Review on Detection of Cyberbullying using Machine Learning

Prof.Swati R. Khokale, Vikrant Gujrathi, Rutik Thakur, Apoorva Mhalas, Shivam Kushwaha

***Abstract*- Internet has touched every aspect of human life, bringing ease in connecting people around the globe and has made information available to huge strata of the society on a click of a button. With advancement, came unforeseen banes of cyber offences. Cyber bullying is a form of electronic communication, which harms the reputation or privacy of an individual, or threatens, or harasses, leaving a long lasting impact. Although it has been an issue for many years, the recognition of its impact on young people has recently increased. Through machine learning, we can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyberbullying content. We comprehensively review cyberbullying prediction models and identify the main issues related to the construction of cyberbullying prediction models. This paper provides insights on the overall process for cyberbullying detection and most importantly overviews the methodology. Though data collection and feature engineering process has been elaborated, yet most of the emphasis is on feature selection algorithms and then using various machine-learning algorithms for prediction of cyberbullying behaviours. Supervised Learning Algorithm, which gives highest accuracy, was used for the detection of cyber bullying activity over the internet.**

Keywords – Cyberbullying, Machine Learning, Supervised Learning Algorithm.

1. Introduction

Cyberbullying is bullying that takes place in the digital world and can occur on various forums where people view, participate in, or share content. Bullying or harassment can be identified as a repeated behaviour and an intent to harm others. Examples of cyberbullying include derogatory, threatening or harassing messages, pictures, audios and videos. Once such content is posted, they live perpetually in the cyber world. Due to the ease of posting such content, cyberbullying empowers a bully to humiliate and hurt the victim in online communities without ever getting recognized. Furthermore, the fear of getting punished or being a social pariah stops victims and bystanders from reporting incidents. Bullying is most common among kids and youngsters. The effects of cyberbullying are often devastating on such population and the result in victims having lower self-esteem. Bullying can also cause many negative effects such as impacts on the mental and physical health, depression and anxiety, and can lead to suicidal tendencies. As a consequence of such cyberbullying behaviour, the victims may miss or even drop out of school. Hence, cyberbullying is an epidemic that needs to be controlled quickly and effectively. The Cyberbullying Research Centre defines Cyberbullying as "the act of using the Internet, cell phones, video games or other technology gadgets to send, text or post content to hurt or embarrass another person". They emphasize that the process must be wilful i.e. the action is intended to do harm, and repeated i.e. a single action is not considered bullying. Cyberbullies carefully single out and humiliate their target on social media publicly, oftentimes embarrassing them or sending hurtful messages. What makes cyberbullying so dangerous is that it gives bullies the ability to embarrass anyone they want in public at any hour of the day. According to 'Tweens, Teens and Technology 2014 Report' by McAfee, 50% of Indian youth have had some experience with cyberbullying. According to a survey, it has been identified that a significant number of suicides have been committed by teens who were exposed to cyberbullying. Teens feel demoralized and get frustrated when they encounter such cyber aggressive comments which act as a barrier for participation and socializing. Most networking sites today prohibit the use of offensive and insulting comments. But this partially being carried out and filtered to a limited extent. As there are enormous amounts of data available it is impossible to take help of human moderators to manually flag each insulting and offensive comment. Thus

automatic classifiers that is fast and effective to detect such type of comment is required which will further reduce cyberbullying. To overcome this problem Machine Learning Algorithm, prove to be very helpful, both in detection and in reporting.

2. Literature Survey

According to survey [1], Age and gender of the bully also plays an important part in prediction, however most users have incomplete or fake profile details, which affects accuracy of the model. SVM-based cyberbullying model is more reliable but not as accurate as rule-based Jrip. However, the SVM-based cyberbullying model is more accurate than NB and tree-based J48.

According to the survey conducted by Amanpreet Singh and Maninder Kaur [2], Majority of the work involved supervised learning techniques in the area of detection of bully content, and that SVM and naïve Bayes are the commonly used classification techniques for attaining effective outcomes for the detection of bully content on social networks. Most of the research papers targeted in the present study are primarily based on bullying as text on the contrary videos and images can also be used as an online system of harassment, and their effects maybe even more harmful.

Rohit Pawar [3] suggested that most of the work of Detecting cyberbullying is done in single language. So it is also necessary to consider detecting cyberbullying in multiple languages. Hence, they built a model to detect cyberbullying in multilingual languages i.e., Hindi and Marathi. According to [3], When both words and emoticons in the text messages are considered as features, cyberbully detection improves. Logistics Regression (LR) outperforms SGD and MNB for the binary classification problem and continues to work better even when the data size grows.

