SUPERVISED FEATURE SELECTION BASED EXTREME LEARNING MACHINE (SFS-ELM) CLASSIFIER FOR CYBER BULLYING DETECTION IN TWITTER

Sherly T T
Research Scholar, PG and Research Department of Computer Science
,Dr. N.G.P. Arts and Science College Coimbatore, India
sherly.dec7@gmail.com

Dr.B. Rosiline Jeetha
Research Guide, PG and Research Department of Computer Science,
Dr.N.G.P. Arts and Science College, Coimbatore, India
jeethasekar@gmail.com

***Abstract*—** Cyber bullying detection that are prevailing commonly in social networks like Twitter is one of the focussed research area. Text mining and detecting cyber bullying has several research challenges and lot of research scope to work with. This research work makes use of supervised feature selection by ranking method in order to choose the features from the tweets. After that extreme learning machine (ELM) classifier is employed in order to perform the detection of cyber bullying tweets. Performance metrics such as accuracy and time taken for classification are chosen in order to evaluate the efficiency of the classifiers namely ELM and the proposed SFS-ELM. Implementations are done in MATLAB tool. From the obtained results it is evident that the proposed SFS-ELM produces better results than that of ELM

***Keywords*—** *Cyber bullying, Twitter, feature selection, classification, detection, extreme learning machine, machine learning.*

I. INTRODUCTION

Text mining is the field of data mining that has several varieties of text analytics. Text mining is the procedure of getting information / knowledge from text. There are lot of techniques used in performing text mining research. As far as the real time real time scenario is taken into account, many texts are in the form of semi – structured that tends to more scope of research in field of text data mining. Cyber bullying detection in twitter is of the interesting research area where it contains lot of research dimensions in text mining. Cyber criminals are making use of social network for doing several kinds of cybercrimes which includes phishing (Aggarwal et al., 2012), spamming (Yardi et al., 2009), spread of malware (Yang et al., 2012), and cyberbullying (Weir et al., 2011).

Due to the raise of social networking sites and internet particularly cyber bullying became a major threat for the users (O'Keeffe & Clarke-Pearson, 2011). Salmivalli, 2010 came up with the definition stating that Cyberbullying is performed by making use of information and communication technology by a single person or as a cluster of users to indulge in bothering erstwhile user(s). In the literature Xu, Jun, Zhu, & Bellmore, 2012 mentioned that Cyber bullying is recognized as a severe country's health issue that leads to victims disclose a noticeably elevated peril of with zilch to live for though provoking.Twitter is a widespread online social network facility that makes users to send and read 140-character messages. The Twitter network presently contains over 590 million users, of which 297 million users are fanatically barter a few words through this network and generate approximately 610 million tweets per day. Around 82% of these ardent users of twitter are sending their tweets through the help of their smart phones and tablets. Even supposing twitter turned into an imperative, near real-time communication channel (Kavanaugh et al., 2012), a study resolute that Twitter is turning into a "cyberbullying playground" (Xu et al., 2012).

Twitter changed its face into new targets for cybercrime, and wicked users try to carry out illegal activities such as cyber-attacks, bullying, fraudulent information, organized

crimes, and even terrorist attack planning on these systems (Yu et al., 2015). In addition Twitter is more prone to malwares, spam messages and other offensive materials (Akoglu et al., 2010; Gao et al., 2012; Hassanzadeh and Nayak, 2013; Rahman et al., 2012; Shrivastava et al., 2008). It is obvious that cyberbullying causes not only monitory loss and also affects a person's behaviour patterns. Such activities of the cyberbullies in Twitter necessitate the scope of cyber forensics in social networks arena. With all this kept in mind, this research work aim to propose supervised feature selection based extreme learning machine (SFS-ELM) classifier for cyber bullying

