Cyberbullying Detection in Social Networks using Machine Learning Models

Nivethitha R[1] and Dr.L.S.Jayashree[2]
{20mz05@psgtech.ac.in[1], lsj.cse@psgtech.ac.in[2]}

PG scholar, Dept of Computer Science and Engineering, PSG College of Technology, Coimbatore, India[1]., Professor, Dept of Computer Science and Engineering, PSG College of Technology, Coimbatore, India[2]

Abstract. Innovation is increasing rapidly in this dynamic world. Innovation gives rise to communication. Cyberbullying happens through communication. Cyberbullying is an act of targeting a person or a group of people to insult them in the form of electronic messages in social media and the depression forces them to commit suicide. It is a very big problem which should be destroyed. These activities should be stopped. Social media is a major environment to create negative comments and it is the birthplace for cyberbullying. Nearly 80% of young people in the world commit suicide because of depression from cyberbullying. This paper has articulated all the points and ideas that are used in the detection of cyberbullying in social media and also its impacts.

Keywords: Cyberbullying, text classification, python, machine learning, offensive language detection.

1 Introduction

Technology is everything in today's world. Everyone depend upon technology for everything. Cyberbullying is intentionally insulting, threatening people in the form of electronic messages. Cyberbullying act disturbs people's joy, happiness and stops them to lead a happy life. It tortures lot of people psychologically and emotionally in the world to commit suicide. Children and teens are mostly affected by this act and the awareness should be spread among people to stop this and to save people's lives. They should feel free to say and talk about this to their parents. Group of people from age 12 to 18 stand as target. As the number of social media users increases, the act of cyberbullying has also been increased. Cyberbullying can be prevented only when parents monitor their children when they use electronic devices like laptops, mobile phones and tablets. Children should be restricted to use these devices and should start moving to do offline activities. Not all parents will have knowledge of what is cyberbullying and how to stop them. Insteadteachers should start teaching about them in schools. Since the person who bully others is physically not known because of social media where people may have fake accounts, the activities of such bully goes on increasing. The rise of social media gave birth to the new problem of cyberbullying. Lot of researchers are working in this field to stop this illegal activity. One-third of the cyberbullying affects people only by social media so children should use them carefully. cyberbullying detection helps people to use social media safely. Cyberbullying detectionwill stop or atleast reduce these activities. Researchers are working hard in this field but even some

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314577

of the problems are not taken care like mixing of languages where people combine two or three languages to bully others. All the countries across the world is suffering from Cyberbullying and India stands first with 37%. If a particular person is bullied for many times then the person finds himself difficult to communicate with other people and also to find new friends. If a person is bullied, they should have the courage to open up about this to their parents or teachers so that they can feel free and they will be able to face it for the next time when it happens. A lot of people undergo severe depression and mental stress after they are being cyberbullied. Not only teenagers are affected by cyberbullying. Anti-bullying laws should be taught to everyone to face their lives boldly. There are several ways ,plans and ideas to prevent the persons who involved in such illegal cyberbullying activities.

2 Literature Survey

This section reviews some of the contents from related works. Several algorithms were used in cyberbullying detection.

Nideeksha et al.[1] suggested the detection of cyberbullying using machine learning models. The dataset is collected first. After collecting the dataset, it is given to the Naïve Bayes classifier for classification and prediction. The model created is then stored in a pickle file which helps us later to reload the file. In discord platform, when a text is received, the data is pre-processed and the feature extraction is done by TextBlob where polarity and subjectivity values are checked for sentiment Analysis. If it is a cyberbullying message, the bully is warned for the action to the channel. When user gives a Thumbs up reaction, the particular message is deleted. Before that, it is given to the classifier for training with positive label. By increasing the usage of the model, the accuracy can further be increased. Detection of cyberbullying in audios and videos has always been a major problem.

John Hani et al. [2] proposed the detection of cyberbullyingin social media. The dataset is first pre-processed using tokenization and by correcting the words. Feature extraction is done by Term Frequency-Inverse Document Frequency where score is assigned based on the similarity of words and by n-grams where n stands from 1 to 5. Based on the n-grams given to words, the accuracy also differs for both the models. It is divided into 80% and 20% for training and testing. Then it is given to the two classifier models Support Vector Machines and Neural Networks. Neural Networks has high accuracy than the other model because of the number of layers present. Duration of development of Neural Network is high and also NN is computationally expensive.

