pp. 2729-2738.

Ho, S. M., Hancock, J. T., Booth, C., and Liu, X. 2016b. "Computer-mediated deception: Strategies revealed by language-action cues in spontaneous communication," *Journal of Management Information Systems* (33:2), pp. 393-420.

Ho, S. M., Hancock, J. T., Booth, C., Liu, X., Liu, M., Timmarajus, S. S., and Burmester, M. 2016c. "Real or Spiel? A decision tree approach for automated detection of deceptive language-action cues," Hawaii International Conference on System Sciences (HICSS-49), IEEE Computer Society, Kauai, Hawaii, 2016c, pp. 3706-3715.

Ho, S. M., Hancock, J. T., Booth, C., Liu, X., Timmarajus, S. S., and Burmester, M. 2015. "Liar, Liar, IM on Fire: Deceptive language-action cues in spontaneous online communication," Proceedings of 2015 IEEE International Conference on Intelligence and Security Informatics, IEEE, Baltimore, MD, 2015, pp. 157-159.

Ho, S. M., Liu, X., Booth, C., and Hariharan, A. 2016. "Saint or Sinner? Language-action cues for modeling deception using support vector machines," in *Social Computing,* *Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), LNCS 9708,* K. S. Xu, D. Reitter, D. Lee and N. Osgood (eds.), Springer International Publishing Switzerland: Washington DC, pp. 325-334.

Kasture, A. 2015. *A predictive model to detect online cyberbullying*, Auckland University of Technology, Auckland, New Zealand.

Kavanaugh, A., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., Natsev, A., and Xie, L. 2012. "Social media use by government: From the routine to the critical," *Government Information Quarterly* (29:4), pp. 480-491.

Langos, C. 2012. "Cyberbullying: The challenge to define," *Cyberpsychology, Behavior and Social Networking* (15:6), pp. 285-289.

Li, Q. 2007. "New bottle but old wine: A research of cyberbullying in schools," *Computers in Human Behavior* (23:4), pp. 1777-1791.

Mesch, G. 2009. "Parental mediation, online activities and cyberbullying," *CyberPsychology & Behavior* (12:4), pp. 387-393.

Nand, P., Perera, R., and Kasture, A. 2016. ""How bullying is this message?" A psychometric thermometer for bullying," Proceedings of the 2016 26th International Conference on Computational Linguistics (COLING'16), Osaka, Japan, 2016, pp. 695-706.

Ong, R. 2015. "Cyber-bullying and young people: How Hong Kong keeps the new playground safe," *Computer Law & Security Review* (31), pp. 668-678.

Patchin, J. W., and Hinduja, S. 2019. "Cyberbullying Research Center," Cyberbullying Research Center.

Peluchette, J. V., Karl, K., Wood, C., and Williams, J. 2015. "Cyberbullying victimization: Do victims' personality and risky social network behaviors contribute to the problem?," *Computers in Human Behavior* (52), pp. 424-435.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., and Trippett, N. 2008. "Cyberbullying: Its nature and impact in secondary school pupils," *Journal of Child Psychology and Psychiatry* (49:4), pp. 376-385.

Xu, J.-M., Kwang-Sung, Zhu, X., and Bellmore, A. 2012. "Learning from bullying traces in social media," Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT'12), Association for Computational Linguistics, Montreal, Canada, 2012, pp. 656-666.