Detection Of Cyberbullying Using Machine Learning Technique

Ms.kavya shetty[1], Ms. Shravya[2], Ms. Sujatha kharvi[3]

Mr. Ritesh Kumar[4]

[1,2,3]*Students, Dept. Of CSE, Yenepoya Institute of Technology, Moodbidri, India-574225*
[4]*Assistant Professor, Dept. Of CSE, Yenepoya Institute of Technology, Moodbidri, India-574225*

---***---

Abstract - Cyberbullying is a serious threat to both the short and long-term well-being of social media users. We study detection of cyberbullying in photosharing networks, with an eye on developing earlywarning mechanisms for the prediction of posted images vulnerable to attacks. Cyberbullying disturbs harassment online, with alarming implications. Addressing this problem in online environments demands the ability to automatically detect cyberbullying and to identify the roles that participants assume in social interactions. It exists in different ways, and is in textual format in most social networks. Cyberbullying is the use of technology as a medium to bully someone. Cyberbullying constitutes a threat to adolescents' psychosocial wellbeing that developed alongside technological progress. A system is proposed to give a double characterization of cyberbullying. Deep learning based models have discovered ways in the identification of digital harassing occurrences, asserting that they can beat the restrictions of the ordinary models, and improve the discovery execution. However, numerous old-school models are accessible to control the incident, the need to successfully order the tormenting is as yet weak.

***Key Words*: Cyberbullying, deep learning, machine learning, content based cybercrime.**

1. INTRODUCTION

To date, people all over the world utilize internet as a tool for communication amongst them. Online tools such as social networking sites (SNSs) are the most popular socializing tool especially for adolescents as SNSs tightly integrated in their daily practices since it can be a medium for users to interact with each other withoutany limitation of time or distance.Nevertheless, SNSs can give negative consequences if users misuse them and one of the common negative activities that occurs in SNSs is cyber bullying which is the focus of this paper.

Cyber bullying involves a person doing threatening act, harassment, etc. towards another person. Meaning of cyber bullying is a group(s) or an individual(s) of peoples that adopt telecommunication advantages to intimidate other persons on the communication networks. However, most of the researchers in cyber bullying field take into account definition of cyber bullying from. According to, definition of cyber bullying formulated as "willful and repeated harm inflicted through the medium of electronic text". Cyber bullying, can takes into a few forms: flaming, harassment, denigration, impersonation, outing, boycott and cyber stalking. The most severe type of cyber bullying is flaming and the less severe is cyber stalking as stated in. Flaming occurs between two or more individuals that argue on some incidents that involve rude, offensive and vulgar language and occurred within electronic message. Flaming is the most severe type of cyber bullying because if online fight between internet's users take part, it could be difficult to recognize cyber bully and victim on that time. Harassment occurs repeatedly sending of harmful message to a victim. Denigration is posting about victim that untrue, rumors or cruel. Impersonation happens when cyber bully disguises into a target and post bad information about that particular target with intention to bullying the target. Outing occurs when cyber bully share victim's secrets or private information which can embarrassing victim. Boycott is exclude a person within social interaction in social media with a purpose. Willard mentioned cyber stalking occurs when cyber bully send harmful messages repeatedly. The cyber stalking is less severity than other categories since cyber bully (cyber stalker) could be detected directly once they send annoying messages towards victim. The main roles involved in cyber bullying occurrences are cyber bully and victim. Given the aforementioned types of cyber bullying, there are various reasons why it happens. Apart from cyber bully and victim presences, proliferation of other roles may accentuate. According, they were classified the role of bullying into eight roles. These are of bully, victim, bystander, assistant, defender, reporter, accuser and rein forcer.