Alwin T. Aind et al. [3] suggested the Q-Bully which is a reinforcement learning model for Cyberbullying detection.Dataset is pre-processed first. New approach Q-Bully is proposed. The algorithm used is Q-learning reinforcement. Each word is considered as a state and the action performed to move to the other state are offensive action and non-offensive action. It goes through all the words in a sentence and gives positive reward if it goes in the right direction and negative reward if goes in wrong direction. The rewards are stored in the Q-Table. One disadvantage with this model is that it does not give increase in accuracy for large datasets as compared to small datasets. The accuracy achieved is 93% for ten thousand comments and 89.5% for the whole dataset. Future work of this paper can be extended by implementing Deep Q Learning by which the model can further be optimized.

K.Nalini et al. [4] proposed the Naïve Bayes classifier model to predict cyberbullying with the use of LDA. The dataset from twitter is used. It is first pre-processed where the

retweets, stop words and URL are removed. The key terms are identified by LDA which gives score based on the severity of the abusive words. It is given to Naïve Bayes,J48 and KNN classifiers where the highest accuracy scored is 70% for the Naïve Bayes classifier and it is also checked with the evaluation parameters like precision, recall, true positive and false positive.

WalisaRomsaiyud et al [5] suggested the cluster patterns for the detection of cyberbullying. The dataset is created and then it is divided into two clusters namely abusive and non-abusive messages by K-Means clustering technique. Feature extraction is done from abusive message and is classified into eight categories like communicative, personal information, compliment etc. and the data is given to the Naïve Bayes classifier model where set of features is present for eight categories for prediction of the abusive message. This particular model helps in increasing the accuracy and also the reliability of the system. One disadvantage is that the cluster process used here does not work in parallel.

Vikas S Chavan et al. [6] proposed the cyberbullying detection by machine learning in social media. The dataset undergoes normalization techniques where unwanted strings are removed and then the words are corrected. N-gram models and Term Frequency- Inverse Document Frequency score are used for standard feature extraction. Additional feature extraction like skip-grams and capturing the pronouns are used. It is then fed into the support vector machine and logistic regression classifying models. Before the additional feature extraction technique was applied, the accuracy value was quite low. The best accuracy is achieved for logistic regression which is 82% before additional feature extraction and increased to 86% after additional feature extraction. Recall, precision were used for evaluation metrics.

Vijay Banerjee et al. [7] suggested the uses of deep learning model for cyberbullying detection. It uses Convolutional Neural Network where the dataset is given to the input layer, the embedding layer converts them into vector representation for understandable by the machine. Convolutional layer makes use of Rectified Linear Unit. For feature reduction, maxpooling layer is used. Dropout layer is used after the embedding layer and before the output layer. Value of the dropout layer is 0.5. It is then given to the fully connected layer or the dense layer which is the output layer and the last layer for classification purpose. The accuracy scored in this model is 93.97%. The training process may take a lot of time if the system does not contain a better GPU.

MonirahA.Al-Ajlan et al. [8] suggested the detection of cyberbullying in twitter by deep learning.Twitter dataset is collected and then data is cleaned as part of pre-processing techniques to avoid errors. Then it is fed to the Convolutional Neural Networks where metaheuristic optimization like ant colony optimization or genetic algorithm from evolutionary computing technique is also given to find the similar values since it reduces the manual work to set the parameter values. CNN then classifies the text whether it is cyber-bullyed or not. One good advantage is that feature extraction and feature selection is eliminated in this model.

Cyberbullying detection is compared among all the machine learning models which is proposed by Amgad Muneer et al. [9]. The dataset is first collected and then pre-processed to remove any noise or duplicates present using removal of stopwords and stemming. Features are extracted using Term Frequency- Inverse Document Frequency and Word2Vec and the data is split into training and testing data for feeding them into the machine learning models. One good aspect is that Light Gradient Boosting Machine has the highest accuracy among all models evaluation metrics like precision, recall, F1 Score and additionally prediction time is

also taken into account. Best prediction time is given for Logistic regression. The disadvantage is that more feature extraction techniques can be used to increase the accuracy.

Noviantho et al. [10] suggested the detection of cyberbullying. The dataset which contains text conversation is pre-processed and it is also balanced on the classification of two classes, four classes and ten classes based on the severity of abusive words present and then N-gram is generated for 1 to 5. Feature extraction is done using Term Frequency -Inverse Document Frequency. It is given to the classification models like Support Vector Machine and Naïve Bayes and compared with all the N-gram models. SVM poly kernels with 5-grams has scored the highest accuracy among all the Support Vector Machine kernels. One disadvantage is that classifying the text conversation seems to be quite difficult because of the presence of shortened words which this model cannot recognize.

Eloi Brassard – Gourdeau et al. [11] proposed the detection of toxicity of a message. The presence of toxic word is found using Sentiment detection which makes use of many lexicons like SentiWordNet, subjectivity clues etc. The message is first pre-processed for the presence of any negation words and then pre-processed for the presence of Sentiment carrying idioms. For example: two words when seen as individual may give positive meaning but when seen as idiom, may give negative meaning. It is then given to the deep Neural Network for classification. When Sentiment detection is also added, the accuracy of the system is highly increased to find the toxicity. Major limitation is that it focuses only on single line messages and not on one whole message.

