International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

A Review on Cyberbullying Detection using Machine Learning

Nideeksha B K[1], P Shreya[2], Sudharani Reddy P[3], Mohamadi Ghousiya Kousar[4]

[1,2,3]*B.E. Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, India*
[4]*Assistant Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology Bengaluru, India*

Abstract - *Social media is being extensively used today. This has led to a form of bullying that is Cyberbullying. Bullies use various network sites to attack victims with offensive comments and posts. This has been so devastating that many youngsters undergo depression, commit suicide, lose their self-confidence, and much more. With anonymity and lack of supervision this form of bullying has increased exponentially. It is also very challenging and difficult to monitor such cases. This leads us to find a way to help people out and protect them from such vulnerable attacks. Machine Learning has various algorithms that help us in detecting cyberbullying with some algorithms outperforming the others thereby leading us to the best algorithm.*

***Key Words*: cyberbullying, machine learning, convolutional neural network, deep learning, feature extraction, text classification.**

1. INTRODUCTION

Cyberbullying has been a major cause of worry for the amount of serious impact it has on people. Although social media is a secure place for communication it is prone to cyberbullying. It is found to be more dangerous than traditional bullying because the humiliation is visible to an unlimited online audience. Since the physical appearance of the victim is not required it can go on nonstop. Many networking sites don't even need a real name to be registered as a user making the bullies braver. The victims who have undergone bullying lose their self-confidence, become antisocial and this has a bad effect on their mental health as well. This leads us to detect cyberbullying. Hence machine learning techniques are employed in this paper. The steps involved are:

- Collecting datasets from networking sites which consist of videos, images, posts, and comments.
- Pre-processing data so that it contains only relevant information
- Classification of data into positive and negative instances of cyberbullying.

2. LITERATURE SURVEY

In 2020, Vimala Balakrishnan *et al.* [1] presented an automatic cyberbullying detection taking Twitter users' psychological features into account. The three main stages discussed in improving cyberbullying detection are Twitter data collection, feature extractions, and cyberbullying detection and classification. The annotated dataset contained 9484 tweets, out of which 4.5% of users are labelled as bullies, 31.8% as spammers, 3.4% as aggressors, and 60.3% as normal. However, the final dataset contained 5453 tweets as a result of the pre-processing step which included removing non-English tweets, profiles containing no data, and special characters. The features extracted were text features, user features, and network features. The model was executed using WEKA 3.8 with 10-fold cross-validation. Since Naïve Bayes performed poorly during preliminary experimental analysis it was eliminated while Random Forest and J48 continued to perform well. The classifiers were trained using manually annotated data.

In 2020, Jaideep Yadav *et al.* [2] proposed a novel pre-trained BERT model developed by Google researchers that generates contextual embeddings and task-specific embeddings. In the proposed method, for the base model, a deep neural network called the Transformer is used. The Bert contains 12 layers to encode the input data and is built on top of a base model. The data is tokenized and padded accordingly and is fed into the model which generates the final embeddings. The classifier layer classifies the embeddings generated by the previous layers and generates the final output accordingly. Using a pre-trained BERT model they were able to achieve efficient and stable results in comparison to the previous models to detect cyberbullying.

In 2020, Sudhanshu Baliram Chavan *et al.* [3] proposed the approach to detect cyberbullying on Twitter. The required dataset was collected from sources like GitHub, Kaggle. Initially, the data is pre-processed and features are extracted using a TFDIF vectorizer algorithm. These tweets are then passed through the naive Bayes and SVM model and are classified accordingly. When a tweet is categorized as bullying, ten other tweets from that users' account will be fetched and passed through naive Bayes and SVM classifiers again. If the overall probability of that user's tweets lies above 0.5 then it will be considered as a bullied tweet. Based on the accuracy score and the results it was evident that the SVM model outperformed the naive Bayes with the accuracy score of 71.25%.

In 2019, John Hani *et al.* [4] presented a supervised learning approach to detect cyberbullying. As a part of the preprocessing step, data is cleaned by removing the noise and unnecessary text. This is performed using tokenization, lowering text, stop words along with encoding cleaning and word correction. The second step is the feature extraction step which is done using TF IDF and sentiment analysis technique including NGrams for considering different combinations of the words like 2- Gram, 3-Gram, and 4-Gram. The cyberbullying dataset from Kaggle is split into ratios (0.8, 0.2) for train and test. SVM and Neural networks are used as classifiers that run on a different n-gram language model. Accuracy, recall and precision, and f-score are the performance measures. It is found that Neural Network performed better than the SVM classifier. Neural Network achieved an average f-score of 91.9% and SVM achieved an average f-score of 89.8%.

