Cyberbullying Detection Using Machine Learning

Polasa Jahnavi[1], Siliveri Rohith Vardhan[2],Shashank Kandhaktla[3]

[123] Student, Computer Science & Engineering, Anurag University, Telangana, India

---***---

Abstract - *In the current digital era, The issue of cyberbullying is spreading and has the potential to seriously harm people's mental health, interpersonal relationships, and academic achievement. To recognise and stop such activity, there is a growing demand for automated cyberbullying detection systems. This paper proposes a machine learning-based approach for detecting cyberbullying in all forms. cyberbullying is a developing problem that can take many different forms, including text, photos, and PDF documents. Cyberbullying detection is being developed to evaluate various forms of content and spot instances of the behaviour using machine learning techniques like SVM[15], k closest neighbour, Decision Tree[17], and Random Forest[18]. To identify cyberbullying in photographs and text, we used image recognition algorithms, OCR[19], and natural language processing approaches. We trained the aforementioned machine learning algorithms on a sizeable dataset of labeled cyberbullying content. The prevalence of cyberbullying might be considerably reduced by the suggested method, and the online space could become safer and more welcoming as a result.*

Key Words: **Machine learning, Natural language processing, Cyberbullying, Text, Image, Documents.**

1. INTRODUCTION

Cyberbullying is the intentional use of electronic communication tools like social media, texting, email, or instant messaging to harass, threaten, or hurt someone. Cyberbullying may take many different forms, such as sending threatening or unpleasant messages, spreading rumours or gossip online, disclosing private or embarrassing information, making up false accounts or posing as someone else, or even cyberstalking. Cyberbullying may have severe repercussions for the victim, such as mental discomfort, sadness, anxiety, and in severe cases, suicide. Cyberbullying is a major problem that has to be acknowledged, addressed and prevented.

The practice of finding and highlighting instances of cyberbullying using technology is referred to as cyberbullying detection. Cyberbullying detection can be done manually, with human moderators reviewing online material and identifying instances of cyberbullying, or using automated techniques, such as machine learning algorithms. Automatic cyberbullying detection technologies examine online information such as social media postings, comments, and messages to discover patterns of behaviour that are suggestive of cyberbullying. These systems may be trained on big datasets of labeled data containing instances of cyberbullying and non-cyberbullying material.

Once trained, the cyberbullying detection tool may automatically flag instances of cyberbullying and take action to block or delete the offending material, either by notifying a human moderator for additional review or by blocking or removing the offending content. Automatic cyberbullying detection technologies can be effective in recognising and reducing cyberbullying on a large scale, but they must be accurate, unbiased, and respectful of individual privacy.

2. RELATED WORKS

In our project, we have implemented this model to detect bullying by browsing the web for published articles.

This section examines the most current automated Cyberbullying Detection Classification methods.

Table -1: literature survey on Cyberbullying Detection.

Research Papers on Cyberbullying Detection			
S.no	Title	Dataset	Methodology
1	Cyberbullying Detection on Social Networks Using Machine Learning Approaches[14] (Adya Bansal, Akash Baliyan, Akash Yadav)	Datasets from Twitter comments and remarks.	NLP(Natural Language Processing. ML(Machine Learning)

2	Cyberbullying Detection in Social Networks Using Deep Learning Based Models[13] (Maral Dadvar and Kai Eckert)	Datasets from Wikipedia Twitter comments and remarks and Formspring.	Deep Neural Network-Based Models
3	A multilingual system for detecting cyberbullying in Arabic content using machine learning.[4] (Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni.)	Datasets are Arabic language text data.	ML(Machine Learning)

3. THE PROPOSED MODEL

This paper proposes a machine learning-based approach to detect cyberbullying in various forms, including text, photos, and PDF documents. Machine learning techniques such as SVM, k nearest neighbour, Decision Tree, and Random Forest are used, as well as image recognition algorithms, OCR[19], and natural language processing approaches. The suggested strategy has the potential to greatly reduce cyberbullying.

Detects cyberbullying in the following:

- Text
- Image
- Text file(.txt)
- PDF(.pdf)

The purpose of this work is to identify cyberbullying and to make the internet a safer and more inclusive place by avoiding cyberbullying and safeguarding individuals from its negative consequences. This is accomplished by reporting instances of cyberbullying for assessment and subsequent action by human moderators, who may then take necessary actions to address the problem and help victims.

For text analysis, the system will use topic modelling techniques to identify the emotional content and themes of the text. The system will also analyse the text for the use of derogatory language and other harmful expressions commonly used in cyberbullying. This method for detecting cyberbullying using machine learning is to train a classification model on a labeled dataset of cyberbullying incidents. The algorithm may then be used to predict whether or not fresh text descriptions include cyberbullying. To train the model features such as profanity, insulting language, and hostile tone may be retrieved from the text.

