A SURVEY ON CYBERBULLYING DETECTION

Sherly Prakash
MET's College of Advanced Studies MALA

A SURVEY ON CYBERBULLYING DETECTION

Sherly T.T [1] Dr B. Rosiline Jeetha [2]

Research Scholar [1],Reserach Guide [2] Bharathiyar University[1] , Dr.N.G.P. Arts & Science College,Coimbatore
Sherly.dec7@gmail.com [1] , jeethasekar1@gmail.com[2]

ABSTRACT

Cyberbullying can be defined as the use of information and communication technology (ICT) by an individual or a group of users to hassle other users. Cyberbullying has also been extensively recognized as a serious national health problem. Cyberbullying is a significantly unrelenting version of conventional forms of bullying with negative effects on the victim. Also social networking plays an important role in this cyberbullying. This survey paper projects on cyberbullying detection and the recent research works carried out in this research dimension. Certain machine learning algorithms are also applied apart from several methods

Keywords: Cyberbullying, detection, machine learning, social networks, twitter.

1. Introduction

Cyberbullying can be defined 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. Cyberbullying and is capable enough to take multiple forms such as threats, exclusion, name-calling in different contexts which includes social networking sites and mobile phones. Cyberbullying frequently occurs among adolescents on social networking sites (SNS) and has been related to several emotional, psychological and physical problems, as well as it leads to poor academic performance and an increase in suicidal ideation. Diverse impacts on victims have been observed, whether due to factors characterising cyberbullying events or to differences in the resilience of the victims. There exist several strategies that are recommended for preventing and intervening in situations involving cyberbullying. Although the problem of cyberbullying has been studied in social sciences and child psychology for over ten years, there has been only few attempts to study the problem using methods from the field of Artificial Intelligence or Natural Language Processing and other computer science research areas.

2. Related Works

Mohammed Ali Al-garadi et al., proposed a set of unique features network, activity, user, and tweet content derived from Twitter. A supervised machine learning solution has been proposed based on the feature for cyberbullying detection in the Twitter. The evaluation results depicted that the proposed detection model provides a feasible solution to detecting Cyberbullying in online communication environments. The data have been collected from Twitter between January 2015 and February 2015. 2.5 million geo-tagged tweets within a latitude and longitude bounding box of the state of California have been fetched using the sampled API service of Twitter. The authors categorised the features as network, activity, user, and content, to detect cyberbullying behaviour, and used NB, SVM, random forest, and KNN for machine learning. All the four classifiers have been tested in various four settings, namely, basic classifiers, classifiers with feature selection techniques, classifiers with SMOTE alone and with feature selection techniques, and classifiers with cost-sensitive alone and with feature selection techniques. AUC has been considered for the measure of performance. AUC has high robustness for evaluating classifiers. Precision, recall, and f-measure were also used as reference measures. Random forest using SMOTE alone showed the best AUC (0.943) and f-measure (0.936). • Research reports have addressed the obsessive social media involvement turning into a compulsive behaviour among university students.

Adel M. Aladwani and Mohammad Almarzouq have addressed and evaluated the issue by a contingency based. The model postulated that the compulsive social media use arises due to self-awareness factors, and these relationships are moderated by the influence of technological factors. The data has been collected from 407 students of University in Kuwait to test the model. The results of the proposed work indicated that self-esteem has a significant negative influence and the interaction anxiousness has a significant positive influence, on compulsive social media use. It also revealed that the social media use has a significant direct control of influence on problematic learning outcomes; and that social media complementarities plays a moderating role in the model.

Christopher P. Barlett et al conducted three studies to validate a new positive attitude towards cyberbullying measure. The authors developed a new measure, a self-report assessment that consists of nine items. The data for the study have been collected as a larger Scale Validation study where all questionnaires were completed online using the *Qualtrics data collection software*. Their first study used exploratory factor analysis and two distinct factors, Harmful Cyberbullying Attitudes and General Cyberbullying Characteristics, has been found. In their second study the factor structure was replicated using confirmatory factor analysis. The final study further replicated the results of second study and showed that the HCA measure predicted cyberbullying perpetration above and beyond other cyberbullying attitude measures. It has been also stated that the measure proposed showed concurrent and predictive validity. The new HCA attitude measure has predicted two measures of cyberbullying perpetration while controlling for the other similar attitudes, and thus it has been suggested that the new measure of Cyberbullying attitudes is a significant predictor of cyberbullying perpetration while controlling for other existing measures.

