ISSN No:-2456-2165

PSO-GA Algorithm Based Cyberbullying Detection

Md. Tofael Ahmed
Department of Computer Science
South Asian University
New Delhi, India

Md. Imran Hossain Showrov
Department of Computer Science
South Asian University
New Delhi, India

Abstract:- Online social networks (OSNs) have seen a rapid growth in recent time. This growth in popularity has also paved the way for the threat of cyberbullying to grow to an extent that was never seen before. Cyberbullying is more alarming day by day in our present life. But the technical solution for preventing and detection of cyberbullying is very limited. Most of the studies on cyberbullying detection focus on detection, not prevention. In this paper, we show that the automatic cyberbullying detection both individual and group communication.

Keywords:-*Cyberbullying detection; SVM classifier; PSO-GA.*

I. INTRODUCTION

Social networks are the place where huge people interact, share, discuss and disseminate knowledge and information for mutual benefits. Cyberbullying is a form of harassment or bullying with a use of electronic devices such as mobile phones, laptops etc. through social networks. Nowadays cyberbullying increases rapidly especially among teenagers with the development of the internet [1]. Cyberbullying analysis among youths is performed with the influence of disability status on participant's ability detect the presence of cyberbullying. Both participants with and without the disability are displayed high victimization prevalence rate. But youth with disability status displayed significantly higher rates on cyberbullying victimization. It also highlights the importance of studying predators and victim status on cyberbullying. It shows the importance of cyberbullying detection and prevention. Hence cyberbullying detection and prevention is an emerging need in the social network. Automatic cyberbullying detection is performed based on cyber victimization and fine-grained classification [2]. The cyberbullying events are classified into the thread, blackmail, insult, curse, defense, defamation, and encouragement based on word unigram, word bigrams, character trigram and sentiment lexicon. The features for classification are extracted from bag-of-words which give the lexical feature of words. Most challenges that found in cyberbullying detection including classification of messages, victim detection, and prevention of cyberbullying have been performed by researchers.

Automated cyberbullying detection is performed by the clustering method [3] with enhanced Naïve Bayes classifier. The entire dataset is partitioned into clusters using k-means clustering approach and eight classes are predicted based on the frequency of word with multinomial model feature vector in specific partition and probability of that word to occur in other documents. Clustering is performed based on data similarity and Naïve Bayes classifier classifies data into eight sub-categories as follows: (1) activities approach (2) communicative (3) desensitization (4) compliment (5) isolation (6) personal information (7) reframing and (8) relationship. Cyberbullying detection is performed by incorporating the weighting scheme in feature selection and using graph model to identify most active cyberbullying predator and victim in a social network [4].

The computational complexity is increased due to using multiple classifiers in the system. To overcome the above problems our proposed work concentrates on efficient cyberbullying detection and prevention in the social network. In social networks, bullying takes place at individuals communication (i.e.) user-user communication (messages) and a group communication (i.e.) user-group communication (posts). Our proposed work concentrates on bullying detection and prevention at both communications.

A. Challenges and Issues in cyberbullying detection

Although lots of work carried out by researchers in cyberbullying detection, there are still many research areas to be explored. Existing methods focused only on cyberbullying detection, the cyberbullying prevention is not considered. Issues and challenges in cyberbullying detection are discussed in [6]. The issues such as the type of cyberbullying, motive and impact of cyber victimization and prevention possibilities are discussed by the author. Life Exposure theory (LET) based cyberbullying observation is performed to protect individuals to become a target victim of cyberbullying [17]. The observation of social network is performed based on Value-Inertia-Visibility-Accessibility (VIVA) model. In LET online activities and risk online behaviors are taken in account to detect individual activities. Psychological and behavioral factors are associated with cyberbullying [18]. Victims use the internet to create a new relationship and to escape from the real world which is the weakness of victims and used by predators. But these methods are proposed only theoretically and not well suitable for real-time.