In 2018, Monirah Abdullah Al-Ajlanet and Mourad Ykhlef [6] proposed a novel algorithm CNN-CB which is based on a convolutional neural organization and adapts the idea of word embedding. The architecture comprises four layers - Embedding, Convolution Layer, Max Pooling Layer, and Dense Layer. The first layer, word embedding, creates a vector space of vocabulary which is the input to the subsequent layer, the convolutional layer, which compresses the input vector without losing significant features. The third layer, the Max pooling layer, takes the output of the second layer as its input and finds the maximum value of the chosen region to save just significant highlights. The last layer, the Dense layer, does the classification. This gave a precision of 95%.

In 2018, Monirah A. Al-Ajlan *et al.* [7] proposed optimized Twitter cyberbullying detection based on deep learning (OCDD) which does not extract features from tweets instead, it represents a tweet as a set of word vectors that are fed to a convolutional neural network (CNN) for classification. Hence the feature extraction and selection phases are eliminated in this approach. To represent the semantics between words, word embedding is used and is generated using (GloVe) technique. CNN uses a lot of parameters and to optimize these values, a metaheuristic optimization algorithm is used to find optimal or near-optimal values that will be used for classification. CNN showed great results.

In 2017, Yee Jang Foong and Mourad Oussalah [10] presented an automated cyberbullying detection that uses natural language processing techniques, text mining, and machine learning. For dataset ASKfm, a social media platform where users can anonymously ask questions and view a sample of a user's profile is used. As a part of the pre-processing procedure web links and unknown characters are removed, incorrect wordings in case any are corrected, and also lexicons are replaced with equivalent textual expressions. A combination of features has been used which includes TF-IDF, Unusual capitalization count, LIWC, and Dependency parser. The data set is split into a 70% training set and 30% testing set. SVM was used as a classifier which was trained with a linear kernel on the training data. To label the training posts Amazon Mechanical Turk Service was used. The combination of features mentioned above yielded the highest performance in terms of accuracy, precision, recall, F1, and F2 scores.

In 2016, X. Zhang *et al.* [11] proposed a novel approach based on a pronunciation-based convolutional neural network (PCNN). Word-to-Pronunciation conversion is done to group a set of words spelled incorrectly, which have the same meaning and pronunciation, together with the corrected word. Two separate CNN is used to establish a baseline. For the first baseline feature set, word-embedding based on Google's word-vector was used. For the creation of the feature set of the second baseline, CNN Random, arbitrarily generated vectors were used. The phoneme codes were arbitrarily introduced into vectors for the feature set for PCNN. To handle class imbalance three techniques were implemented- threshold moving, cost function adjust, and a hybrid solution, out of which cost function adjusting is most effective.

In 2016, Michele Di Capua *et al.* [12] presented an unsupervised approach to detect cyberbullying using a design model inspired by Growing Hierarchical SOMs. Firstly, features are divided into four groups: Syntactic features, Semantic features, Sentiment features, Social features. Growing Hierarchical Self Organizing Map (GHSOM) network algorithm, which is well suited for a large collection of documents that has to be classified, is used. It uses a hierarchical structure of multiple layers, where each layer consists of a variety of independent SOMs. A single SOM is employed at the root layer. For every unit, during this map, a SOM could be added to the subsequent layer of the hierarchy. GHSOM Network is trained and tested concerning a K-folded dataset, applying a K-fold partitioning of data.

In 2014, Sourabh Parime and Vaibhav Suri [13] presented an approach of using data mining and machine learning techniques to detect cyberbullying. Text mining is performed on unstructured data using machine learning techniques to extract knowledge from the text which includes multiple stages like document clustering, data pre-processing, attribute generation for which an in-built classifier is used to generate labels from the features fed into it and occurrences are counted and a weight is assigned to each label and irrelevant attributes are removed which helps to estimate the nature of the comments. Sentiment analysis is used for determining the tone of the given text. Two classes of data are considered one with positive emotions and the other with negative emotions. These are stored into a vector and used to train a supervised learning algorithm SVM.