2. METHODOLOGY

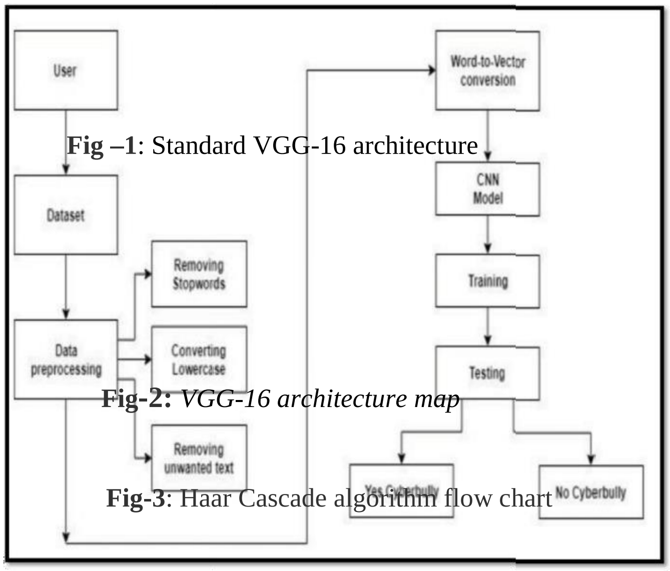


Fig –1: Standard VGG-16 architecture

Fig-2: *VGG-16 architecture map*

Fig-3: Haar Cascade algorithm flow chart

Figure: 2.1 Schematic Diagram of Detecting the Cyberbullying

2.1. DATA PRE-PROCESSING

In figure 2.1 Data pre-processing is the cleaning of the data. It is the first and the most important step required in any process. It is the conversion of a raw form of data into a required form of data for properly training the model. For instance, in the raw form the data is "You look so ugly and fat change the style", after pre-processing the data is like "look ugly fat change style". The pre- processed data takes out all the unwanted words like as, what, who, with, is, the etc. and special characters like @ () [] ?/; etc which are not required for training in the model. Data is separated into sentences and each sentence is made to make equal number of words by padding a common word which helps in the uniformity of the data. The model accepts the data in the form of a vector, the process makes the data into its lowercase format and converts that data into its vector form.

2.1.DATASET ORIGIN

Formspring.me this site especially prone to cyber bullying is the option for anonymity. To obtain this data, we crawled a subset of the Formspring.me site and extracted information from the sites of 18,554 users. The users we selected were chosen randomly The number of questions per user ranged in size from 1 post to over 1000 posts. We also collected the profile information for each user. Labeling the data we ensured that there was no overlap between the two sets of files. We used the same procedure to identify class labels both the training and the testing sets.

2.1.DEVELOPING FEATURES FOR INPUT

we wanted to develop a model based on textual features. This section describes the identification and extraction of features from each Form spring post. We were determined to avoid a bag-of- words approach for several reasons. First, the feature space with a bag-of-words approach is very large. Second, we wanted to be able to reproduce the model in code, and having each term as a feature would make that impractical. Third, we wanted to be able to understand why a post was considered as containing cyber bullying as this will inform the development of a communicative model for Cyber bullying detection .

2.1.LEARNING THE MODEL

Weka is a software suite for machine learning that creates models using a wide variety of well- known algorithms When working with decision trees, it is important to consider the size of the tree that is generated, as well as the accuracy of the model

JRIP: JRIP is a rule based algorithm that creates a broad rule set then repeatedly reduces the rule set until it has created the smallest rule set that retains the same success rate.

IBK: The instance-based (IBK) algorithm implemented in Weka is a k-nearest neighbor approach.

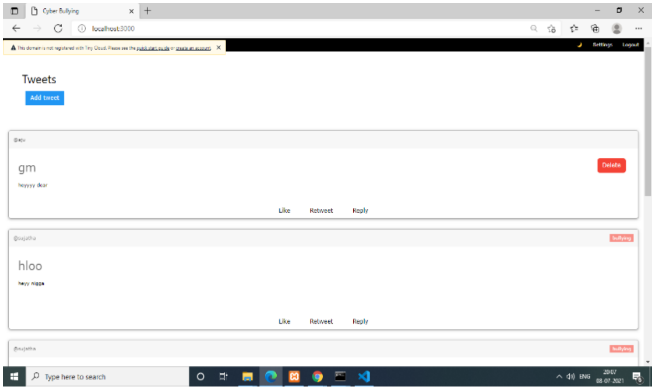
SMO: We wanted to use a support vector machine algorithm for testing also. In Section V we show that SMO was the least successful algorithm for our experiments.

