Cyber Harrasment Detection On Social Network

Vijayalakshmi M[1], Meera Hussain L[2], Aakash Balan R[3], Harrish J[4], Siva Kumar D[5]
[1]Assistant Professor
[1, 2, 3, 4, 5] CSBS, Sethu Institute of Technology

Abstract- *Online social systems have become an important part of everyday life. Peoples used to share personal content with friends circle in the social networks. Unfortunately, Social Networks provide very little support to prevent from a harassment and an unwanted messages on user walls. To overcome this issue, the Deep convolutional Neural Network algorithm is proposed in this system. This algorithm automatically block the unwanted text post message efficiently. This method is implemented in java platform with the front end Netbeans compiler and Mysql database as backend.*

Keywords- Cyber Harassment, Online Social Network, DCNN, Message Blocking.

I. INTRODUCTION

ONLINE Social Networks (OSNs) allow users to create a public or private profile, encourage sharing information and interests with other users and communicating with each other. As a result, OSNs are being used by millions of people and they are now part of our everyday life. People use OSNs to keep in touch with family, friends, and share personal information, as well as for business purposes [1]. Users of an OSN build connections with their friends, colleagues and people over time. These connections form a social graph that controls how information spreads in the social network. Although there is a dramatic increase in OSN usage – Facebook and Instagram, for instance, has now 1.55 billion monthly active users, 1.31 billion mobile users, and 1.01 billion daily users1 there are also a lot of security/privacy concerns. One of the main sources of these concerns is that OSN users establish new relationships with unknown people with the result of exposure of a huge amount of personal data.

Unfortunately, very often users are not aware of this exposure as well as the serious consequences this might have. Also, some users are less concerned about information privacy; therefore, they post more sensitive information on their profiles without specifying appropriate privacy settings and this can lead to security risks [2]. As a result, today's social networks are exposed to many types of privacy and security attacks. These attacks exploit the OSN infrastructures to collect and expose personal information about their users, by, as an example, successfully convincing them to click on specific malicious links with the aim of propagating these links in the network. These attacks can either target users personal information as well as the personal information of their friends [3]. Another widely used attack is the generation of fake profiles, which are generated with the sole purpose of spreading malicious content. In addition, there is a growing underground market on OSNs for malicious activities in that, for just a few cents, anyone can buy Facebook likes, share, Twitter and instagram followers, and fake accounts. Although many solutions, targeting one specific kind of attacks, have been recently proposed (see for instance, having a more general solution that can cope with the main privacy/security attacks that can be perpetrated using the social network graph is missing.

Therefore, this paper is to assign a risk score to each user, by taking into account both the user's activities and friendship patterns in the network [4]. The goal is to compare the behavioral patterns of users with other users in the network to find anomalous behaviors. The key idea is that the more the user behavior diverges from what it can be considered as a 'normal behavior', the more it should be considered risky (i.e., with high risk score). Following this principle requires to address two main issues. The first is the definition of a user behavioral profile able to catch those user's activities and interactions that are considered meaningful for the risk assessment. The second issue regards how to model a 'normal behavior'. In doing this, we have to consider that OSN population is really heterogeneous in observed behaviors. However, similar to real world, we expect that similar users (e.g., similar in activity level, gender, education, country, and so on) tend to follow similar rules (e.g., moral, social) with the results of similar behavioral models [5].

Based on the above principle, this article proposed a Deep convolutional Neural Network (DCNN)algorithm is proposed in this system. This algorithm automatically block the unwanted text post message efficiently. This paper is organized as follows. The second section is a summarization of related work in the literature is given. While the third section is presented the methodology of proposed work for harassment detection. Section IV presents the experimental study and shows obtained results. Finally the paper ends with a conclusion.

II. REALTED WORK

Several approaches for online Harassment and cyber Harassment detection have been proposed in the literature.
Huang et al. [5] has shown that considering the social relationships between the users on Twitter can improve the cyberbullying classification results. They build a relationship graph for each tweet and extract social features from the number of: links, edges, and nodes in relationship graph.