NLP approaches may also be used to preprocess text data and extract aspects such as sentiment, grammar, and semantics. These characteristics can be utilised to increase the classification model's accuracy.

For image analysis, the system will use deep learning techniques to recognise patterns and identify visual cues associated with cyberbullying. The system will look for specific types of images, such as those with derogatory captions or those depicting harmful acts, to identify instances of cyberbullying.

For PDF documents, the system will use OCR to convert the text into a machine-readable format, and then apply the same text analysis techniques used for other types of content.

Once the system identifies an instance of cyberbullying, it will flag the content for review and further action by human moderators, such as reporting the content to appropriate authorities or removing it from the platform.

3.1 ALGORITHMS

- Support Vector Machine (SVM[15]) is a powerful supervised learning algorithm used for classification and regression analysis. The goal of SVM[15] is to find the best hyperplane that separates the data points into different classes with the largest possible margin.

- K-Nearest Neighbours (KNN[16]) is a simple and widely-used non-parametric classification algorithm in machine learning. KNN[16] works by finding the K closest data points in the training set to a given input data point and assigns a label to the input data point based on the most common label among its K nearest neighbors.

- Decision Tree[17] is a popular machine learning algorithm used for classification and regression analysis. It works by recursively splitting the data into subsets based on the most informative features until a decision is made about the class label or predicted value of a given input data point.

- Random Forest[18] is an ensemble learning algorithm used for both classification and regression analysis. Random Forest works by constructing a multitude of Decision Trees at training time and outputting the class that is the mode of the classes of the individual trees.

3.2 OPTICAL CHARACTER RECOGNITION(OCR)

OCR stands for Optical Character Recognition, which is a technology used to convert printed or handwritten text into a machine-readable format. OCR involves scanning the text using an optical scanner or a smartphone camera and then using image processing techniques to extract the text from the image. OCR works by analysing the shape and size of the characters in the image and comparing them to a database of known characters. The OCR software then uses pattern recognition algorithms to identify the characters in the image and convert them into machine-readable text.

4. EXPERIMENT AND RESULTS

4.1 DATASET

There are several publicly available datasets for cyberbullying detection research, including

1. Fine-Grained Balanced Cyberbullying Dataset[1]: It was created by academics at the University of Cagliari and includes 25,000 Facebook and Twitter posts. The dataset includes both cyberbullying and non-cyberbullying messages, and it is balanced to provide an equal amount of good and bad examples.

2. Aggression Parsed Dataset [2]: This dataset contains 20,000 tweets, labeled as containing cyberbullying or not, and is often used for evaluating machine learning models for cyberbullying detection.

3. Hate Speech and Offensive Language dataset: This dataset contains tweets that are labeled as containing hate speech or offensive language. It can be used for cyberbullying detection as well.

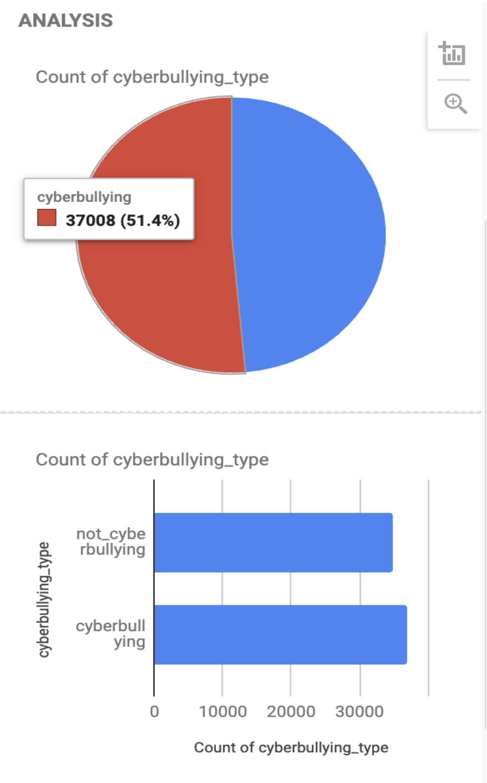

Fig 1: Dataset statistics

4.2 PERFORMANCE EVALUATION

The SVM algorithm achieved an accuracy of 86.862, the k-NN algorithm achieved a higher accuracy of 96.962, the Decision Tree[algorithm achieved an even higher accuracy of 97.972, and the Random Forest algorithm achieved a similar accuracy of 0.972.

Overall, the data indicate that the Decision Tree and Random Forest algorithms performed the best for cyberbullying detection, outperforming SVM and k-NN.