An experimental study has been made by Sara Bastiaensens et al in order to examine the influence of contextual factors on bystanders' behavioural intentions to assist the victim or reinforce the bully in cases of harassment on Facebook. Three contextual influences; severity of the incident, the identity of other bystanders who are present and the behaviour of these other bystanders, have been examined by the authors. In addition, gender was added to the models as a control variable. 453 second year students of Flemish secondary schools participated in the study. The experimental results showed that the bystanders had higher behavioural intentions to help the victim when they witnessed a more severe incident. On the other hand, the bystanders had higher behavioural intentions to join in the bullying while other bystanders were good friends rather than acquaintances. Also, an interaction effect was identified between other bystanders' identity and behaviour on behavioural intentions to join in the bullying, and, both helping and strengthen behavioural intentions differed according to gender.

R. Forssell investigates the prevalence of cyberbullying and face-to-face bullying in Swedish working life and its relation to gender and organisational position. A large sample of 3371 respondents has been involved in the study. A cyberbullying behaviour questionnaire (CBQ) has been used in the study, the result shows that 9.7% of the respondents has been labelled as cyberbullied in accordance with Leymann's cut-off criterion. 0.7% of the respondents has been labelled as cyberbullied and 3.5% of the respondents have been labelled as bullied face-to-face. Their study also revealed that men than women were exposed to Cyberbullying to a higher degree. Individuals with a supervisory position were found more exposed to cyberbullying than persons with no managerial responsibility.

Kassandra Gahagan investigated the experience of the college students with cyberbullying on Social Networking Sites (SNS). College students of 196 numbers from a north western university have participated in the study. Their study revealed that19% of the college students had been bullied on SNS and 46% indicated that they had witnessed cyberbullying on SNS. 61% of the college students who witnessed cyberbullying on SNS did nothing to intervene. In their study, the college students were also questioned about their perceived responsibility when they witnessed cyberbullying on SNS. Two diverging themes have been emerged that indicated some college students believed their responsibility to intervene was circumstantial, while others supposed that there is a constant clear level of responsibility for college student cyberbullying bystanders on SNS.

Manuel Gámez-Guadix et al, examined the possible presence of an identifiable group of stable victims of cyberbullying. His work also analysed whether the stability of cyber victimization is associated with the perpetration of cyberbullying and bully–victim status, and tested whether stable victims report a greater number of psychosocial problems compared to non-stable victims and non-involved peers. A sample of 680 Spanish adolescents (410 girls) have been used in completing the self-report measures on cyberbullying perpetration and victimization, depressive symptoms, and problematic alcohol use at two time points that were separated by one year. The cluster analyses results suggested the existence of four distinct victimization profiles. Stable-Victims (5.8% of the sample) reported victimization at both Time 1 and Time 2. It has been found that they were more likely to fall into the bully–victim category and presented more cyberbullying perpetration than the rest of the groups. Time1-Victims(14.5% of the sample) and Time 2-Victims (17.6% of the sample) presented victimization only at one time. Non-Victims (61.9% of the sample) presented minimal victimization at both times. Overall, the Stable victims group exhibited higher scores of depressive symptoms and problematic alcohol use over time than the other groups, whereas the Non-Victims exhibited the lowest of these scores. It has been stated that their findings have major implications for prevention and intervention efforts intended at reducing cyberbullying and its consequences.

Casey R. Guillot examined the potential longitudinal associations between anhedonia and internet-related addictive behaviours in 503 at-risk emerging adults. Demographic characteristics and descriptive statistics of the experiment showed that at Time 1 assessment T1 and Time 2 assessment T2, 21.4% and 21.8% of participants, respectively, endorsed addiction to at least one of the three online activities, internet browsing, social media, or online shopping, that comprise the IA scale. In regard to covariates, both male gender (OR ¼= 3.1, p = 0.008) and high school graduation (OR = 3.9, p = 0.008) were prospectively connected with video game addiction. The result indicated that trait anhedonia prospectively forecasted greater levels of compulsive internet use and addiction to online activities as well as a greater likelihood of addiction to online/offline video games. Their findings suggested that anhedonia may contribute to the development of internet-based addictive behaviors in the emerging adult population, and thus, interventions that target anhedonia in upcoming adulthood may help prevent or treat internet addiction.