Wan Noor Hamiza Wan Ali [4] suggested that Features can be categorized into mainly four categories i.e., content, sentiment, user and network-based features. Most of the time Profanity feature, BoW features, Latent Semantic features and bullying features were used by the researchers to get the higher precision, Recall and F-measure value. According to [4], Out of all other Classification algorithm, SVM is most frequently used by all the researchers because it is suitable for high-skew text classification such as to detect Cyber bullying using content-based feature.

In "An Empirical Study and Analysis of the Machine Learning Algorithms Used in Detecting Cyberbullying in Social Media" [5], A larger dataset from a widely popular platform Twitter were used to compare and determine which supervised machine learning approach performs best by fine-tuning our data pre-processing techniques to provide higher accuracy. Automatic approach of Labeling thousands of hundreds of tweets were adopted and tweets were labeled as positive and negative tweets based on no. of positive words and negative words present in tweet. After labeling each tweets with their polarities, the training set is stored in a NoSQL non-relational database named Mongodb. Naïve bayes and SVM algorithm were used as a classification model out of which SVM algorithm gave higher accuracy than Naïve Bayes algorithm.
SVM was correctly able to predict sentiment with an accuracy of 89.54% and Naive Bayes having an accuracy of 73.03%.

According to the survey conducted by Semiu Salawu [6], In most of the cases Cyber Bullying is defined as "Repetitive action" of harassing someone online so if an act has to be repeated before being considered Cyber bullying, then the detection system must maintain a history of previous messages and perhaps introduce the timestamps of messages exchanged as a feature to satisfy the "repeated acts" criterion. Hence, for this System a rule can be created to only flag a user as a Cyber bully if he or she exceeds a threshold of bullying messages over a set period of time.
Mostly all Cyber Bullying Detection models are only limited to detection of Cyber Bullying detection without recourse to the preventive actions to be taken once bullying is detected. So it is necessary to take preventive action into consideration while building Cyber Bullying Detection model. There are many approaches are available for building Cyber Bullying Model such as Supervised Learning approach, Lexicon-Based Approach, Rules-Based Approaches, Mixed-Initiative Approaches. Out of all available Techniques, Supervised Learning approach is most widely used for building Cyber Bullying model.

According to Batoul Haider [7], Cyber bullying can be categorized into many categories such as Flaming, Masquerade, Disintegration, Impersonation, Harassment etc. and depending upon severity of bullying Culprit can be punished. In [7], Datasets were built by scrapping comments from OSN platforms by using some already available data management platforms such as HP

Autonomy products- Intelligent Data Operating Layer (IDOL). It also concludes that work related to Arabic language is scarce due to its complex morphological nature of Arabic hence, they proposed a Multilingual Cyber Bullying Detection machine for detecting Cyber bullying done in Pure English, Pure Arabic and Mixed environment such as Arabish and Arabizi texts on OSN platforms such as Facebook, Twitter etc.

Homa Hosseinmardi [8] Found that Most of the previous work focuses on classifying the data as bullying content or not rather than predicting. In their proposed system, Predictions are made using previous media posts such as images or videos that might prove to be an onset of potential cyberbullying in the multi-modal online social network. Logistic regression classifier was used to train a predictor with the forward feature selection approach which produced 98% accurate results.

Mohammed Ali Al-garadi [9] suggested that Features such as age and gender are inadequate and are not extensive or discriminative enough to analyze the dynamics of online social network data. It also suggested that including the new acronyms and words can improve the performance of a cyberbullying classifier and maybe used as the first clue in detecting cyberbullying engagement. Random forest using SMOTE alone showed the best AUC (0.943) and f-measure (0.936).

A. Mangaonkar [10] suggested that all the obscene text on social media may not be cyberbullying, therefore, care must be taken in deciding whether a tweet constitutes cyberbullying or not, even if it contains obscenities. Author also found out that Logistic Regression does perform a little better than the generative model (Naive Bayes) for both balanced and unbalanced datasets, whereas SVM is unable to perform satisfactorily And the usage of a collaborative cyberbully detection paradigm yields better results than the sequential paradigm and also proves to be more efficient time-wise.