II. RELATED WORKS

Mohammed Ali Al-garadi et al., presented a set of inimitable features; which includes network, activity, user, and tweet content plagiaristic from Twitter. A supervised engine wisdom procedure has been projected based on the feature for cyberbullying detection in the Twitter. The evaluation results of the authors work offered a practicable result for Cyberbullying detection that most commonly occurs in internet using scenario with help of their proposed detection model. The authors obtained data collected from Twitter during January 2015 and February 2015 and made use for their evaluation process. The collected twitter data almost consists of two and half million geo-tagged tweets that falls in the geographic region of California with the help of application programming interface service of Twitter. The authors categorised the features as system, action, client, and content, to spot cyberbullying behaviour, and used NB, SVM, random forest, and KNN for engine wisdom. All the four classifiers have been evaluated in four various settings, to be specific, essential classifiers, classifiers with highlight determination systems, classifiers with SMOTE alone and with highlight choice procedures, and classifiers with cost-touchy alone and with highlight choice strategies. AUC has been considered for the measure of performance. AUC has high robustness for evaluating classifiers. Accuracy, remind, and f-measure were also used as orientation methods. Random forest using SMOTE alone proven the finest AUC (0.943) and f-measure (0.936).

R. Forssell investigated the occurrence of cyberbullying and face-to-face bullying in Swedish operational time and its family member towards gender and organisational spot. A large sample of 3371 respondents has been involved in the study. A cyberbullying behaviour questionnaire (CBQ) has been used in the study; 9.7% of the respondents have been categorized as cyberbullied in harmony with Leymann's cut-off criterion, 0.7% of the respondents as cyberbullied and 3.5% of the respondents as bullied confronting each other. Their study also revealed that men when compared with women were exposed to a high degree of Cyberbullying. People with a supervisory position were observed with additional exposure on cyberbullying than persons with no administrative duty.

Manuel Gámez-Guadix et al, examined the possibility of the nearness of an identifiable gathering of stable casualties of cyberbullying. The author analysed the ¬¬¬solidity of cyber victimization linked through the perpetration of cyberbullying and bully–victim position. The psychosocial problems of non-stable fatalities and non-involved nobles have been compared with stable victims. The authors used a taster of 680 Spanish adolescents which includes 410 girls in completing the self-report method on cyberbullying perpetration and persecution, depressive symptom, and challenging alcohol employ at two time points that were split by one year. The cluster analyses grades suggested the existence of four distinct victimization profiles. Stable-Victims (5.8% of the sample) were observed with persecution at mutually Time 1 and Time 2. The authors also found that the steady fatalities were more expected to drop into the bully–victim class and accessible extra cyberbullying perpetration than the relax of the groups. Time1-Victims (14.5% of the sample) and Time 2-Victims (17.6% of the sample) offered persecution only at one time. Non-Victims (61.9% of the sample) offered least persecution at both times. Overall, the authors observed that the steady fatalities set with higher scores of depressive side effects and

hazardous use of alcohol more than time than the erstwhile groups, while, the Non-Victims with the lowest of the scores. Their result have been observed with main implications for deterrence and intrusion hard work intended at dropping cyberbullying and its penalty.

In the literature by Chen et al in 2012 proposed a mechanism in order to identify offensive language. The mechanism has been built up with a lexical syntactic feature and confirmed an elevated precision when compared with the conventional learning based mechanisms.

Dadvar et al., in the year 2013 conducted a study on the YouTube database. The authors made use of support vector machine in order to detect Cyberbullying. The authors stated in the literature that by making use of user-based content improved the detection accuracy of SVM. Using data sets from MySpace, Dadvar et al. built up a sexual orientation based cyberbullying location approach that utilized the sexual category include in upgrading the separation limit of a classifier. Dadvar et al. what's more, Ordelman et al. included age and sexual orientation as components in their approach; in any case, these elements were constrained to the data gave by clients in their online profiles. Besides, most reviews established that lone a couple of clients gave finish data about themselves in their online profiles. On the other hand, the tweet substance of these clients were broke down to decide their age and sexual category (D. Nguyen, Gravel, Trieschnigg, & Meder, 2013).

A few reviews on cyberbullying discovery used dishonor words as a component (Kontostathis, Reynolds, Garron, and Edwards, 2013), along these lines altogether enhancing the model execution. A current review (Squicciarini, Rajtmajer, Liu, and Griffin, 2015) proposed a model for distinguishing cyberbullies in MySpace and perceiving the pairwise communications between clients through which the impact of spooks could spread. Nalini and Sheela proposed an approach for distinguishing cyberbullying messages in Twitter by applying an element determination weighting plan (Nalini and Sheela, 2015). Chavan and Shylaja included pronouns, skip-gram, TFeIDF, and N-grams as extra components in enhancing the general grouping exactness of their model (Chavan and Shylaja, 2015).