In 2011, Kelly Reynolds *et al.* [14] presented a language-based method for detecting cyberbullying. Data for the dataset is collected from the website Formspring.me which is a question and answer-based website where users openly invite others to ask and answer questions. This website is highly populated by teens and college students increasing

the percentage of bullying content. Amazon's Mechanical Turk labelled a post as "yes" if it was a cyberbullying post else "no". Out of 2696 posts in the training set, 196 received a final class label of "yes," and out of 1219 posts in the test set, 173 were recognized as cyberbullying. The SUM and TOTAL features that were used to measure the overall badness of a post were included in both the versions of the datasets which were NUM and NORM. Weka a software suite for machine learning which uses J48, JRIP, IBK, and SMO algorithms. Using 10- fold cross-validation it is observed that the NORM training set outperforms the NUM training set except for the SMO algorithm.

In 2011, Roi Reichart *et al.* [15] proposed the approach to detect cyberbullying on social media using a range of binary and multiclass classifiers. They have used a dataset from the YouTube comment section and grouped it into labels of sexuality, physical appearance, race, and intelligence they have trained various supervised models like JRip, SVM, J48, and Naive Bayes. They have experimented with a binary classifier trained for specific labels and multiclass classifiers on all combined labels. On examining the kappa statistic, accuracy it was evident that the label-specific classifiers outperformed the multiclass classifiers in detecting cyberbullying.

Table -1: Tabular Summary of Literature Survey

AUTHORS	TITLE AND YEAR OF PUBLICATION	ALGORITHM (CLASSIFIERS)
Balakrishnan,Vimala *et al.* [1]	Improving cyberbullying detection using Twitter user's psychological features and machine learning (2020)	Random Forest, J48
J. Yadav, D. Kumar *et al.* [2]	Cyberbullying Detection using Pre-Trained BERT Model (2020)	Pre-trained BERT
Sudhanshu Baliram Chavan *et al.* [3]	Detecting A Twitter Cyberbullying Using Machine Learning (2020)	SVM, Naive Bayes
John Hani, Mohamed Nashaat *et al.* [4]	Social Media Cyberbullying detection using machine learning (2019)	SVM, Neural Network
R. Pawar and R. R. Raje *et al.* [5]	Multilingual Cyberbullying Detection System (2019)	MNB, LR, SGD
Monirah Abdullah Al-Ajlanet and Mourad Ykhlef [6]	Deep Learning algorithm for cyberbullying detection (2018)	CNN-CB
M. A. Al-Ajlan and M. Ykhlef [7]	Optimized Twitter Cyberbullying Detection based on Deep Learning (2018)	CNN
B. Haidar, M. Chamounet *et al.* [8]	Arabic Cyberbullying Detection: Using Deep Learning (2018)	Feed Forward Neural Network
Noviantho, Sani Muhamad Isa *et al.* [9]	Cyberbullying Classification using Text Mining (2017)	SVM with poly kernel, Naive Bayes
Yee Jang Foong and Mourad Oussalah [10]	Cyberbullying System Detection and Analysis (2017)	SVM
X. Zhang *et al.* [11]	Cyberbullying detection with a pronunciation based convolutional neural network (2016)	CNN, PCNN
Michele Di Capua, E. Di Nardo *et al.* [12]	Unsupervised Cyberbullying detection in social networks (2016)	GHSOM Network Algorithm
Sourabh Parime and Vaibhav Suri [13]	Cyberbullying detection and prevention: Data Mining and psychological perspective (2014)	SVM
Kelly Reynolds, A. Kontostathis *et al.* [14]	Using Machine Learning to Detect Cyberbullying (2011)	J48, JRIP, IBK and SMO
Roi Reichart, Dinakar, K *et al.* [15]	Modelling the Detection of Textual Cyberbullying (2011)	JRip, SVM, J48 and naive bayes

4. METHODOLOGY

The general procedure of Cyberbullying detection consists of data collection, data pre-processing, feature extraction, feature selection, and lastly classification which is given below.