In “Automatic detection and prevention of cyberbullying” [11], A shallow error analysis revealed that implicit realizations of cyberbullying are fairly hard to recognize, as they are devoid of lexical cues such as profanity. Therefore, use of more advanced features (e.g., syntactic patterns, semantic information) in addition to lexical features is required. The data scarcity of large discrepancies in performance are presumably due to the extent to which a category is lexicalized. SVM combined with fine grain sampling produces optimum results.

Van Royen [12] focuses on the misuse of pictures and considers it as high priority and suggests strategies for automatic detection such as a cross-media detection approach focusing both on visual and textual Cyber bullying. The model provides Automatic warnings before uploading and other alternatives for automatic detection to allow for prevention of harm should be examined as well. A response grading system was developed, through which cases were classified according to assessments of severity and subsequently linked to appropriate follow-up measures. also the importance of automatic monitoring implementation, evaluation of the follow-up strategies is also discussed.

Chavan [13] have devised methods to detect cyberbullying using supervised learning techniques. They have presented two new hypotheses for feature extraction to detect offensive comments directed towards peers which are perceived more negatively and result in cyberbullying. Their initial experiments show that using features from their hypotheses in addition to traditional feature extraction techniques like TF-IDF and N-gram increases the accuracy of the system.

Ying Chen [14], has put forth that existing methodologies do not involve either hand authorizing syntactic rules in identifying name-calling harassments or incorporation of users writing styles, structure and any specific cyber bullying content as a feature to predict potential bulling content. They found out that the study of user’s profile, writing style and structure can give better insights towards the motive of the users and to thus predict potential cyber bullying threats. They concluded that Lexical Syntactical Feature (LSF) framework performs significantly better than existing methods in offensive content detection. It achieves precision of 98.24% and recall of 94.34% in sentence offensive detection, as well as precision of 77.9% and recall of 77.8% in user offensive detection

In "Using Machine Learning to Detect Cyberbullying" [15] the dataset was Extracted from formspring.me and to obtain these datasets they crawled data a subset of the Formspring.me site and extracted information from the sites of 18,554 users. Amazon’s Mechanical Turk service was used to determine the labels for their truth sets. Mechanical Turk is an online marketplace that

allows requestors to post tasks (called HITs) which are then completed by paid workers. A variety of lexical features were extracted from the Formspring.me post data, and several data mining algorithms that are available in the Weka toolkit were used to develop a model for the detection of cyberbullying while comparing their results, they were focused more on recall than on precision. Number of "bad" words (NUM) and the density of "bad" words (NORM) were used as features for input to the learning tool. Therefore, they extracted two different training sets, one containing the count information, and one containing normalized information. Using these features they were able to attain the accuracy of 78.5% in detection of Cyber Bullying Content.

3. Proposed System

The main objective of the proposed system is to detect cyber bullying that occurs on various social media platforms. For the Same we collected the data from different sources such as Twitter, Youtube, Reddit, Wikipedia, etc. The data on these sites is present in the form of tweets and comments from different online users. Next Step is Data preprocessing which means we will process our data before feeding it into our machine. In Data preprocessing, we first remove any irrelevant data from our dataset, then we treat outliers and lastly we handle any missing data which may be present in our dataset. We used 70% of the dataset for training and 30% for testing purpose. A better practice is to use 60% for training, 20% for cross validation, 20% for testing. Due to having such a large dataset both for training and testing, we did not find the necessity to use K-folds cross validation technique. After splitting our dataset into training and testing dataset, we will train our machine using this training set which will help our classifier algorithm in learning to classify the data into positive and negative tweets/comments. After the machine has been trained, the testing dataset will be used to test the accuracy of our machine learning model. For evaluation, we calculated the accuracy of the classifier, precision, recall and f-score of the positive, negative and neutral tweets. Precision and recall are the metrics used to determine classifier output quality. Precision is the measure of how relevant the results are and recall is the measure of how many relevant results are returned. F-score is the average of both the precision and recall.