III. PROPOSED WORK

III.1 Supervised Feature Selection using Ranking Method

As far as supervised approach is concerned, a class of tweets are obtained as the basic unit or context for computing connotation scores for language. Connotation compute fundamentally portrays how ordinary a fastidious words' incidence is in a set of tweets when judge to the previous sets of tweets. When there are unexpected words in the tweet are present at that point significance measure brings about high importance scores. In this portion it is analogous to the Multinomial Naive Bayes in which the all the tweets in a class is converged into a solitary tweet and afterward the probabilities are assessed from this one huge class tweets. In supervised meaning measure, parameter represents tweets that fit in to class and speaks to the total preparing set. It is assumed that a component shows up times in the dataset, and times in the tweets of class. The length of dataset (i.e. preparing set) and class measured by the aggregate term frequencies is and individually. is the rate of the length of the dataset and the class which figure in (3). In view of these the quantity of false cautions is characterized in(4)

$$L = \sum_{d \in S} \sum_{w \in d} tf_w \quad \ldots (1)$$

$$B = \sum_{d \in c_j} \sum_{w \in d} tf_w \quad \ldots (2)$$

$$N = \frac{L}{B} \ldots (3)$$

$$NFA(w, c_j, S) = \binom{k}{m} \cdot \frac{1}{N^{m-1}} \ldots \quad (4)$$

The connotation attain of the phrase w in a set c_j is distinct as:

$$meaning(w, c_j) = -\frac{1}{m}\log NFA(w, c_j, S) \dots (5)$$

In enjoin to make simpler the calculation connotation formula can be rewrite as:

$$meaning(w, c_j) = -\frac{1}{m}\log\binom{k}{m} - [(m-1)\log N] \dots (6)$$

The well-built the connotation attain of a phrase w in a set c_j can be perceived as that the given word w is further meaningful, important or edifying for that class. It is firmly to mention that, the phrases with well-built connotation score be in contact to extra important, noteworthy or enlightening words for that particular set. On the other hand, for characteristic choice is required for an approach to consolidate these class-based scores into one and select top R features. In enjoin to do this ranking method is applied. Ranking perform sort the elements by utilizing their importance scores for each class. For instance, the rank of the main component on each sorted rundown will be 1 and the last component will be the word reference estimate. In this research work rank of the components is used in each class rather than their importance scores. When consolidating these class based records into a solitary component list, for each element the most noteworthy rank among all classes are picked as in (7).

$$score(w) = \max_{c_j \in C}(Rank(w, c_j)) \dots (7)$$

III.2. Extreme Learning Machine Classifier

Once when the feature selection task is completed, ELM is employed for performing classification task. Given a set of N training samples (x_i, t_i) and 2L concealed neurons altogether (that is, each of the two shrouded layer has L shrouded neurons) with the initiation work g (x). At first arbitrarily introduce the association weight matrix between the info layer and the first shrouded layer W and the inclination matrix of the first concealed layer B, and afterward figure the weight matrix β between the second concealed layer and the yield layer.

$$g(W_H H + B_1) = H_1 \dots (8)$$

where W_H indicates the weight matrix between the first shrouded layer and the second concealed layer. It is assumed that the first and second concealed layers have a similar number of neurons, and subsequently W_H is a square matrix. The documentation H indicates the yield between the first concealed layer as for all N preparing tests. The matrices B_1 *and* H_1 individually speak to the inclination and the normal yield of the second shrouded layer.

