Available in the SSRN eLibrary: http://ssrn.com/link/2019-CTFC.html Engineering Research Network (EngRN)
Mahatma Education Society's Transactions and Journals' Conference Proceedings ISBN 978-93-82626-27-5

DETECTION OF CYBERHECTORING ON INSTAGRAM

Tanmayee Patange, Jigyasa Singh, Aishwarya Thorve, Yadnyashree Somaraj Madhura Vyawahare
(PCE, New Panvel, India, Affiliated to University of Mumbai)

Abstract
Cyberhectoring is a growing problem affecting more than half of the population. Cyberhectoring is affecting mostly among teenagers. This problem has to be tackled which is been done by many researchers. The main goal of this is to understand and automatically detect the incidents of cyberhectoring. This paper focuses on collecting data sets of Instagram i.e. images and their associated comments. A detailed analysis of the labelled data, including a study of relationships between cyberbullying and a host of features such as profanity, temporal commenting behavior, linguistic content and image content is made. The collected data is then processed and classified using classification algorithms and is further classified into bullying and non bullying content. Using the labelled data, we further design and evaluate the performance of classifiers to automatically detect incidents if cyberhectoring.

Keywords:
Cyberhectoring, Cyberbullying, Automated detection, Machine Learning, CNN.

***Submitted on**: 31/10/2018*
***Revised on**: 15/12/2018*
***Accepted on:** 24/12/2018*
***Corresponding Author** Email: tanmaysp@student.mes.ac.in

I. INTRODUCTION

A developing assortment of examination into cyberbullying in on the web social systems has been catalyzed by increasing commonness and extending outcomes of this sort of maltreatment. To date, automated recognition of cyberbullying has focused on investigations of content in which tormenting is suspected to be available.However, given the increase in media accompanying text in online social networks, an increasing number of cyberbullying incidents are linked with photos and media content, which are often used as targets for harassment and stalking. For the purpose of detecting cyberbullying, techniques such as Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Bag Of Words, Word2Vec and OFFensiveness can also be used. We can analyze these techniques which are in association with our system. We can recognize the importance of these propelled highlights in identifying events of cyberbullying in posted remarks. We will be able to give results on assignment of pictures and subtitles themselves as potential focuses for cyberbullies. Utilizing highlights of the posted pictures and inscriptions.

II. OBJECTIVES

The objective of detection of cyberhectoring is being able to reduce the amount of bullying on Instagram. The objectives of this work is as follows:

1. To study the psychological impact on teenagers and attempt to reduce it.
2. To understand the behavior and reaction of the victim and the guilty on offensive and bullying content on social media.
3. To identify the intensity of bullying done with the help of text in caption or objects present in an image or both.

III. LITERATURE SURVEY

We have learned various techniques that can be used in association with our system.
The detection of cyber hectoring in text can be done using various algorithms like Word2Vec, OFFensiveness, Bag Of Words (BOW) and the detection of cyber hectoring in image can be done using various algorithms CNN (Convolutional Neural Network) in Caffe. It can be used to prevent sharing of harmful or offensive content by

detection. Although Warning mechanism is not provided [1].

Steps taken for detection of bullying on social media is learned. It provides guideline for the detection of cyber bullying. Although Data models are not classified into predefined categories [2].

Detection of bullied images and texts by behavioral analysis using limited classifiers is done. Prediction of onset cyberbullying incidents is also mentioned. Although It detects only one profanity word [3].

Author is using deep learning for Systematic Analysis of Cyberbullying on various SMPs .Although Limited Information about the profiles on various SMPs. Current DataSet doesn't provide information about severity of Bullying[4].

IV. METHODOLOGY

The data in the captions of an image or that particular image itself is detected if they both or anyone consists of any sensitive or offensive information. This can be done using various algorithms like Word2Vec, OFFensiveness, Bag Of Words (for text detection) and Convolutional Neural Networks CNN (for image detection). The algorithms used for text and images will be implemented using trained data sets which will be pre-defined data sets and these will be integrated for the purpose of showing connectivity or relation of captions with the images. This pre-defined data sets and integrated algorithms will be used to detect bullying content in the testing data sets. The detected text or image will appear as "Bullying Content Present" before displaying the actual image or text. Thus, we can say that the testing data sets is the input to the system and the message that displays the presence of cyber bullying is the output of the system.

The techniques which can be used for the detection of cyber hectoring in text in the caption of the image and that in the image itself is done by integrating the algorithms which can be used for individual text or individual image. The Output that can be obtained in CNN is in the form of text which is obtained from the input image having any kind of sensitive object in it, it can be detected using its algorithm. Now this text (object defined in terms of text) can be given as an Input to the Techniques used for text.

Rate of cyberhectoring amongst the teenagers is increasing with the increase in the usage of social media. The main goal of this project is to understand and automatically detect the incidents on cyberbullying. In recent times, techniques such as Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Bag Of Words, Word2Vec and OFFensiveness have also been used. We analyze these techniques which are in association with our system.

Various techniques and approaches can be proposed and developed to detect cyber bullying. The proposed approaches have focused only on the text and some only on the images. For the purpose of detection of cyber hectoring, techniques are divided into three major categories:

1) Detection of sensitive text.
2) Detection of objects in an image.
3) Detection of text and image together.

