IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review On Detection Of Cyber-Bully Messages Using Machine Learning Algorithms

Balram Singh Yadav[1], Dr. Saurabh Sharma[2]

[1]Research scholar, Sant Baba Bhg singh University
[2]Assistant Professor, Sant Baba Bhg singh University

Abstract : The evolution and growth of social networking and modern web technology have made an individual's online presence permanent. People frequently express their thoughts, ideas, and emotions through social networking links, with the most popular activity being the discussion of everyday events, which may include private or public conversations. Bullying that involves technology is referred to as cyber-bullying. Bullying attacks target teenagers and young people who spent lot of time on social networking sites. The rise of social media, especially Twitter, has caused confusion about the meaning of free expression, which has given rise to a number of worries. Cyber-bullying is one of these issues, a severe global problem that has an impact on both people and societies. There have been numerous reported attempts to intervene, prevent, or lessen cyber-bullying; however, these efforts are unworkable since they depend on the victims' interactions. Victims of this behavior may experience hopelessness and other potentially fatal issues. It is necessary to create monitoring and detection procedures for potentially hazardous Internet behavior. By using machine learning, we can create algorithms to automatically identify cyber-bullying content and recognize language patterns used by bullies and their victims. Therefore, it is crucial to identify cyber-bullying without the victims' participation. Additionally, numerous machine learning classifiers were applied, including the K-nearest neighbor technique, linear regression, decision trees, and the Support Vector Machine classifiers, Random Forest, Naive Bayes, and AdaBoost. The most precise supervised learning algorithm was used to identify communications that contained cyber-bullying.

Keyword: cyber-bully, cybercrime, social media, traditional bullying, social Networking sites, Harassment.

I. Introduction

A new kind of harassment has emerged as a result of the increasing use of social media. A purposeful or aggressive act committed by a single person or group of people utilising repeated communication messages again and again to victim who is powerless to defend them is referred to as bullying [1]. Most people now depend on technology to survive, and it has become an essential element of our lives. The sharing of ideas is made easier by the internet. A large number of people spend a daily lot of time on social media. Communication between individuals is no exception, since technology has changed how people connect and introduced a new layer of communication. These communities are being abused by many people. Nowadays, bullying affects a lot of children. Bullies harass people online using tools like Facebook,Twitter and email. Studies show that in India has 37% of children engage in cyber-bullying, while 14% of children in India experience bullying on a regular basis. The victim of cyber-bullying is affected emotionally and psychologically. Bullies are also given the anonymity they need on social media to commit their heinous crimes. Bullying gets worse over time as it happens more frequently. Therefore, the sufferer will gain if it is prevented.

II. Cyber bullying and its impact on society:

Cyber-bullying is the act of harassing, threatening, or bullying someone through contemporary means of communication with one another and with anyone and everyone on the globe via social media apps and sites. Threatening someone is a form of cyber-bullying that goes beyond making up a false identity, publishing or posting an embarrassing image or video, or spreading negative rumors about someone. Social media cyber-bullying has horrible consequences, with some unfortunate victims even passing away as a result. These people exhibit emotions, self-assurance, and fear. Therefore, this problem needs a complete solution. Online bullying needs to stop. Machine learning can be used to find and stop the problem, but you need to look at it from a different angle.

There are several categories of cyber-bullying as stated by [8] and [9] :

- **Flaming:** starting a form of online fight.
- **Masquerade:** where there is a bully pretending to be someone else, in order to perform malicious intents.
- **Denigration:** sending or posting gossip to ruin someone’s reputation.
- **Impersonation:** Pretending to be someone else and sending or posting material to get that person in trouble or danger or to damage that person’s reputation or friendships.
- **Harassment**: Repeatedly sending profane and cruel messages.
- **Outing:** Publishing someone’s embarrassing information, images or secrets.
- **Trickery:** Talking someone into revealing secrets or embarrassing information for the sake of sharing it online.
- **Exclusion:** Intentionally and cruelly excluding someone from an online group.
- **Cyber-stalking:** Repeated, intense harassment and denigration that includes threats or creates significant fear.

III. Machine learning algorithms

The most important stage in the pipeline for text categorization is choosing the best classifier. We are unable to choose the best model for a text classification implementation without having a solid conceptual grasp of each approach. In particular, Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors and Support Vector Machine (SVM) are a few machine learning methods (KNN).