Hosein Jafarkarimi et al identified the influential factors that have impact on individuals' ethical decision-making and also proposed a model of the factors that are significant in the ethical decision-making process in the SNS context. Their study employed the Theory of Planned Behavior (TPB). A scenario based questionnaire has been used to predict the behavioral intention, which included personal normative beliefs, moral intensity and perceived threat of legal punishment to the main constructs of TPB namely attitude, subjective norms, and perceived behavioural control. The probable effect on the proposed model has been investigated by moderating effects of several factors, including age, gender, level of income, ego strength, locus of control and religion. Four scenarios have been embedded in the survey instrument. The collected data for the 441 questionnaires has been analysed using the partial least squares structural equation modelling technique. Their results showed attitude to be the most influential factor, followed by subjective norms, perceived behavioural control, personal normative beliefs, and moral intensity.

Matthew Pittman and Brandon Reich used a mixed-design survey to test the potential possibility of image-based platforms such as Instagram and Snapchat and text-based platforms such as Twitter and Yik Yak, in ameliorating loneliness due to the enhanced intimacy they offer. Their study has been made on young adults as this population is at once most likely to use social media and to suffer from loneliness. Their quantitative results suggested that loneliness may decrease, while happiness and satisfaction with life may increase, as a function of image-based social media use. In contrast, text-based media use appears ineffectual. Qualitative results suggested that the observed effects were being due to the enhanced intimacy offered by image based social media use.

Hannah L. Schacter at al examined how cyber victims' disclosures on Facebook influence bystanders' attributions of blame, empathy, and intention to intervene on behalf of a victim following a cyberbullying incident. Participants of 118 numbers have been randomly assigned to view the Facebook profile of a cyber-victim who posted an update ranging in personal disclosure and valence. Their results indicated that viewing the high disclosure profile regardless of valence, caused participants to blame the victim more and feel less empathy for the victim, which in turn predicted lower likelihood of bystander intervention with the bullying incident.

Andreas Weiler et al , studied the run-time and task-based performance of several state-of-the-art event detection techniques for Twitter. The authors presented a two-pronged approach. The first approach ensures the comparable run-time performance results by providing streaming implementations of all techniques based on a data stream management system. Second, propose several new measures have been proposed to assess the relative task-based performance of event detection techniques. Finally, scoring functions have been defined based on selected measures that revealed how the different techniques relate to each other as well as where their strengths and weaknesses lie. In order to reproducibly compare run-time performance, their approach was based on a general-purpose data stream management system, whereas task-based performance is automatically assessed based on a series of novel measures.

A previous study proposed an approach for offensive language detection that was equipped with a lexical syntactic feature and demonstrated a higher precision than the traditional learningbased approach (Chen, Zhou, Zhu, & Xu, 2012). A YouTube databased study (Dadvar, Trieschnigg, Ordelman, & de Jong, 2013) applied SVM to detect cyberbullying, and determined that incorporating user-based content improved the detection accuracy of SVM. Using data sets from MySpace, Dadvar et al. developed a gender-based cyberbullying detection approach that used the gender feature in enhancing the discrimination capacity of a classifier. Dadvar et al. and Ordelman et al. included age and gender as features in their approach; however, these features were limited to the information provided by users in their online profiles. Moreover, most studies determined that only a few users provided complete information about themselves in their online profiles. Alternatively, the tweet contents of these users were analysed to determine their age and gender (D. Nguyen, Gravel, Trieschnigg, & Meder, 2013). Several studies on cyberbullying detection utilized profane words as a feature (Kontostathis, Reynolds, Garron, & Edwards, 2013), thereby significantly improving the model performance. A recent study (Squicciarini, Rajtmajer, Liu, & Griffin, 2015) proposed a model for detecting cyberbullies in MySpace and recognizing the pairwise interactions between users through which the influence of bullies could spread. Nalini and Sheela proposed an approach for detecting cyberbullying messages in Twitter by applying a feature selection weighting scheme (Nalini & Sheela, 2015). Chavan and Shylaja included pronouns, skip-gram, TFeIDF, and N-grams as additional features in improving the overall classification accuracy of their model (Chavan & Shylaja, 2015).

3. Findings and Conclusions

Most networking sites today prohibit the use of offensive and insulting comments. But this partially being carried out and filtered to a limited extent. As there is enormous amount of data available it is impossible to take help of human moderators to manually flag each insulting and offensive comments. This leads the motivation of researchers to provide a solution to the cyberbullying problem. Among many research contributions made, supervised machine learning solution, self - report assessment based techniques, contextual influences based analysis, questionnaire based analysis, clustering techniques, statistical methods are discussed in the literatures. From this survey it is evident that much scope exists to detect cyberbullying using text mining.

References

[1] M. A. Al-garadi, K. D. Varathan, S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter

network," Computers in Human Behavior, vol. 63, pp. 433 - 443, 2016.