Zhao et al. [6] proposed a method named Embeddingenhanced Bag-of-words (EBoW) that combines: the bag of words features, latent semantic features, and bullying features based on word embedding. They used a linear support vector machine to classify the tweet as bullying or not and compared the performance of their method with the bag-of-words model, a semantics-enhanced bag of words model, Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA).

Despoina et al. [7] first provided a dataset of 9484 tweets. Then, they decomposed the features into three groups : 1) user-based features by consider basic metrics such as the number of tweets, age of account, the number of lists user subscribed to, and account verification status, 2) textbased features based on number of hashtags used, uppercase text, emoticons count, URLs count, and sentiment. 3) networkbased features focused on popularity of a user reciprocity that measures which users reciprocate the follower connections they receive from other users.

T. Marwa et al [8] presented the problem as a classification task and investigates the effectiveness of deep learning to detect online harassment in Large Human-Labeled corpus specially designed for harassment research purpose. To this end, models are considered namely Long short-term memory(LSTM) ,Bidirectional Long Short-Term Memory (BLSTM),Convolutional neural Network (CNN), and compared with other classification models. Obtained results are very encouraging.

K. Rizwan et al [9] employed machine learning and natural language processing to tackle online harassment. This study proposed a real time machine learning based algorithm which detects harassment actively and alert user to take action against it. For detection mechanism, Naïve Bayes classification is used.

M. T. Shahria et al [10] presented an approach to reduce such problems has been enlightened. A Convolutional Neural Network (CNN) based model has been built to detect sexual harassment at workplaces so that it can immediately be verdict or resolved accordingly. Due to the insufficiency of sexual harassment at workplaces dataset, we created our own dataset from social media videos for the model. Conclusively, following transfer learning methods upon three well known pre-trained models of Keras have been implied for getting better accuracy insight of model.

III. PROIPOSED METHODOLOGY

In this section, the proposed methodology is presented. the overall system architecture is shown in figure 1 which has a dataset collection, feature extraction, DCNN classifier and harassment prediction.

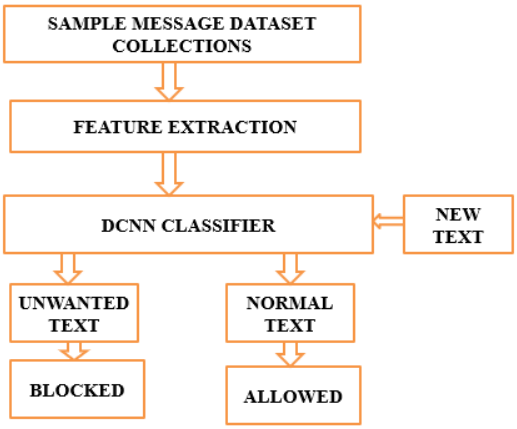


Figure 1: proposed system architecture

The dataset is collected and then the feature extraction is done with pre-processed numerical values has to convert the range of feature values between 0 and 1. Then the proposed system implemented the deep learning algorithm of convolutional neural network (DCNN) techniques that used for the dynamic blocking process. This algorithm automatically block the unwanted text post message and in the classification methods using the supervised methods. This algorithm creates the more number of category, under this category its stored more data.

This system is implemented with a several modules namely

- Network scenario
- Filtering rules
- Online setup assistant for FRS thresholds
- Blocked unwanted message

Network scenario

Given the social network scenario, creators may also be identified by exploiting information on their social graph. This implies to state conditions on type, depth and trust values

of the relationship(s) creators should be involved in order to apply them the specified rules. All these options are formalized by the notion of creator specification.