Batoul Haidar et al. [12] proposed cyberbullying detectionin Arabic language by the Deep learning model. A lot of work is done for cyberbullying detection in English but it is also quite important to find in Arabic language since lot of young people uses them. Whole dataset and small portion of the dataset is taken, they undergo tokenization where non-Arabic words are also removed and after word embedding it is split into 80% and 20% as training dataset and testing dataset then given to the Feed Forward Neural Network where it is compared with all the hidden layers starting from 3 to 9. The accuracy for small dataset was found to be 66.67% which is considered to be very low among all the deep learning models so it is trained and tested with the whole dataset. Then the accuracy scored was 91.17%. It is quite clear that the accuracy scored is low for small dataset and it drastically increases for large dataset which is a limitation.

Various textual features are identified for the detection of cyberbullying is proposed by Jianwei Zhang et al. [13]. Twitter dataset is collected where the labels are positive for cyberbullying words and negative for non-cyberbullying words. The data us pre-processed and then features are extracted using many techniques like Word2Vec, N-grams, Doc2Vec, specific characteristics of twitter. After extraction, the machine learning models like Support vector machine, logistic regression, random forest, decision tree are generated. Then by using accuracy, precision and recall as the evaluation parameters the best model is found which is logistic regression here. One drawback found is that the text used here is limited and large amount of text can be obtained using Bootstrapping method. This method refers to extracting the bullying words from the existing tweets iteratively in order to create new tweets.

AzaldenAlakrot et al. [14] suggested the detection of abusive language in Arabic. Youtube dataset is taken which consists of both offensive and inoffensive messages. The dataset is first pre-processed where tokenization is done and each word is taken as individual tokens. Next filtering is done to remove all the punctuation marks and the stopwords are also removed. Normalization is done to avoid any misspelling and extra normalization is done to improve the performance of the system. Data is divided into training and testing data and it is fed into the Support Vector Machine where it is compared with after pre-processing, before

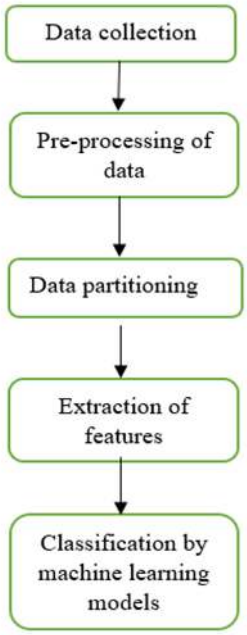
pre-processing, pre-processing with stemming, N-grams and N-grams with stemming. It is found that pre-processing with stemming has the highest accuracy and it is beneficial to use this. The limitation is that N-gram with stemming is not advised to use since it has negative effect in recall.

A multi-language system for cyberbullying detection classifier model is proposed by Batoul Haidar et al [15]. The dataset is taken by fetching the comments from facebook and twitter. The size of the dataset was found to be 4.93GB. The dataset is then cleaned and then pre-processed where in the cleaning all the language comments were removed except English and Arabic. Arabic tweet was contained in one dataset and English tweets in another dataset. WEKA toolkit is used because it the only toolkit that deploys Arabic language. Dataset is then divided into training and testing then it is given to the two machine learning models like support vector machines and naïve bayes as they perform well compared to other models. Both the datasets are given to the model and they are tested. Recall for both the models seems to be very low. It is just tested that cyberbully detection in Arabic language can be found. Recall of both the models is very low because of misclassification. Limitation is that the performance of the system should be improved.

3 Findings

a) If the training and testing dataset size is not mentioned, then the accuracy which we get from result is hard to believe.
b) Dataset can be increased to improve the accuracy of the classifier models.
c) There is high rate of false positivity when multilingual system is used.
d) The effectiveness of the system can be improved and increased when fuzzy rule set is used to retrieve relevant data.
e) Choosing cluster processes are inappropriate since they don't work in parallel

The main flow of the cyberbullying detection discussed in various papers is discussed below:



The dataset is collected at first and it is pre-processed in order to remove noise, error if it is present. Then it is divided into training and testing dataset. The features are extracted from the dataset so that the features are reduced. In some papers, selection of features is also done after feature extraction for increasing the accuracy and also to avoid overfitting. It is loaded into the machine learning models for classifying whether it is cyberbullied sentence or not. Here, the accuracy is also predicted to find the best machine leaning model among the set of models given for classification.