II. RELATED WORKS

Many approaches had been employed in cyberbullying detection and prevention. Most of them use machine learning approaches for cyberbullying detection.

A. Cyberbullying detection based on supervised learning approaches

Support Vector Machine (SVM) is employed for cyberbullying detection which uses the user's characteristics, their information, gender information and post harassment behavior [6]. It considers user characteristics to predict victim and predator in cyberbullying event. The features used in SVM classifier are profane words, second person pronoun, other pronouns and TF-IDF value of all words in each post. Then the dataset is divided into male and female posts and two

classifiers trained separately for that groups. But in this worktext level features are not considered and also it supports only limited dataset because it uses two classifiers. Supervised learning approach is employed to detect fake profiles in twitter by analyzing content of comments generated by both profiles [7]. Tweets, time of publication, language, geo-position, and Twitter client. Once the tweets, features and user information extracted the next step is to generate an Attribute-Relation File Format (ARFF) for classification. Based on profiles users are grouped into four groups as follows. (i) Highly Negative (HN), (ii) highly positive (HP), (iii) Positive- Negative and (iv) others. The based on these types of users the fake profile who involves in cyberbullying is detected. This method only concentrates on user characteristics which is not sufficient for fake profile detection. A framework is proposed for detecting cyberbullying contention text as well as the image is carried out using the SVM machine [14]. Abusive images are detected using Local Binary Pattern (LBP) method for feature extraction and bag-of-words model for visual vocabulary. Then SVM classifier is used to classify image is bullying or not. In text detection the dataset is preprocessed and bag-of-words model is applied for feature extraction. Then multivariate Bernoulli Naïve Bayes classifier is employed for text classification. In this method sentiment features are not taken in account and user context also not considered.

B. Cyberbullying detection based on semi-supervised learning approaches

Semi-supervised learning approach is presented for cyberbullying detection in social networks [9]. The augmented training technique is employed to extract and enlarges the dataset from streaming noisy data. The trained samples applied to fuzzy SVM classifier for classification. The augmented technique uses two classifiers (i.e.) Naïve Bayes and Stochastic Gradient Descent Classifier as based classifier and Fuzzy SVM as the main classifier. In this paper, the author assumed that the dataset always noisy, complex and uncertain. The semi-supervised approach under text streaming scenario is build using one-class ensemble learner to detect cyberbullying instances [10]. In this approach keyword based features and swear keyword based (pronoun) features are extracted to train dataset. Then feature strength for each feature is computed in all documents to train data using base classifier. Then trained data is fed into the one-class session based classifier for classification. In this method labeling is not effective since it trains dataset based only on feature set, the semantic relationship is not considered. In [12] labeling of training dataset is performed by "concept based labeling" method which involves two steps: (i) Dictionary-based sentiment lexicon and (ii) Corpus-based sentiment lexicon. Features are extracted based on the SMART method which generates 35 feature vectors for the feature set. In this approach seven emotions are identified in twitter as follows: (i) anger, (ii) embarrassment, (iii) empathy, (iv) fear, (v) pride, (vi) relief and (vii) sadness. In this approach victim and predator is not identified and there is no action taken on the predator.

C. Cyberbullying detection based on other approaches

A multi-agent framework is proposed for cyberbullying detection in Twitter social media [13]. Multi-agent system automates the process of data gathering from user activities in social network and performs an in-depth analysis of the evaluation of the social behavior of a user. Following agents are proposed in this scheme: User, Manager, Analyzer, and Monitor. The data model used in this scheme contains labels, nodes, and relationships. The functionality of each agent is implemented with independent services and allocated in a host. Every event of the user (follow/unfollow) is performed through these agents. The activities, interests of the user are analyzed by these agents. Involvement multi-agent increases the complexity of the system and the time consumption for each event is increased due to multiple agents in the system. Time series model is proposed for cyberbullying detection in which the predator's message is formulated using Singular Vector Decomposition [15]. The severity of the predator's message is annotated manually with the numeric label. Feature weighting and reducing techniques are performed to forecast the level of insult in that message using the neural network. In this process, human involvement is required and labeling manually is a time-consuming process.