Filtering rules (FRs)

In defining the language for FRs specification, we consider three main issues that, in opinion, should affect a message filtering decision. First of all, in OSNs like in everyday life, the same message may have different meanings and relevance based on who writes it. As a consequence, FRs should allow users to state constraints on message creators. Creators on which a FR applies can be selected on the basis of several different criteria; one of the most relevant is by imposing conditions on their profile's attributes. In such a way it is, for instance, possible to define rules applying only to young creators or to creators with a given religious/political view.

Online setup assistant for FRs thresholds

As mentioned in the previous section, we address the problem of setting thresholds to filter rules, by conceiving and implementing within FW, an Online Setup Assistant (OSA) procedure. OSA presents the user with a set of messages selected from the dataset. For each message, the user tells the system the decision to accept or reject the message. The collection and processing of user decisions on an adequate set of messages distributed over all the classes allows to compute customized thresholds representing the user attitude in accepting or rejecting certain contents. Such messages are selected according to the following process. A certain amount of non neutral messages taken from a fraction of the dataset and not belonging to the training/test sets, are classified by the DCNN in order to have, for each message, the second level class membership values.

DCNN METHOD

To classify sexual harassment from a normal workspace environment, we have introduced a CNN-based model and tested our dataset. We preferred CNN over other neural networks because CNN has a unique architecture that works better with images. Like any other CNN-based model, we have a stack of different layers in our model followed by some fully connected layers. First, the model loaded the images from the dataset and stored its pixel values in the input layer. Then we applied a pair of two layers, convolutional layer and separable convolution layer, four times. First, we applied the pair with 32 filters, then with 64 filters, 128 filters, and 256 filters respectively. Each convolutional layer has a filter size of 3, "same" padding and followed by batch normalization. On the other hand, each separable convolutional layer has a filter size of 3, "valid" padding and followed by a max-pooling layer where the pool size is 2. Here we used a combination of convolutional and separable convolutional layers because unlike convolutional layers, separable convolutional layers can run faster and save a lot of computational power. Then we flattened the last layer and added 3 fully connected layers with 256, 128, and 2 neurons respectively. The activation function for the final layer was softmax and for all other layers, were Rectified Linear Unit (ReLU). To avoid overfitting, we applied a dropout of 0.3 in the convolutional layers and 0.5 in the fully connected layers.

Blocked unwanted message

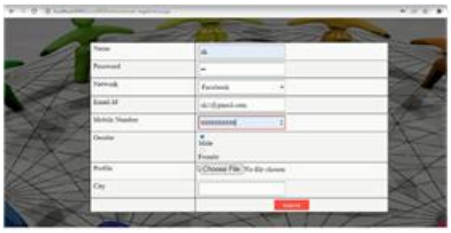
Similar to FRs, our BlacklList (BL) rules make the wall owner able to identify users to be blocked according to their profiles as well as their relationships in the OSN. Therefore, by means of a BL rule, wall owners are for example able to ban from their walls users they do not directly know (i.e., with which they have only indirect relationships), or users that are friend of a given person as they may have a bad opinion of this person. This banning can be adopted for an undetermined time period or for a specific time window. Moreover, banning criteria may also take into account users' behavior in the OSN. More precisely, among possible information denoting users' bad behavior we have focused on two main measures. The first is related to the principle that if within a given time interval a user has been inserted into a BL for several times, say greater than a given threshold, he/she might deserve to stay in the BL for another while, as his/her behavior is not improved. This principle works for those users that have been already inserted in the considered BL at least one time.

IV. RESULT AND DISCUSSION

Fundamentally, the CNN-based model is implemented and developed by training the model with the certain training portion of the dataset. Then, the testing of the model took place using the chunk of the testing labeled dataset.