The normal yield of the second shrouded layer can be ascertained as

$$H_1 = T\beta^{+} \dots (9)$$

where β† is the MP widespread contrary of the matrix β. The manipulative means of β† is the alike as before discussed for H†, namely $\beta^{\dagger} = (\beta^{T}\beta)^{-1}\beta^{T}$ *if* $\beta^{T}\beta$ is non-singular, or alternatively $\beta^{\dagger} = \beta^{T}(\beta^{T}\beta)^{-1}$ *if* $\beta\beta^{T}$ is non-singular. Consequently it is defined the augmented matrix $W_{HE} = [B_1 W_H]$, and calculate it as

$$W_{HE} = g^{-1}(H_1) H_E^{\dagger} \dots (10)$$

where $H_E^{\dagger}$ is the widespread contrary of $H_E = [1H]^T$, 1 denotes a one-column vector of size N whose components are the scalar unit 1, where the notation g(x) denotes the contrary of the calculation of $H^{\dagger}$ proceeds in the fashion described some time recently. The investigations directed to test the execution of the ELM calculation. In order to perform the classification task extensively used logistic sigmoid function g (x) = $1/(1+e^{-x})$ is used. The real yield of the second shrouded layer is calculated as

$$H_2 = g(W_{HE} H_E) \dots (11)$$

and finally, the mass matrix β_{new} flanked by the second shrouded layer and the real layer is calculated as

$$\beta_{\text{new}} = H_2^{\dagger} T \dots (12)$$

where $H_2^{\dagger}$ is the MP widespread contrary of H_2, gotten utilizing the approach talked about some time recently. The ELM yield in the wake of preparing can be communicated as

$$f(x) = H_2 \beta_{new} \dots (13)$$

Algorithm 1. ELM Algorithm

Input: N training samples X = $[x_1, x_2, ..., x_N]^T$, T = $[t_1, t_2, ..., t_N]^T$ and 2L hidden neurons in total with activation function g (x)

1: Haphazardly create the association weight matrix between the information layer and the first shrouded layer W and the inclination matrix of the first concealed layer B and for straightforwardness, W_{IE} is defined as[B W] and likewise, X_E is defined as $[1\ X]^T$.

2: Calculate $H = g(W_{IE}\ X_E)$:

3: Acquire weight matrix between the second shrouded layer and the yield layer β = $H^{\dagger}T$

4: Compute the normal yield of the second concealed layer $H_1 = T\beta^{\dagger}$

5: Decide the parameters of the second shrouded layer (association weight matrix between the first and second concealed layer and the predisposition of the second shrouded layer) $W_{HE} = g^{-1}(H_1)H_E^{\dagger}$

6: Acquire the real yield of the second concealed layer $H_2 = g(W_{HE}\ H_E)$

7: Recalculate the weight matrix between the second shrouded layer and the yield layer $\beta_{new} = H_2^{\dagger}T$

Output: The final output of ELM is

$$f(x) = \{[W_H g(W\ X + B) + B_1]\}\beta_{new}$$

IV. RESULTS AND DISCUSSIONS

Performance metrics such as classification accuracy and time taken for classification are chosen for comparison. 4556 tweets from various topics such as demonetisation, kids, mobilephones, sachin and whatsapp words are searched in Twitter and analysed as positive opinion tweets and negative opinion tweets. The analyzed tweets are presented in Table 1. Implementation are carried out using MATLAB 2012.

A. Figures and Tables

Table 1. Collected Tweets with various search terms and opinion analysis

File Name	Total No. of Tweets	Actual	
		Positive Opinion Tweets	Negative Opinion Tweets
demonetisation.txt	1003	498	505
kids.txt	984	402	582
mobilephones.txt	783	599	184
sachin.txt	994	483	511
whatsapp.txt	792	599	193

- True Positive (TP) → Correctly identified as positive opinion tweets
- False Positive (FP) → Incorrectly identified as positive opinion tweets
- True Negative (TN) → Correctly identified as negative opinion tweets
- False Negative (FN) → Incorrectly identified as negative opinion tweets

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100$$

Table 2. Performance Analysis of the Classifiers

Performance of the existing ELM classifier

File Name	TP	TN	FP	FN	Accuracy (%)	Execution Time (seconds)
demonetisation.txt	413	398	101	91	80.86	192
kids.txt	335	461	91	97	80.89	183
mobilephones.txt	467	159	79	78	79.95	177
sachin.txt	394	417	82	101	81.59	199
whatsapp.txt	443	191	97	61	80.05	169
Performance of the proposed SFS-ELM classifier						
File Name	TP	TN	FP	FN	Accuracy (%)	Execution Time (seconds)
demonetisation.txt	461	474	41	27	93.22	147
kids.txt	409	514	33	28	93.80	139
mobilephones.txt	562	168	14	39	93.23	128
sachin.txt	447	479	36	32	93.16	132
whatsapp.txt	579	162	31	20	93.56	133