III. PROPOSED MODEL

In social networks, bullying takes place at individuals communication (i.e.) user-user communication (messages) and a group communication (i.e.) user-group communication (posts). Our proposed work concentrates on bullying detection and prevention at both communications. Our proposed cyberbullying detection and prevention method designs frameworks for both individual communication and group communication.

A. Cyberbullying prevention at group communication

The collected preprocessed dataset are labeled using the label propagation method, in which labeling is performed through the similarity between data. In the label propagation method, we have to label small amount of data not all data in the dataset. The unlabeled data are labeled by finding similarity between labeled data and unlabeled data. In the dataset, feature selection is performed using Particle Swarm Optimization with Genetic Algorithm (PSO with GA) which improves the accuracy of the classifier. The training dataset classified by Neural-Support Vector Machine (Neural-SVM). The optimal parameters for SVM classifier are determined by Bayesian Optimization method which uses the prior knowledge of output in cross-validation. Finally, the user's posts (test message) are fed into classifier to categorize the message. In classifier, we employ a neural network to support multiple users' posts at a time. If the user post is classified as a bully then the server blocks that post otherwise it allows posting into the social network.

B. Cyberbullying prevention at individual communication

The index list is clustered based on string length which is calculated by regular expression. The incoming messages are preprocessed and the similarity is measured between the preprocessed message (contains only keywords) and index list using Hamming Distance which requires string length to be equal. Since we have clustered index list based on string length, the determining similarity by hamming distance is effective. If the message contains any word with distance 0 then the message is identified as bullying message and it will

be blocked and sender id is stored with count (how many bullying messages received from that user id). Otherwise, the message is displayed to the user. After receiving the message if the user finds any abusive words in that message then the index is updated with new bullying words and that particular sender id is stored. If the count of a particular user id reaches the threshold value then that particular id is reported to the server. In the server, if the number of reports on particular user id reaches threshold then that user id is marked as predator and server blocks that particular user id.

C. Algorithms developed

Algorithm 1 explains or proposed cyberbullying detection and prevention work in group communication.

Algorithm 1: Cyberbullying detection at group communication

```
Begin
  Collect dataset from social network
  Label some data in dataset manually
  Employ label propagation method
  Preprocess dataset with tokenization, stemming, stop word
  removal, polysemy and n-gram similarity
//Feature selection using PSO with GA algorithm
  Initialize feature set as particles with global best value
  For all particles
  Compute fitness value for each feature subset based on
  subset length, classification accuracy and number of
  features.
  Update global best and local best value
  Apply mutation and crossover operators
  Repeat the process until stop criterion meet
  End for
  Select optimal feature set
//Classification
  Initialize classifier parameters using Bayesian optimization
  technique.
  Fed training dataset into neural SVM classifier
  Classify dataset into bullying and non-bullying
  Classify test data from user based on training dataset
  If(user post=bullying)
     Block the post and store user Id
  Else
     Display the post to user
  End if
//Reporting process
  If(reports on particular user id>threshold)
      Block the user
  Else
      Store user id with reports counts
  End if
End
```

Algorithm 2 explains our proposed cyberbullying detection method for individual communication.

Algorithm 2: Cyberbullying detection at individual communication

```
Begin
   Construct index list with harmful words
   Cluster the index list based on string length
   Preprocess incoming message with WordNet ontology
   Find similarity between preprocessed message and index
   list using hamming distance
     If(any word in index file is present in message)
       Block the message and store user id with count
        If(message from particular user id>threshold)
          Report that user to server
     Else
       Display the message to receiver
          If(receiver find any bullying word in message)
             Update the index file with new bullying word
          Else
             End if
End
```

IV. EXPERIMENTAL SETUP AND RESULTS

A. Dataset

In this paper, we collected the datasets described below for the experiment on cyberbullying detection, which is available from the workshop on Content Analysis for the Web 2.0 [17]. The dataset contains data collected from two different social networks: Formspring.me and MySpace. Formspring.me is a discussion-based site where users broadcast their message. MySpace is a popular social networking website. Datasets were provided in the form of a text document for each conversation set between two users.