4 Conclusion

The related works of all the papers with their methodologies discussed above will be quite useful for cyberbullying detection. Cyberbullying is one of the severe problems in the society that even leads to death. Social media is the only way by which people connect and communicate with each other. This social media is becoming the environment for people to develop abusive comments for the hatred ones. The inappropriate way of using social media paves way for this cyberbullying. Therefore, it is important to detect cyberbullying in the social media. Till now, few implementations are done for cyberbullying detection and researchers need to concentrate more in cyberbullying since it destroys the life of people and also it makes people to live a coward life. From this survey, it is obvious that much scope exists in detection of cyberbullying and also many research contributions were made from all the machine learning models like supervised, unsupervised and reinforcement.

References

[1] Nideeksha B K, P Shreya, Sudharani Reddy P, MohamadiGhousiyaKousar. (2021). "Cyberbullying Detection Using Machine Learning", International Research Journal of Engineering and Technology (IRJET) .Volume: 08- Issue: 08.Aug 2021.

[2] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, Ammar Mohammed.(2019). "Social Media Cyberbullying Detection using Machine Learning". International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 10, No. 5, 2019.DOI:10.14569/IJACSA.2019.0100587

[3] Alwin T. Aind, AkashdeepRamnaney, DivyashikhaSethia.."2020 International Conference for Emerging Technology" (INCET) Belgaum, India.5-7 June 2020. DOI:10.1109/INCET49848.2020.9154092. IEEE.

[4] K. Nalini and L. Jaba Sheela. (2016). "Classification using Latent Dirichlet Allocation with Naive Bayes Classifier to detect Cyber Bullying in Twitter". Indian Journal of Science and Technology, Vol 9(28).DOI: 10.17485/ijst/2016/v9i28/93825.July 2016.

[5] WalisaRomsaiyud, KodchakornnaNakornphanom, PimpakaPrasertsilp, PiyapornNurarak, PiromKonglerd. (2017). "Automated Cyberbullying Detection using Clustering Appearance Patterns". 9th International Conference on Knowledge and Smart Technology (KST).pp. 242-247, DOI:10.1109/KST.2017.7886127.

[6] Vikas S Chavan, Shylaja S S. (2015). "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network", International Conference on Advances in Computing, Communications and Informatics (ICACCI).DOI: 10.1109/ICACCI.2015.7275970. IEEE.

[7] Vijay Banerjee, JuiTelavane, Pooja Gaikwad, Pallavi Vartak. (2019). "Detection of Cyberbullying Using Deep Neural Network". 5th International Conference on Advanced Computing & Communication Systems (ICACCS). DOI: 10.1109/ICACCS.2019.8728378.IEEE 2019.

[8] Monirah A. Al-Ajlan, Mourad Ykhlef. 2018. "Optimized Twitter Cyberbullying Detection based on Deep Learning". 21st Saudi Computer Society National Computer Conference (NCC).DOI: 10.1109/NCG.2018.8593146.IEEE 2018.

[9] Amgad Muneerand Suliman Mohamed Fati. 2020."A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter", future internet, 29 October 2020.https://doi.org/10.3390/fi12110187.

[10] Noviantho, Sani Muhamad Isa, Livia Ashianti, "Cyberbullying Classification using Text Mining", 1st International Conference on Informatics and Computational Sciences (ICICoS). DOI: 10.1109/ICICOS.2017.8276369.IEEE. 2017.

[11] Eloi Brassard-Gourdeau, Richard Khoury. (2019). "Subversive Toxicity Detection using Sentiment Information". Proceedings of the Third Workshop on Abusive Language Online, pages 1–10, August 1, 2019.

[12] Batoul Haidar, Maroun Chamoun, Ahmed Serhrouchni, "Arabic Cyberbullying Detection: Using Deep Learning", 7th International Conference on Computer and Communication Engineering (ICCCE). 2018. IEEE. DOI: 10.1109/ICCCE.2018.8539303.

[13] Jianwei Zhang, T. Otomo, L. Li and S. Nakajima. (2019). "Cyberbullying Detection on Twitter using Multiple Textual Features". 10th International Conference on Awareness Science and Technology (iCAST), pp. 1-6, 2019.doi: 10.1109/ICAwST.2019.8923186. IEEE.

[14] AzaldenAlakrot, Liam Murray, Nikola S Nikolov. (2018). Towards Accuarte detection of Offensive Languages in Online Communication in Arabic", 4th International Conference on Arabic Computational Linguistics, Science Direct, Procedia Computer Science142 315–320. Doi: 10.1016/j.procs.2018.10.491

[15] Batoul Haidar, Maroun Chamoun, Ahmed Serhrouchni.(2017). "A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning", Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 275-284. DOI: 10.25046/aj020634.