1. Detection of sensitive text

The techniques in this category which can be used to detect the bullying occuring in the text. Their short description is given below:

1a. Bag Of Words

To focus on the main topics and jargon used in the captions for images, we analyzed word frequency, using a Bag of Words model. The "Bag of words" model (BoW) is a baseline text feature wherein the given text is represented as a multiset of its words, disregarding grammar and word order. Multiplicity of words are maintained and stored as a word frequency vector. We applied standard word stemming and stop listing to reduce the dictionary size, then created a word vector in which each component represents a word in our dictionary and its value corresponds to its frequency in the text. Finally, we create a word vector, where each component represents a word in the dictionary we have generated and its value corresponds to its frequency[1].

$$BoW3 = BoW1 \cup Bow2$$

where BoW1 and BoW2 is the input of first sentence and second sentence respectively. The "union" of two documents in the bags-of-words representation is, formally, the disjoint union, summing the multiplicities of each element.

1b. Offensiveness

This technique is used for indicating that the occurrence of second person pronouns in close proximity to offensive words is highly indicative of cyberbullying, we use an "offensiveness level" (OFF) feature. We first use a parser to capture the grammatical dependencies within a sentence. Then for each word in the sentence, a word offensiveness level is calculated as the sum of its dependencies' intensity levels.

$$Os = Ow * Dj$$

where O_w = 1 if word *w* is an offensive word, and 0 otherwise. For word *w*, there are *k* word dependencies, and *d* =2 if dependent word *j* is a user identifier, *d* = 1.5 if it is an offensive word, and 1 otherwise [1].

1c. Word2Vec

Word2Vec is used for computing a continuous vector representation of individual words, commonly used to calculate word similarity or predict the co-occurrence of other words in a sentence. Here we generate a Word2Vec comment feature vector by concatenating each word's vector, based on the observation that performing simple algebraic operations on these result in similar words' vectors. For testing purposes, we apply pre-trained vectors trained on data[1].

$$M = U. \Sigma . V^T$$

where U is the topic matrix V is the image matrix and Σ is the matrix of singular values.Vector column of V sufficiently close:

$$1 < (v_i \cdot v_j) \div (||v_i|| ||v_j||) < 0.05$$

2. Detection of objects in an image

The techniques in this category are used to detect the bullying occuring in an image. Their short description is given below:

2a Convolution Neural Network (CNN)

Convolutional Neural Networks are used to detect objects present in a particular image. They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network still expresses a single differentiable score function: from the raw image pixels on one end to class

scores at the other. And they still have a loss function (e.g. SVM/Softmax) on the last (fully-connected) layer and all the tips/tricks we developed for learning regular Neural Networks still apply[1].

3. Detection Of Image and text together: In this technique, the combination of category 1 and 2 is used. The Output that is obtained in CNN is in the form of text which is obtained from the input image having any kind of sensitive object in it, it is detected using its algorithm. Now this text (object defined in terms of text) is given as an Input to the category one. So here, the output of category 2 is given as input to category 1[1].

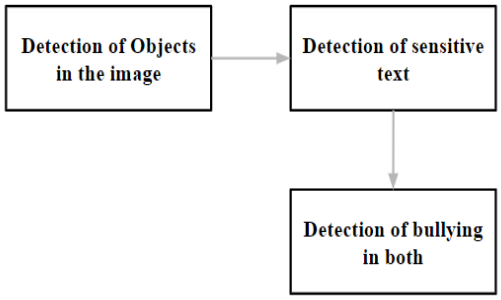

Fig. 1.1: Hybrid technique

V. SUMMARY

We have considered the discovery of cyberhectoring in photo sharing systems, with an eye on the advancement of early cautioning components for recognizing pictures powerless against assaults. With regards to photograph sharing, we have refocused this exertion on highlights of the pictures and inscriptions themselves,finding that subtitles specifically can fill in as a shockingly great indicator of future cyberhectoring for a given picture. This work is a primary advance toward creating programming apparatuses for informal organizations to screen cyberhectoring. The system we proposed will be used to detect cyber bullying in the text in the captions and in the images.

REFERENCES

1. David Miller, Haoti Zhong, Hao Li,Anna Squicciarini,Sarah Rajtmajer, Christopher Griffin, Cornelia Caragea, "Content-Driven Detection of Cyberbullying on the Instagram Social Network", IJCAI, 2016, Philadelphia, Pennsylvania.
2. Rekha Sugandhi ,Anurag Pande, Siddhant Chawla, Abhishek Agrawal, Husen Bhagat, "Methods for Detection of Cyberbullying", IEEE-2015, Marrakech, Morocco.
3. Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra, "Prediction of Cyberbullying Incidents on the Instagram Social Network", arXiv,Boulder,2015, CO 80309 USA
4. Sweta Agrawal,Amit Awekar, "Deep learning for prediction of cyberbullying across Multiple Social Media Platforms", ECIR-2018,Guwahati.
5. Richard Han1, "Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network", Boulder, CO, USA, 2015.
6. Semiu Salawu, "Approaches to Automated Detection of Cyberbullying: A Survey", IEEE-2017, Ireland.
7. Liew Choong Hon, Kasturi Dewi Varathan, "Cyberbullying Detection on Twitter", ISSN 2015, Malaysia.
8. Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, Veronique Hoste, "Automatic Detection and Prevention of Cyberbullying",arXiv 2015.
9. Krishna B. Kansara, Narendra M. Shekokar, "A

Framework for Cyberbullying Detection in Social Network", 2015

10. Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Veronique Hoste, "Automatic Detection of Cyberbullying in Social Media Text",arXiv-2018.
11. Qianjia Huang, Vivek K. Singh, Pradeep K. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis". 2014.