B. Evaluation Matrix

The performance of proposed approach is evaluated using the standard evaluation metrics precision and recall. Precision and recall can be calculated as below-

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Here, TP represents true positive, FP refers false positive, and FN is false negative.

The Kappa statistic (k) is a metric that is based on the difference between the observed accuracy and the Expected accuracy that the samples randomly would be expected to reveal. It is calculated as

$$k = \frac{O_a - E_a}{1 - E_a}$$

Where, O_a is the Observed accuracy and E_a is the Expected accuracy. O_a is the proportion of the test data on which the experts agree. If kappa statistics is greater than 0.75, it means the result is excellent. On the other hand, kappa statistics is less than 0.40 means the result is poor.

The Mean Absolute Error (MAE) of a model with respect to the test set is the mean of the absolute values of the

ISSN No:-2456-2165

prediction errors. These statistical values compare the true values to the estimated values. The Root Means Square Error (RMSE) is the measure of difference between the values predicted by a model and the actual values.

$$MAE = \frac{\sum_{i=1}^{n} |p_i - a_i|}{n}$$

$$RMSE = \frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}$$

Where a_i is the actual target value for test instance i, p_i is the predicted target value for test instances i and n is the number of test instances. The MAE and RMSE values range from 0 to ∞. For the evaluation of the models lower values are considered as better values.

C. Evaluation Results

In order to evaluate the proposed model, we have used dataset collected from Formspring.me and MySpace. At first, we have calculated precision and recall for both individual as well as for group-based detection model. Then we have computed the error rate.

The results obtained from the group-based model shows a precision of 0.608 and a recall of 0.586. For the group-based model, the result shows 68.6% accuracy. In the individual based cyberbullying model, it shows a precision of 0.756 and a recall of 0.685. The accuracy obtained is 72.5%.

In order to evaluate the error rate, we have computed Kappa Statistics, Mean Absolute error, Root Means Square Error. For the group-based model, the kappa statistics is 0.1691; the MAE and RMSE values are 0.4583 and 0.5021 respectively. For individual model, Kappa Statistics is 0.1821; the MAE and RMSE values are 0.5436 and 0.4321 respectively.

Table 1. Results Obtained from Group Based Cyberbullying Detection Model

Classifier	True Positive	False Positive	Precision	Recall	AUC value	Accuracy (%)
SVM	0.586	0.417	0.608	0.586	0.608	67.6

Table 2. Results Obtained from Individual Based Cyberbullying Detection Model

Classifier	True Positive	False Positive	Precision	Recall	AUC value	Accuracy (%)
SVM	0.679	0.321	0.756	0.685	0.721	72.5

Table 3. Error rate for Group-based based cyberbullying detection Model

Classifier	Kappa Statistics	Mean Absolute error	Root Means Square Error
SVM	0.1691	0.4583	0.5021

Table 4. Error rate for Individual based cyberbullying detection Model

Classifier	Kappa Statistics	Mean Absolute error	Root Means Square Error
SVM	0.1821	0.5436	0.4321

V. CONCLUSION

It can be seen from the earlier discussions that cyberbullying is one of the major issues in online social networks, mainly faced by the teenagers. Therefore, detection of cyberbullying messages and thereby users involved in such non-social activities is the need of the hour, so that effective cyberbullying prevention mechanism could be developed. Though on prima facie the cyberbullying detection seems to be a simple text classification task, it is mainly challenging due to the various complexities associated with the data available on online social networks.

