Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS 2020 EU — Proceedings of the Seventh Annual DFRWS Europe

Detecting Cyberbullying "Hotspots" on Twitter: A Predictive Analytics Approach

Shuyuan Mary Ho[a,*], Dayu Kao[b], Ming-Jung Chiu-Huang[b], Wenyi Li[a], Chung-Jui Lai[b]. [a] *Florida State University, Tallahassee, FL, 32306, USA;* [b] *Central Police University, Guishan, Taoyuan 333, Taiwan, ROC*

Keywords
Cyberbullying
Predictive analytics
Logistic regression
Language-action cues
LIWC
Social media
Twitter

ABSTRACT

The ability to discover cyberbullying "hotspots" on social media is vitally important for purposes of preventing victimization. This study attempts to develop a prediction model for identifying cyberbullying "hotspots" by analyzing the manifestation of charged language on Twitter. A total of 140,000 tweets were collected using a Twitter API during September 2019. The study reports that certain charged language in tweets can indicate a high potential for cyberbullying incidents. Cyberbullies tend to share negative emotion, demonstrate anger, and use abusive words to attack victims. The predictor variables related to "biology," "sexual," and "swear" can be further used to differentiate cyberbullies from non-cyberbullies. The study contributes to the detection of cyberbullying "hotspots," by providing an approach to identify a tendency for cyberbullying activity based on computational analysis of charged language. The contribution is significant for mediation agencies—such as school counseling and law enforcement agencies.

1. Introduction

"Hotspots" refers to "an area that has a greater than average number of events, or an area where people have a higher than average risk of victimization" (Gonzales et al., 2005). The global spread of information & communication technology (ICT)—such as Twitter—not only has enabled the freedom to express thoughts and opinions without borders, but also poses cyber-risks to its users, and turns Twitter into a world-wide "cyberbully playground."

Cyberbullies tend to verbally abuse their victims. The abusive language-action cues that cyberbullies employ is the focal point of the study. We thus ask: *What is the probability—based on the charged language-action cues in tweets—for detecting the tendency of a person to be a cyberbully? Is it possible to computationally detect cyberbullying "hotspots"?*

2. The study

We first identified and extracted manifestations of charged language—in terms of certain mean and abusive words (Nand et al., 2016). Then, the prediction model was derived by applying logistic regression analysis based on the classifications proposed in Kasture's (Kasture and advisorsNand, 2015) study.

We collected tweets—based on the words suggested in Nand et al.'s (Nand et al., 2016) study. Our study selectively focused on only 14 emotionally-charged words: die, faggot, fat, fuck, kill, loser, shit, slut, suck, whore, bitch, cunt, dick, and pussy. We thus set the Search API Python program to collect 10,000 tweets for each of the 14 emotionally-charged words—with a total of 140,000 tweets making up the dataset. This dataset was collected from September 7, 2019 to September 13, 2019.

Unformatted tweets and duplicates in the dataset were eliminated. Our final dataset contains 54,894 tweets in total (09/07/19-09/13/19).

3. Data analysis

Our study adopts Kasture's (Kasture and advisorsNand, 2015) dataset as a baseline for classifying data into cyberbullying versus non-cyberbullying designations. Kasture (Kasture and advisorsNand, 2015) developed a predictive model to detect cyberbullying on Twitter. Kasture (Kasture and advisorsNand, 2015) collected 1313 tweets, and used human judgment to tag these tweets. A total of 376 tweets were classified as being cyberbullying tweets. Kasture's (Kasture and advisorsNand, 2015) data was reprocessed and reconverted into numerical values using LIWC 2015.

3.1. Variables

Ten (10) categories among 90 features generated from LIWC 2015 were selected as the independent variables (IVs) (Kasture and advisorsNand, 2015). Cyberbullying is set as the dependent variable (DV).

3.2. Logistic regression prediction model

Kasture's (Kasture and advisorsNand, 2015) dataset was used as the baseline model. That is, text that was marked in Kasture's (Kasture and advisorsNand, 2015) dataset as being cyberbullying is coded as 1

* Corresponding author.

E-mail address: smho@fsu.edu (S.M. Ho).

https://doi.org/10.1016/j.fsidi.2020.300906

Table 1
Logistic regression model using Kasture's [31] dataset as the baseline.