[2] A. Weiler, M. Grossniklaus, M. H. Scholl, "An evaluation of the run-time and task-based performance of event detection techniques for Twitter," Information Systems, vol. 62, pp. 207 - 219, 2016.

[3] M. A. Al-garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, A. M. Al-Kabsi, "Using online social networks to track a pandemic: A systematic review," Journal of Biomedical Informatics, vol. 62, pp. 1 - 11, 2016.

[4] H. Jafarkarimi, R. Saadatdoost, A. T. H. Sim, J. M. Heed, "Behavioral intention in social networking sites ethical dilemmas: An extended model based on Theory of Planned Behavior," Computers in Human Behavior, vol. 62, pp. 545 - 561, 2016.

[5] A.H. Hamer, E.A. Konijn, "Can emotion regulation serve as a tool in combating cyberbullying?," Personality and Individual Differences, vol. 102, pp. 1 - 6, 2016.

[6] K. Gahagan, J. M. Vaterlaus, L. R. Frost, "College student cyberbullying on social networking sites: Conceptualization, prevalence, and perceived bystander responsibility," Computers in Human Behavior, vol. 55, pp. 1097 - 1105, 2016.

[7] R. Forssell, "Exploring cyberbullying and face-to-face bullying in working life – Prevalence, targets and expressions," Computers in Human Behavior, vol. 58, pp. 454 - 460, 2016.

[8] H. L. Schacter, S. Greenberg, Jaana Juvonen, "Who's to blame?: The effects of victim disclosure on bystander reactions to cyberbullying," Computers in Human Behavior, vol. 57, pp. 115 - 121, 2016.

[9] S. Bastiaensens, H. Vandebosch, K. Poels, K. V. Cleemput, A. DeSmet, I. D. Bourdeaudhuij, "Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully," Computers in Human Behavior, vol. 31, pp. 259 - 272, 2014.

[10] C. R. Guillot, M. S. Bello, J. Y. Tsai, J. Huh, A. M. Leventhal, S. Sussman, "Longitudinal associations between anhedonia and internet-related addictive behaviors in emerging adults," Computers in Human Behavior, vol. 62, pp. 475 - 479, 2016.

[11] M. Pittman, B. Reich, "Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words," Computers in Human Behavior, vol. 62, pp. 155 - 167, 2016.

[12] A. M. Aladwani, M. Almarzouq, "Understanding compulsive social media use: The premise of complementing self-conceptions mismatch with technology," Computers in Human Behavior, vol. 60, pp. 575 - 581, 2016.

[13] M. G. Guadix, G. Gini, E. Calvete, "Stability of cyberbullying victimization among adolescents: Prevalence and association with bully–victim status and psychosocial adjustment," Computers in Human Behavior, vol. 53, pp. 140 - 148, 2016.

[14] J. L. Pettalia, E. Levin, J. Dickinson, "Cyberbullying: Eliciting harm without consequence," Computers in Human Behavior, vol. 29, pp. 2758 - 2765, 2013.

[15] C. P. Barlett, K. Helmstetter, D. A. Gentile, "The development of a new cyberbullying attitude measure," Computers in Human Behavior, vol. 64, pp. 906 - 913, 2016.

[16] O. T. Aricak, A. Ozbay, "Investigation of the relationship between cyberbullying, cybervictimization, alexithymia and anger expression styles among adolescents," Computers in Human Behavior, vol. 55, pp. 278 - 285, 2016.

[17] Y. Chen, Y. Zhou, S. Zhu, H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, pp. 71 – 80, 2012.

[18] M. Dadvar, D. Trieschnigg, R. Ordelman, F. Jong, "Improving Cyberbullying Detection

with User Context," Advances in Information Retrieval, vol. 78, pp. 693 – 696, 2013.

[19] D. Nguyen, R. Gravel, D. Trieschnigg, T. Meder, "How Old Do You Think I Am?; A Study of Language and Age in Twitter," Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp. 439 – 448, 2013.

[20] A. Kontostathis, K. Reynolds, A. Garron, L. Edwards, "Detecting cyberbullying: query terms and techniques," Proceedings of the 5th Annual ACM Web Science Conference, pp. 195 – 204, 2013.

[21] A. Squicciarini, S. Rajtmajer , Y. Liu, C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 280 – 285, 2015.

[22] K. Nalini , L. J. Sheela, "Classification of Tweets Using Text Classifier to Detect Cyber Bullying," Advances in Intelligent Systems and Computing, vol. 338, pp. 637 – 645, 2015.

[23] V. S. Chavan, S. S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2354-2358, 2015.