Naïve bayes: Known as one of the most effective and efficient inductive learning algorithms, Naive Bayes has been successfully applied as a classifier in various social media research. The Naive Bayes classification for text, which can categorise any sort of data, including text, network characteristics, phrases, etc., has been used frequently in document categorization assignments since the 1950s. By taking samples from this model, it is possible to come up with new data that is similar to the training data.

K-Nearest Neighbors : KNN is the simplest instance-based learning algorithm suitable for multi-class situations for supervised learning technique. This algorithm separates a fresh sample from its neighbours based on distance. The system allocates an item to the class that has the highest frequency among its K-nearest neighbours after locating its K-nearest neighbours within the training set. KNN is a non-parametric lazy learning algorithm that doesn't make any assumptions about how the underlying data is distributed.

Decision trees : Supervised machine learning techniques like decision trees constantly divide the data based on a specific criterion. Decision nodes and leaves are the two components that make up the tree. The leaves stand for choices and results. And the decision nodes divide the data.

Random forest : An ensemble algorithm called Random Forest (RF) is used to tackle regression and classification issues. RF constructs many decision tree classifiers using a arbitrary subset of data samples and characteristics. Through decision trees with a majority vote, fresh samples are classified. The main benefits of RF are that it operates effectively on huge datasets, provides good accuracy even when a large amount of the data is missing, and is a useful tool for guessing missing data..

Support Vector Machine (SVM): "Support Vector Machine" is the supervised learning algorithm called SVM can categorize both linear and nonlinear data. Finding the optimum separators for differentiating classes in the search space is the central idea of SVM. When employing necessary training tuples, support vectors are the data points that separate one or more hyperplanes. The nonlinear SVM classifier is sometimes used when there is no way to separate two data points with a straight line.

IV. Related Work with existing Algorithm’s

Many ideas propose technological solutions that can accurately and automatically detect cyber-bullying as described in Table 1. Nandhini et. al. [2] built a model utilizing data from MySpace.com that achieved 91% accuracy using the Naive Bayes machine learning approach. They then proposed a model [3] that employs genetic processes and the Naive Bayes classifier to attain 87% accuracy. Romsaiyud et. al. [4] investigate word extraction using the Naive Bayes classifier and clustering of loaded patterns. The algorithm divided the datasets into eight classes using two major methods: (1) utilising k-mean clustering to iteratively group the full datasets into clusters; and (2) capturing any given partition with the frequency of words using a multinomial model feature vector. The suggested strategy improved the experiment's precision and dependability. Furthermore, Bunchanan et. al. [5] .'s approach made use of a dataset that was manually categorised from War of Tanks game discussion. Compared to simple Naive classification that includes sentiment analysis as a feature and their results were poorer than those of the human categorised dataset. Additionally, Isa et. al.. [6] suggested a method in which they obtained their dataset from Kaggle and employed two classifiers, Naive Bayes and SVM. The Naive Bayes classifier had an average accuracy of 92.81 percent compared to the 97.11 percent of SVM with poly kernel's. However, because the dataset's training and testing sizes were not provided, the results might not be accurate. The dataset used by Dinakar et. al.. to identify explicit bullying language relating to sexuality, IQ and ethnicity and culture came from the YouTube comment area.

When SVM and Naive Bayes classifiers were used, SVM had an accuracy of 66% and Naive Bayes had an accuracy of 63%. Di Capua et. al.. [8] proposed a brand-new method for detecting cyber-bullying using an unsupervised approach Growing Hierarchical SOMs. On FormSpring get 73% accuracy , YouTube 69% accuracy and 72% accuracy on twitter. A model to detect cyber-bullying developed by Haidar et. al. [9] Naive Bayes, and SVM on Arabic language and achieved 90.85% precision and SVM achieved 94.1% as precision.

Table 1 Researchers implemented cyber bully detection using Machine learning algorithm