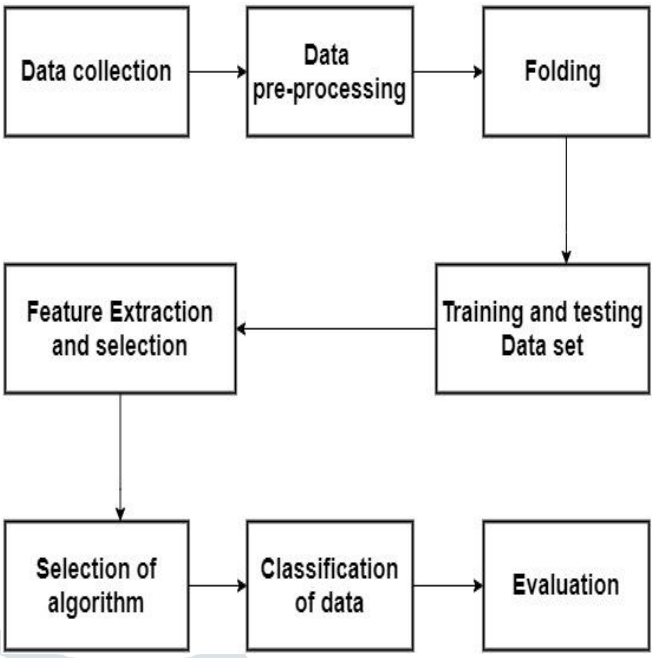

Fig. Proposed system

4. Conclusion

According to our survey, we came to a conclusion that most of the work previously done were detecting Cyber bulling on single platform so we're trying to target multiple social media platforms rather than focusing on single platform and as Supervised learning is most widely used technique for detecting Cyber bullying hence we're also choosing the same technique for dealing with this problem. Out of all classification algorithm Support vector machine (SVM) classifier, Logistic Regression, Naïve Bayes algorithm and XGboost classifier are the most efficient one and for evaluating accuracy of these algorithms we are using Precision, Recall and F-measure evaluation metrics.

References

[1] Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza, G.; Nweke, H.F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H.A.; Gani, A. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. IEEE Access 2019, 7, 70701–70718.

[2] Singh, Amanpreet & Kaur, Maninder. (2019). Content-based Cybercrime Detection: A Concise Review. 8. 1193-1207.

[3] R. Pawar and R. R. Raje, "Multilingual Cyberbullying Detection System," 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA, 2019, pp.

040-044, doi: 10.1109/EIT.2019.8833846.
[4] W. N. Hamiza Wan Ali, M. Mohd and F. Fauzi, "Cyberbullying Detection: An Overview," 2018 Cyber Resilience Conference (CRC), Putrajaya, Malaysia, 2018, pp. 1-3, doi: 10.1109/CR.2018.8626869.
[5] M. Sintaha and M. Mostakim, "An Empirical Study and Analysis of the Machine Learning Algorithms Used in Detecting Cyberbullying in Social Media," 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2018, pp. 1-6, doi: 10.1109/ICCITECHN.2018.8631958.
[6] S. Salawu, Y. He and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey," in IEEE Transactions on Affective Computing, vol. 11, no. 1, pp. 3-24, 1 Jan.-March 2020, doi: 10.1109/TAFFC.2017.2761757.
[7] B. Haidar, M. Chamoun and F. Yamout, "Cyberbullying Detection: A Survey on Multilingual Techniques," 2016 European Modelling Symposium (EMS), Pisa, 2016, pp. 165-171, doi: 10.1109/EMS.2016.037.
[8] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016, pp. 186-192, doi: 10.1109/ASONAM.2016.7752233.
[9] Al-garadi Mohammed Ali, Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior, 63, 433–443. doi:10.1016/j.chb.2016.05.051
[10] A. Mangaonkar, A. Hayrapetian and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015 IEEE International Conference on Electro/Information Technology (EIT), Dekalb, IL, USA, 2015, pp. 611-616, doi: 10.1109/EIT.2015.7293405.
[11] Van Hee, Cynthia & Lefever, Els & Verhoeven, Ben & Mennes, Julie & Desmet, Bart & Pauw, Guy & Daelemans, Walter & Hoste, Véronique. (2015). Automatic detection and prevention of cyberbullying.
[12] Van Royen, K., Poels, K., Daelemans, W., & Vandebosch, H. (2015). Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics, 32(1), 89–97. doi:10.1016/j.tele.2014.04.0
[13] V. S. Chavan and Shylaja S S, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 2015, pp. 2354-2358, doi: 10.1109/ICACCI.2015.7275970.
[14] Y. Chen, Y. Zhou, S. Zhu and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, Amsterdam, Netherlands, 2012, pp. 71-80, doi: 10.1109/SocialCom-PASSAT.2012.55.
[15] K. Reynolds, A. Kontostathis and L. Edwards, "Using Machine Learning to Detect Cyberbullying," 2011 10th International Conference on Machine Learning and Applications and Workshops, Honolulu, HI, USA, 2011, pp. 241-244, doi: 10.1109/ICMLA.2011.152.