2.1.OUTCOME OF PROJECT

We compare the results using the NUM training set to the NORM training set .These tables report the recall for identifying cyber bully using 10- fold cross validation. We see that NORM training set generally out performs the NUM training set for all repetitions and for all algorithms, with the exception of the SMO algorithm. We also see that as the weighting of positive instances increases, the NORM success rates are slightly higher than the NUM success rates. We conclude from these results that the percentage of "bad" words in a post is more indicative of cyber bullying than a simple count.

3. RESULTS

We conclude from these results that the percentage of "bad" words in a post is more indicative of cyberbullying than a simple count. The first evaluation we present is on the ability of the models to detect cyberbullying content in messages. Another comparison is between the labeling strategies and We also discussthe textual categories.



Cyber Bullying

localhost:3000

This domain is not registered with Tiny Cloud. Please see the quick start guide or create an account.

Tweets

@raju

gm

heyyyy dear

@sujatha

hloo

heyy nigga

@sujatha

Fig 2.Result of cyberbullying

And in this if we know that it is a bad comments or a word it is named as bullying and we are taking action on that and if it is not a bad one we consider that as a not bullying statement.

4. CONCLUSION AND FUTURE WORK

In this paper cyberbullying is one of the most critical inter-net crimes, and research has demonstrated its critical impact on the victims. In this paper, a novel idea is proposed where any cyberbullying tweet remarks are identified as cyberbullying comment or not. The system uses a accurate method of CNN implementation using keras and helps in achieving precise results. The proposed systemcan be used by government or organiztion- parents, guardians, institutions, policy makers and enforcement bodies. This can help the users by preventing them for becoming victims to this harsh consequence of cyberbullying.Since the domain of online bullying is a never-ending process, it is required that the methodologies require constant upgrading and updating to the current situation. Our proposed methodology, can come useful in handling crises and can even be enhanced to provide full-time support. Finally, it can even prevent a potential crisis. Some of the salient enhancements which can be included In this we have studied detection of cyberbullying in bully words . warning mechanism to identifying bully words. In the context of words we have refocused this effort on features of bully words.This work is a foundational step toward developing software tools for social networks to monitor cyberbullying.

REFERENCES

[1] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, "Towards understanding cyberbullying behavior in a semi-anonymous social network," in International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2014.

[2] K. E. Bauman and E. S. T., "On the importance of peer influence for adolescent drug use: commonly neglected considerations," Addiction, vol. 91, no. 2, 1996.

[3] C. J. Ferguson, C. S. Miguel, and R. D. Hartley, "A multivariate analysis of youth violence and aggression: The influence of family, peers, depression, and media violence," The Journal of Pediatrics, vol. 155, no. 6, 2009.

[4] R. Slonje, P. K. Smith, and A. Frisen, "The nature of cyberbullying, ´ and strategies for prevention," Computers in Human Behavior, vol. 29, no. 1, 2013.

[5] Y. Chen, L. Zhang, A. Michelony, and Y. Zhang, "4is of social bully filtering: identity, inference, influence, and intervention," in CIKM, 2012.

[6] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text." in International Conference On Web And Social Media (ICWSM), 2014.

[7] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in International Conference on Machine Learning (ICML), 2014.

[8] R. Reh ˇ u˚ˇrek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in LREC Workshop on New Challenges for NLP Frameworks, 2010.

[9] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Conference on Empirical Methods on Natural Language Processing (EMNLP), 2014.

[10] O. Owoputi, B. O’Connor, C. Dyer, K. Gimpel, and N. Schneider, “Part-of-speech tagging for twitter: Word clusters and other advances,” Technical Report, Machine Learning Department. CMU-ML-12-107., 2012.

[11] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, 1995.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, 2011.

[13] F. Giglietto, L. Rossi, and D. Bennato, “The open laboratory: Limits and possibilities of using facebook, twitter, and youtube as a research data source,” Journal of Technology in Human Services, vol. 30, no. 3-4, 2012.

[14] N. O. Hodas, F. Kooti, and K. Lerman, “Friendship paradox redux: Your friends are more interesting than you.” International Conference on Web and Social Media (ICWSM), 2013.