Table 2 portrays the performance analysis of the proposed SFS-ELM and existing ELM classifiers. It can be observed that the overall accuracy of the SFS-ELM classifier is improved by 13%. The implementation of the proposed SFS-ELM and existing ELM classifiers are implemented using MATLAB. The performance analysis in terms of accuracy is shown in **Fig.1** It is to be noted that the execution time of the proposed SFS-ELM classifier is comparatively lesser than that of the existing ELM classifier. The performance analysis in terms of execution time is shown in **Fig.2.**

Fig. 1. MATLAB Result Graph for Performance Analysis of the Classifiers in terms of Accuracy (in percentage)

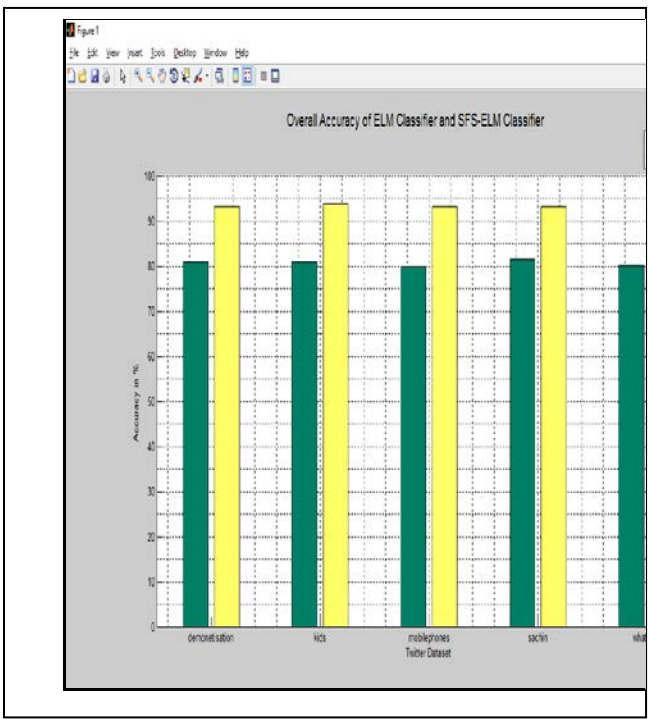
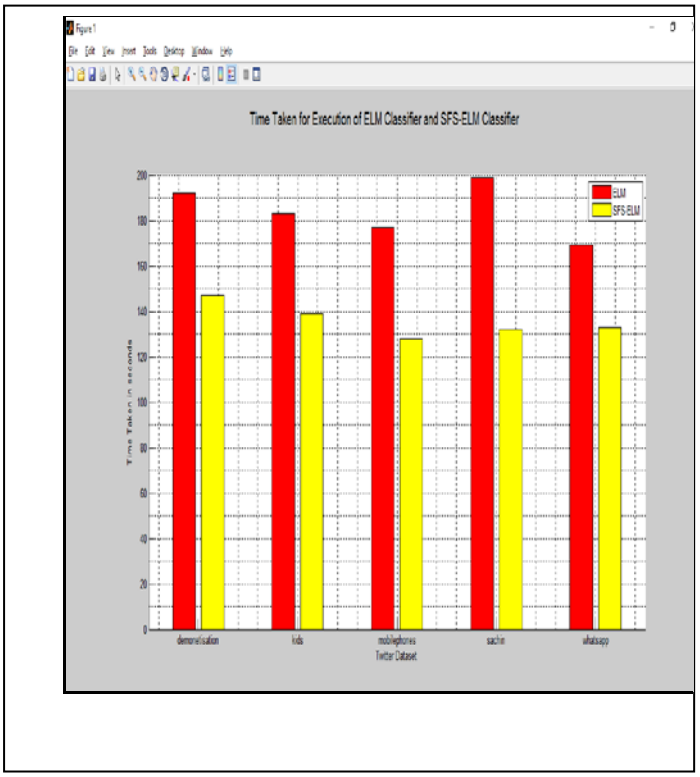