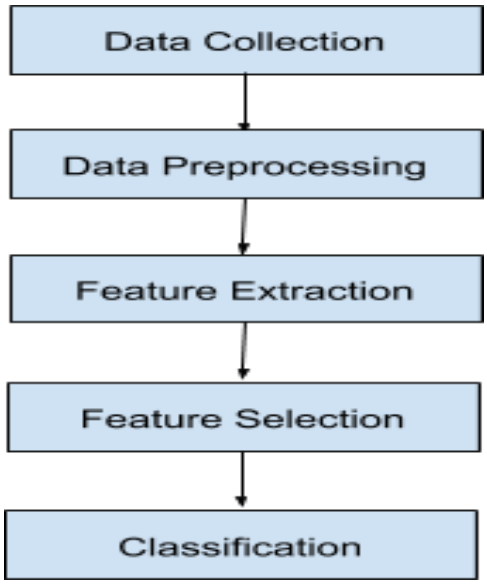

Fig -1: Methodology

Data collection is a process of gathering data for training and testing the model, this can be existing datasets or messages and remarks from a particular web-based media. The data collected might contain noise and must be pre-processed for appropriately preparing the model. Data pre-processing includes the removal of hyperlinks, hashtags, usernames, as, what, who, etc., and special characters that are not required for training in the model. Feature extraction is crucial for recognizing cyberbullying text. Investigation of the features might be done using Bag of Words, TF-IDF Vectorizer, n-grams, Word2Vec, Doc2Vec, emotion values of messages, etc. In the feature selection stage, important features must be selected to increase the accuracy and to reduce overfitting. This can be done using Filter, Wrapper, Embedded, and Hybrid methods. Lastly, the classification must be done using JRip, SVM, J48, naive Bayes, neural networks, etc.

5. FUTURE SCOPE

As for future work, we would like to implement the proposed approach to detect cyberbullying in different languages as social media is vast and is not restricted to a single language. We can look for patterns in behaviour on the social media platform instead of a single post. By identifying patterns, we can alert them based on the user's behaviour. There should be more research that has to be conducted where schools, colleges, and various communities, are addressing cyberbullying to determine the best means to avoid it.

6. CONCLUSION

The literature survey done in this paper provides insight into the detection of cyberbullying. Different methods to detect cyberbullying is mentioned. The main steps to detect cyberbullying are mentioned which include data collection, data pre-processing, feature extraction, feature selection, and lastly classification. Multiple techniques can be used for the detection of cyberbullying such as SVM, Naive Bayer's, neural networks, etc. It is also seen in a few papers that models using neural networks give slightly better performance.

7. REFERENCES

[1] Balakrishnan, Vimala & Khan, Shahzaib & Arabnia, Hamid. (2020). Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning. Computers & Security. 90. 101710. 10.1016/j.cose.2019.101710.

[2] J. Yadav, D. Kumar and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700.

[3] R. R. Dalvi, S. Baliram Chavan and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893

[4] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer and Ammar Mohammed, "Social Media Cyberbullying Detection using Machine Learning" International Journal of Advanced Computer Science and Applications(IJACSA),10(5),2019.

[5] 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.

[6] Monirah Abdullah Al-Ajlan and Mourad Ykhlef, "Deep Learning Algorithm for Cyberbullying Detection" International Journal of Advanced Computer Science and Applications(IJACSA), 9(9), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090927

[7] M. A. Al-Ajlan and M. Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning," *2018 21st Saudi Computer Society National Computer Conference (NCC)*, Riyadh, 2018, pp. 1-5, doi: 10.1109/NCG.2018.8593146.

[8] B. Haidar, m. Chamoun and a. Serhrouchni, "arabic cyberbullying detection: using deep learning," 2018 7th international conference on computer and

communication engineering (iccce), kuala lumpur, 2018, pp. 284-289, doi: 10.1109/iccce.2018.8539303.

[9] Noviantho, s. M. Isa and l. Ashianti, "cyberbullying classification using text mining," 2017 1st international conference on informatics and computational sciences (icicos), semarang, 2017, pp. 241-246, doi: 10.1109/icicos.2017.8276369.

[10] Y. J. Foong and M. Oussalah, "Cyberbullying System Detection and Analysis," 2017 European Intelligence and Security Informatics Conference (EISIC), Athens, 2017, pp. 40-46, doi: 10.1109/EISIC.2017.43.

[11] X. Zhang et al., "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 740-745, doi: 10.1109/ICMLA.2016.0132.

[12] M. Di Capua, E. Di Nardo and A. Petrosino, "Unsupervised cyber bullying detection in social networks," *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, 2016, pp. 432-437, doi: 10.1109/ICPR.2016.7899672.

[13] S. Parime and V. Suri, "Cyberbullying detection and prevention: Data mining and psychological perspective," 2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014], Nagercoil, 2014, pp. 1541-1547, doi: 10.1109/ICCPCT.2014.7054943.

[14] 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, 2011, pp. 241-244, doi: 10.1109/ICMLA.2011.152.

[15] Dinakar, K., Roi Reichart and H. Lieberman. “Modeling the Detection of Textual Cyberbullying.” *The Social Mobile Web* (2011).