Homepage

Admin login



User registration

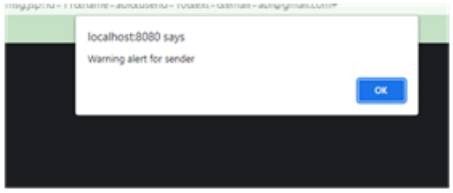
User request for security

Adding BL

Viewing friend list

Sending message to friends

Sending BL words



Warning alert

V. CONCLUSION

To eradicate this social concern the workplaces should always be monitored in a smart way. And, this can be assured if the monitoring device itself can detect the occurrence and thus act accordingly. However, our approach illuminates such a solution to this issue. If this model can be implemented on a large scale, it has the potentiality to mitigate this hideous global issue. Although our model possessed decent accuracy, the performance could be improved with large amounts of datasets. This model needs to be trained and tested upon vast datasets to build a generalized version of it. In the future, we will continue exploring the field to propose a more generalized model to predict sexual harassment. Hence, we will try to make the model lightweight also in a cost-efficient way so that it can be used to detect harassment in real-time and deployed in the security system of any workspace

REFERENCES

[1] Laleh, N., Carminati, B., & Ferrari, E. (2018). Risk Assessment in Social Networks Based on User Anomalous Behaviors. IEEE Transactions on Dependable and Secure Computing, 15(2), 295–308. doi:10.1109/tdsc.2016.2540637

[2] Gao, T., & Li, F. (2019). De-Anonymization of Dynamic Online Social Networks via Persistent Structures. ICC 2019 - 2019 IEEE International Conference on Communications (ICC). doi:10.1109/icc.2019.8761563

[3] Jones, N., Jaques, N., Pataranutaporn, P., Ghandeharioun, A., & Picard, R. (2019). Analysis of Online Suicide Risk with Document Embeddings and Latent Dirichlet Allocation. 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). doi:10.1109/aciiw.2019.8925077

[4] MITTAL, P., PAPAMANTHOU, C., AND SONG, D. Preserving link privacy in social network based systems. In NDSS (2017), ISOC.

[5] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. ACM, 2014, pp. 3–6.

[6] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in Proceedings of the 17th international conference on distributed computing and networking. ACM, 2016, p. 43.

[7] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in Proceedings of the 2017 ACM on Web Science Conference. ACM, 2017, pp. 13–22.

[8] T. Marwa, O. Salima and M. Souham, "Deep learning for online harassment detection in tweets," 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS), 2018, pp. 1-5, doi: 10.1109/PAIS.2018.8598530.

[9] K. Rizwan, S. Babar, S. Nayab and M. K. Hanif, "HarX: Real-time harassment detection tool using machine learning," 2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI), 2021, pp. 1-6, doi: 10.1109/MTICTI53925.2021.9664755.

[10] M. T. Shahria, F. Tasnim Progga, S. Ahmed and A. Arisha, "Application of Neural Networks for Detection of Sexual Harassment in Workspace," 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2021, pp. 1-4, doi: 10.1109/ICAECT49130.2021.9392429.

[11] "Sexual harassment and violence against garment workers in Bangladesh," ActionAid International, Jul. 25, 2019. [Online]. Available: https://actionaid.org/publications/2019/sexual-harassmentand-violence-against-garment-workers-bangladesh. [Accessed: Nov. 10, 2020].

[12] K. B. Clancy, L. M. Cortina, and A. R. Kirkland, "Opinion: Use science to stop sexual harassment in higher education," Proceedings of the National Academy of Sciences, Sep. 2020, vol. 117, no. 37, pp. 22614- 22618.

[13] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," Progress in Artificial Intelligence, Jun. 2020, vol. 9, no. 2, pp. 85-112.

[14] H. Cho, and S. M. Yoon, "Divide and conquer-based 1D CNN human activity recognition using test data sharpening", Sensors, Apr. 2018, vol. 18, no. 4, pp. 1055.

[15] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie, and J. Zhou, "TSE-CNN: A two-stage end-to-end CNN for human activity recognition," IEEE journal of biomedical and health informatics, Apr. 2019, vol. 224, no. 1, pp. 292-299.