Auther Name	Method	Result
Nandhini et. al. [2]	Naive Bayes machine learning	91% accuracy
B Sri Nandhini and JI Sheeba.[3]	Naive Bayes classifier and genetic operations (FuzGen)	87% accuracy
Romsaiyud et. al. [4]	Naive Bayes classifier for extracting the words and examining loaded pattern clustering.	increasing accuracy
Bunchanan et. al. [5]	simple Naive classification & Manually	Poor result
Isa et. al. [6]	Classifier Naive Bayes and SVM	Naive Bayes classifier yielded average accuracy of 92.81% SVM with poly kernel yielded accuracy of 97.11%,
Dinakar et. al. [7]	SVM and Naive Bayes classifiers	SVM 66% and Naive Bayes 63% accuracy
Di Capua et. al.[8]	unsupervised approach applying SVM	SVM on FormSpring and achieving 67% ,GHSOM on YouTube 60% precision, 69% accuracy and 94% recall, Naive Bayes on Twitter 67% accuracy
Haidar et. al. [9]	Naive Bayes and SVM	90.85% precision and SVM 94.1% as precision
Zhao et. al. [10]	SVM	recall of 79.4%.
Parime et. al. [11]	SVM Classifier	
Chen et. al. [12]	feature extraction using Lexical Syntactic Feature and SVM	achieved 77.9% precision and 77.8% recall
Ting et al. [13]	SNM	97% precision and 71% as recall
Harsh Dani et. al. [14]	KNN	0.6105 F1 score and 0.7539 AUC score.
Hee et. al. [15]	SVM on English and Dutch	F1 score of 64% and 61% for English and Dutch
Theyazn H. H. Aldhyani et. al.[17]	Convolutional neural networks integrated with bidirectional long short-term memory networks (CNN-BiLSTM) and single BiLSTM	Detection accuracy of 94%, BiLSTM outperformed the combined CNN-BiLSTM classifier, achieving an accuracy of 99%.
Pradeep Kumar & Fenish Umeshbhai [18]	Deep learning-based convolutional neural network	Accuracy of 89% for the best case,
Dewani A, and et. al. [19]	RNN-LSTM, RNN-BiLSTM and CNN models	RNN-LSTM and RNN-BiLSTM accuracy of 85.5 and 85% and F1 score was 0.7 and 0.67 respectively
Bharti, S. and et al,[20]	Bi-directional long short-term memory (BLSTM) used for classification.	Accuracy, precision and F1 measure of 92.60%, 96.60% and 94.20%, respectively.

Zhao et. al. [10] provided a framework for identifying cyber-bullying; they utilised word embedding to create a list of pre-defined undesirable phrases and gave weight's to acquire bullying traits; they get result 79.4% using SVM as their primary classifier. Parime et. al. [11], manually annotated on the dataset from MySpace, it, and used the SVM Classifier to classify it, proposed a different strategy. Chen et. al. [12], who utilized SVM as their classifier and got 77.8% recall and 77.9% accuracy, also suggested a novel feature extraction method called Lexical Syntactic Feature. Ting et. al. [13] established a method based on Social network mining technique; they included elements such as sentiments and Social network analysis measures, and their data came from social media. Seven experiments were conducted, and the precision and recall were nearly 97% and 71%, respectively. In addition, a fresh framework named SICD was introduced by Harsh Dani et. al. [14] that uses KNN for categorization. Their final grades were 0.7539 AUC and 0.6105 F1. Hee et. al. [15] report the collection and fine-grained annotation of an English and Dutch cyber-bullying corpus, which was done in order to demonstrate the viability of automatic cyber bullying detection. Additionally, they carried out a number of binary classification trials. They examine the application of highly feature-rich linear support vector machines to locate posts that are connected to cyber-bullying. 64% for English and 61% for Dutch are their respective scores. Theyazn H. H. Aldhyani et. al.[17] used Convolutional neural networks integrated with bidirectional long short-term memory networks (CNN-BiLSTM) and single BiLSTM for detection cyber-bully messages and attain 99% accuracy. Pradeep Kumar & Fenish Umeshbhai [18] use the deep learning-based convolutional neural network model and got 89% accuracy for detecting the cyber-bully post. Dewani A et. al.[19] use RNN-LSTM, RNN-BiLSTM models for detecting the cyber-bully method in Urdu-Roman language and got accuracy 85.5 and 85% whereas F1 score was 0.7 and 0.67 respectively. Bharti, S. et. al.[20] use deep learning algorithms for word embedding technique on 35,787 with BLSTM and attain results on the dataset with an accuracy, precision and F1 measure of 92.60%, 96.60% and 94.20%, respectively .