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***

Table 2
Classification using baseline data from Kasture (Kasture and advisorsNand, 2015).

Datasets	time	Cyberbullying		Non-Cyberbullying	
		tweets	percentage	tweets	percentage
Kasture (2015) data	2015	376	28.64%	937	71.36%
Our dataset	2019	21,042	38.33%	33,852	61.67%

(positive) in our prediction model. Likewise, text that was marked in Kasture's (Kasture and advisorsNand, 2015) dataset as non-cyberbullying is coded as 0 (negative) in our prediction model (1 = cyberbully, 0 = noncyberbully).

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$
$$= \beta_0 + \sum_{i=1}^{m} \beta_i \cdot x_i,\ i = 0,\ 1,\ 2\ldots$$

P: is probability function for cyberbullying where y = 0 is non-cyberbullying, and y = 1 is cyberbullying. The x_i is the explanatory variables, and β_i is the coefficient of parameters.

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = -4.270 + .007\mathit{you} + .127\mathit{negemo}$$
$$+ .006\mathit{anger} - .235\mathit{bio} + .281\mathit{body}$$
$$+ .296\mathit{health} + .428\mathit{sexual} + .496\mathit{death}$$
$$+ .169\mathit{swear}.$$

We performed the logistic regression model in R-studio. The results in Table 1 indicate that only “you” and “anger” are not statistically significant to detect cyberbullying. Even though these two variables are not statistically significant, they are still kept in the model because of the extremely small estimated coefficients. Our predictive model is illustrated as:
We classified these probabilities into two categories: Yes (probability greater than or equal to 0.5) and No (probability less than 0.5). “Yes” means cyberbullying, and “No” means non-cyberbullying.

3.3. Classification

We use the above logistic regression model with the 14 abusive words as IVs to differentiate cyberbullying from non-cyberbullying. Table 2 illustrates the results being classified to differentiate cyberbullying from non-cyberbullying—using Kasture's (Kasture and advisorsNand, 2015) data as the baseline.

4. Conclusion

Our study identifies abusive language-action cues as a manifestation of charged language. Fig. 1 illustrates significant differences between tweets

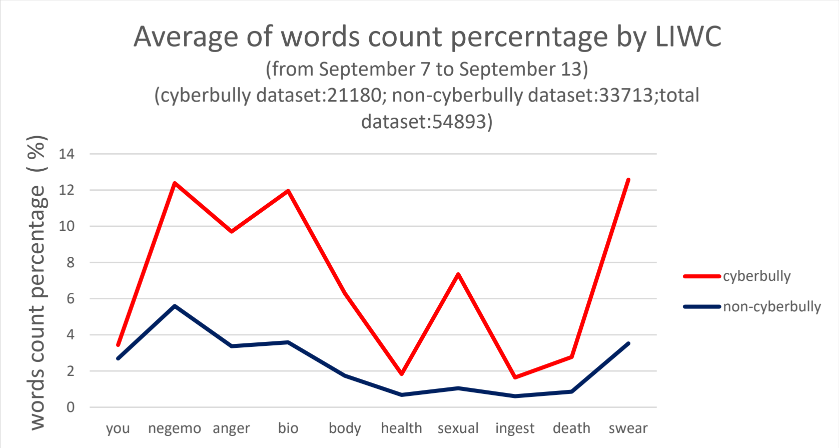

Fig. 1. Comparison between cyberbully and non-cyberbully

with a tendency for cyberbullying language and tweets without a tendency for cyberbullying language.

We conceptualize language-action cues as being an efficacious way to disclose the characteristics of a cyberbully's information behavior, which lends itself to the identification of cyberbullying “hotspots.”

Acknowledgements

The authors wish to thank Florida Center for Cybersecurity (FC2) for the grant 3910-1007-00-B, and the Executive Yuan of the Republic of China for the grants from the Digital Infrastructure Information Security Project-109 of the Forward-looking Infrastructure Development Program.

References

Gonzales, A.R., Schofield, R.B., Hart, S.V., 2005. In: Justice, D.o. (Ed.), Mapping Crime: Understanding Hot Spots. National Institute of Justice, pp. 1–79.

Kasture, A., 2015. A predictive model to detect online cyberbullying. In: advisors, Nand, P. (Eds.), Master of Computer and Information Science. Auckland University of Technology, Auckland, New Zealand.

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