We have considered an application of PSO-GA approach for effective feature engineering and to improve the overall performance of the classification systems in terms of accuracy and other parameters. Though the proposed approaches seem effective to identify spam messages, their efficacy can be further established using different real-world large-scale datasets from different categories of online social networks. Amalgamation of structural information with content-based features also seems to be one of the promising research areas. The system also uses genetic operators like crossover and mutation for optimizing the parameters and obtain precise type of cyberbullying activity which helps government or other social welfare organization to identify the cyberbullying activities in social network and to classify it as Flaming, Harassment, Racism or Terrorism and take necessary actions to prevent the users of the social network from becoming victims.

REFERENCES

[1]. Robin M. Kowalski1 and Allison Toth, "Cyberbullying among Youth with and without Disabilities", Journal of child and adolescent Trauma, Springer, pp.1-9, 2017.
[2]. Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans and V´eronique Hoste, "Detection and Fine-Grained Classification of Cyberbullying Events", Recent advances in Natural Language Processing, Ghent Academic Bibliography, 2015.
[3]. Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak and Pirom Konglerd, "Automated Cyberbullying Detection using Clustering Appearance Patterns", IEEE International conference on Knowledge and Smart Technology, Thailand, 2017.
[4]. Vinita Nahar, Xue Li and Chaoyi Pang, "An Effective Approach for Cyberbullying Detection", Communications in information science and management engineering, 2015.
[5]. Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman and Franciska de Jong, "Improving cyberbullying detection with user context", Springer, 2013.
[6]. Maral Dadvar and Franciska de Jong, "Cyberbullying Detection; A Step Toward a Safer Internet Yard", ACM conference on World Wide Web, 2012.
[7]. Patxi Galan-Garca, Jose Gaviria de la Puerta, Carlos Laorden GomezIgor Santos and Pablo Garca Bringas, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying", Springer, 2013.
[8]. Rahat Ibn Rafiq, Homa Hosseinmardi, Sabrina Arredondo Mattson, Richard Han, Qin Lv and Shivakant Mishra, "Analysis and detection of labeled cyberbullying instances in Vine,a video-based social network", Social network analysis and mining, 2016.
[9]. Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang, "Semi-supervised Learning for Cyberbullying Detection in Social Networks", Database theory and applications, Springer, pp. 160-171, 2014.
[10]. Vinita Nahar, Xue Li, Chaoyi Pang and Yang Zhang, "Cyberbullying Detection based on Text-Stream Classification", In Proc. Eleventh Australasian Data Mining Conference (AusDM13) Canberra, Australia, 2013.
[11]. Kelly Reynolds, April Kontostathis and Lynne Edwards, "Using Machine Learning to Detect Cyberbullying", IEEE Conference on Machine learning and applications, USA, 2012.
[12]. Jun-Ming Xu, Xiaojin Zhu and Amy Bellmore, "Fast Learning for Sentiment Analysis on Bullying", ACM library, 2012.
[13]. E. del Val, C. Martínez, V. Botti1, "Analyzing users' activity in online social networks over time through a multi-agent framework", Soft computing, Springer, Vol.20, Issue.11, pp. 4331-4345, 2016.
[14]. Krishna B. Kansara and Narendra M. Shekokar, " A Framework for Cyberbullying Detection in Social Network", International Journal of current Engineering and Technology, Vol.5, Issue.1, 2015.
[15]. Nektaria Potha and Manolis Maragoudakis, "Cyberbullying Detection using Time Series Modeling", IEEE conference on Data mining workshop, China, 2014.
[16]. Online: http://www2009.eprints.org/255/
[17]. Benjamaporn Kluaypa, Da-Yu Kao, "Protecting Individuals from the Suitable Target of Cyberbullying", IEEE conference on advanced communication technology, South Korea, 2017.
[18]. Raúl Navarro, Elisa Larrañaga and Santiago Yubero, "Differences between Preadolescent Victims and Non-Victims of Cyberbullying in Cyber-Relationship Motives and Coping Strategies for Handling Problems with Peers", Current phycology, Springer, 2016.