V. Conclusion

People are able to communicate with one another thanks to the widespread use of the Internet. Which elements make it more likely that new technology will be misused? Due to the quick advancement of technology, kids are more active than ever online. Therefore, surfing the Internet safely requires that safeguards be taken. It is critical to implement legislative safeguards and educate kids in order to safeguard youngsters from cyber-bullying and its repercussions. Communication between young children and teenagers and their parents is crucial. Parents should talk to their kids about the internet, social media, and any unfavourable consequences they may have. Everyone has to know how to use social media and the Internet responsibly. In addition to schools and universities, education is also offered at workplaces. Adults are also susceptible to cyber-bullying. It is critical to recognise that virtual spaces on the Internet can be just as dangerous as physical spaces.Cyber-bullying may be even more harmful than other forms of abuse. Domina Petric [2019] [16]. This research article explores the use of machine learning to automatically identify posts on social media sites that are related to cyber-bullying. It is now impossible to manually monitor cyber-bullying. The automatic detection of cyber-bullying signals enhances moderation and enables us to take prompt action when required. Research on cyber-bullying has usually focused on identifying cyber-bullying "attacks," which has left out other or more subtle kinds of cyber-bullying as well as posts made by victims and onlookers. However, these posts might also be a sign of online bullying. The main goal of this research is to offer algorithms that automatically identify signals of cyber-bullying on social media, including various forms of cyber-bullying and posts from bullies, victims, and onlookers. As a result, steps must be done to define the various cyber-bullying problems that kids and teens could experience and to present potential fixes. Machine learning makes it straightforward to carry out this analysis.

References

[1] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: Its nature and impact in secondary school pupils. Journal of child psychology and psychiatry, 49(4):376–385, 2008.

[2] B Nandhini and JI Sheeba. Cyberbullying detection and classification using information retrieval algorithm. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), page 20. ACM, 2015.

[3] B Sri Nandhini and JI Sheeba. Online social network bullying detection using intelligence techniques. Procedia Computer Science, 45:485–492, 2015.

[4] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. Automated cyberbullying detection using clustering appearance patterns. In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242– 247. IEEE, 2017.

[5] Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. Computers & Security, 76:197–213, 2018.

[6] Sani Muhamad Isa, Livia Ashianti, et. al. Cyberbullying classification using text mining. In Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on, pages 241–246. IEEE, 2017.

[7] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3):18, 2012.

[8] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. Unsupervised cyber bullying detection in social networks. In Pattern Recognition (ICPR), 2016 23rd International Conference on, pages 432–437. IEEE, 2016.

[9] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Advances in Science, Technology and Engineering Systems Journal, 2(6):275–284, 2017.

[10] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th international conference on distributed computing and networking, page 43. ACM, 2016.

[11] Sourabh Parime and Vaibhav Suri. Cyberbullying detection and prevention: Data mining and psychological perspective. In Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on, pages 1541–1547. IEEE, 2014.

[12] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pages 71–80. IEEE, 2012.

[13] I-Hsien Ting, Wun Sheng Liou, Dario Liberona, Shyue-Liang Wang, and Giovanny Mauricio Tarazona Bermudez. Towards the detection of cyberbullying based on social network mining techniques. In Behavioral, Economic, Socio-cultural Computing (BESC), 2017 International Conference on, pages 1–2. IEEE, 2017.

[14] Harsh Dani, Jundong Li, and Huan Liu. Sentiment informed cyberbullying detection in social media. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 52– 67. Springer, 2017.

[15] Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, et. al. (2018) Automatic detection of cyberbullying in social media text. PLOS ONE 13(10): e0203794. https://doi.org/10.1371/journal.pone.0203794

[16] Domina Petric, 2019, Cyberbullying, doi: 10.13140/RG.2.2.35163.82729

[17] Aldhyani, T.H.H.; Al-Adhaileh, M.H.; Alsubari, S.N. Cyberbullying Identification System Based Deep Learning Algorithms. Electronics 2022, 11, 3273. https:// doi.org/10.3390/electronics11203273

[18] Roy, P.K., Mali, F.U. Cyberbullying detection using deep transfer learning. Complex Intell. Syst. 8, 5449–5467 (2022). https://doi.org/10.1007/s40747-022-00772-z

[19] Dewani A, Memon MA, Bhatti S. Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. J Big Data. 2021;8(1):160. doi: 10.1186/s40537-021-00550-7. Epub 2021 Dec 22. PMID: 34956818; PMCID: PMC8693595.

[20] Bharti, S., Yadav, A.K., Kumar, M. and Yadav, D. (2022), "Cyberbullying detection from tweets using deep learning", Kybernetes, Vol. 51 No. 9, pp. 2695-2711. https://doi.org/10.1108/K-01-2021-0061