Figure Fig.2. MATLAB Result Graph for Performance Analysis of the Classifiers in terms of Execution time (in seconds)



V. CONCLUSIONS AND FUTURE WORKS

This research work proposes design and development of supervised feature selection based extreme learning machine (SFS-ELM) classifier for cyber bullying detection in twitter that aims to improve the detection accuracy and reduce the execution time. Supervised feature selection is carried out by making use of ranking strategy. Extreme learning machine classifier is employed for carrying out detection. Five datasets are collected from Twitter namely demonetisation, kids, mobile phones, sachin and whatsapp and 1003, 984, 783, 994, 792 tweets are collected respectively. SFS-ELM classifier is implemented in MATLAB and the results poses better accuracy with reduced execution time.

References

[1] Aggarwal, A. Rajadesingan, P. Kumaraguru, "PhishAri: Automatic Realtime Phishing Detection on Twitter", eCrime Researchers Summit (eCrime), pp. 1-12, 2012.

[2] 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.

[3] L. Kavanaugh, E. A. Fox, S. D. Sheetz, S. Yang, L .T. Li, D. J. Shoemaker, A. Natsev, L. Xie, "Social Media Use by Government: From the Routine to the Critical', Government Information Quarterly, vol. 29, no.4, pp.480-491, 2012.

[4] 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.

[5] Salmivalli, "Bullying and the Peer Group: A Review", Aggression and Violent Behavior, vol. 15, no. 2, pp. 112-120, 2010.

[6] Yang, R. Harkreader, J. Zhang, S. Shin, S. Gu, "Analyzing Spammers' Social Networks for Fun and Profit: A Case Study of Cyber Criminal Ecosystem on Twitter", Proceedings of the International Conference on world wide web, pp. 71-80, 2012.

[7] 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.

[8] G. R. Weir, F. Toolan, D. Smeed, "The Threats of Social Networking: Old Wine in New Bottles?", Information Security Technical Report, vol. 16, no. 2, pp. 38-43,2011.

[9] G. S. O'Keeffe, K. Clarke-Pearson, "The Impact of Social Media on Children, Adolescents, and Families", Pediatrics, vol. 127, no. 4, pp. 800-804, 2011.

[10] H. Gao, Y. Chen, K. Lee, D. Palsetia, A. N. Choudhary, "Towards Online Spam Filtering in Social Networks", Network and Distributed System Security Symposium Conference, 2012.

[11] J. M. Xu, K. S. Jun, X. Zhu, A. Bellmore, "Learning from Bullying Traces in Social Media", Proceedings of the Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, pp.656-666, 2012.

[12] 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.

[13] 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.

[14] M. Dadvar, D. Trieschnigg, R. Ordelman, F. Jong, "Improving Cyberbullying Detection with User Context," Advances in Information Retrieval, vol. 78, pp. 693 – 696, 2013.

[15] 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.

[16] M. S. Rahman, T. K. Huang, H. V. Madhyastha, M. Faloutsos, "Efficient and Scalable Socware Detection in Online Social Networks", Proceedings of the 21st USENIX conference on Security symposium, pp. 663–678, 2012.

[17] N. Shrivastava, A. Majumder and R. Rastogi, "Mining (Social) Network Graphs to Detect Random Link Attacks," IEEE International Conference on Data Engineering, pp. 486-495, 2008.

[18] 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.

[19] R. Hassanzadeh and R. Nayak, "A Rule-Based Hybrid Method for Anomaly Detection in Online-Social-Network Graphs," IEEE International Conference on Tools with Artificial Intelligence, pp. 351-357, 2013.

[20] R. Yu, X. He, Y. Liu, "GLAD: Group Anomaly Detection in Social Media Analysis", ACM Transactions on Knowledge Discovery from Data, vol. 10, no.2, pp. 39 - 61,2015.

[21] S. Yardi, D. Romero, G. Schoenebeck, "Detecting Spam in a Twitter Network", First Monday, vol. 15, No. 1, 2008.

[22] 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.

[23] 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.)