Table -2: Accuracy of different Algorithms

Accuracy vs Algorithms	
Algorithms	Accuracy
Support vector machine(SVM)	0,861998703343521
K-Nearest Neighbors(k-NN)	0,9622117254793
Decision Tree	0,972307122348801
Random Forest	0,972260813188849

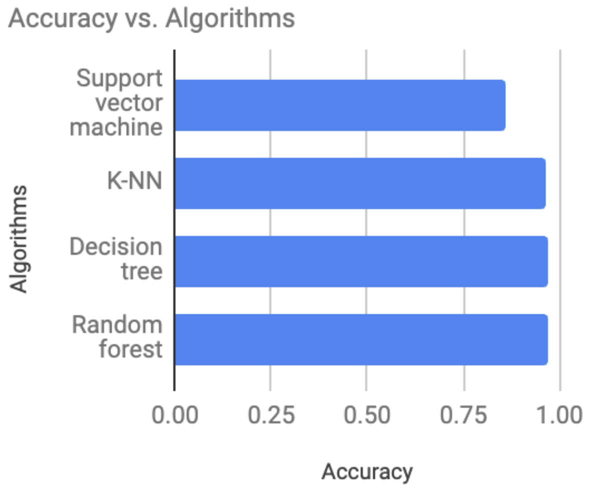

Chart -1: Comparative Study of Machine Learning Algorithms for Cyberbullying Detection

Each algorithm was tested using a dataset containing labeled cyberbullying content, and the reported accuracy is the percentage of occurrences of cyberbullying that was properly recognised.

5. CONCLUSIONS

In conclusion, cyberbullying is a serious problem that can have a significant impact on individuals and society. It is important to develop effective methods for detecting and preventing cyberbullying, and machine learning has emerged as a promising approach for this task. The existing literature on cyberbullying detection using machine learning techniques highlights the potential for deep learning models, natural language processing techniques, and social network analysis to identify patterns of negative behaviour and potential sources of cyberbullying. However, more research is needed to develop more accurate and efficient models that can be deployed in real-world settings. Furthermore, it is important to consider the ethical implications of using machine learning for cyberbullying detection and to ensure that any system developed is fair, transparent, and accountable. Overall, cyberbullying detection using machine learning is a rapidly evolving field, and we will likely see continued progress and innovation in the years to come.

6. FUTURE ENHANCEMENT

Several potential future enhancements could improve the effectiveness of cyberbullying detection using machine learning techniques. Here are a few examples: Many cyberbullying detection systems rely solely on text analysis to identify patterns of negative behaviour. However, incorporating other modalities such as audio, and video could provide additional context and improve accuracy. Cyberbullying behaviours and language can evolve, and static models may not be effective at capturing these changes. Dynamic models that can adapt to changing patterns of behaviour could improve the effectiveness of cyberbullying detection systems. Cyberbullying is a global problem, and many existing systems only analyse text in a single language. The multilingual analysis could improve the ability of these systems to identify cyberbullying across different cultures and languages. Social media platforms have access to vast amounts of data on user behaviour, and collaborating with these platforms to develop more effective cyberbullying detection systems could be a promising avenue for future research.

7. REFERENCES

1. J. Wang, K. Fu, C.T. Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection,” *Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020)*, pp. 1699-1708, December 10-13, 2020.

2. Elsafoury, Fatma (2020), “Cyberbullying datasets”, Mendeley Data, V1, doi: 10.17632/jf4pzyvnpj.1

3. Cyberbullying Detection Using Machine Learning, Aaminah Ali, Adeel M. Syed software Engineering Department, Bahria University, Islamabad, PakistanSoftware Engineering Department, Bahria University, Islamabad, Pakistan.

4. A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning Batoul Haidar*,1, Maroun Chamoun1, Ahmed Serhrouchni2 1Saint Joseph University, Lebanon 2Telecom ParisTech, France.

5. Dadvar, M., Trieschnigg, D., de Jong, F.: Experts and machines against bullies: a hybrid approach to detecting cyberbullies. In: Sokolova, M., van Beek, P. (eds.) AI 2014. LNCS (LNAI), vol. 8436, pp. 275–281. Springer, Cham (2014).

6. Zhang, X., et al.: Cyberbullying detection with a pronunciation-based convolutional neural network. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 740–745 (2016).

7. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. In: Proceedings of the 10th International Conference on Machine Learning and Applications, ICMLA 2011, vol. 2, pp. 241–244 (December 2011).

8. Neelakandan S,1Sridevi M,2Saravanan Chandrasekaran,3Murugeswari K,4Aditya Kumar Singh Pundir,5Sridevi R,6and T.Bheema Lingaiah7 Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media.

9. N. Yuvaraj, K. Srihari, G. Dhiman, et al., "Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking," Mathematical Problems in Engineering, vol. 2021, Article ID 6644652, 12 pages, 2021.

10. S. Mahbub, E. Pardede, and A. S. M. Kayes, "Detection of Harassment Type of Cyberbullying: A Dictionary of Approach Words and its Impact," Security and Communication Networks, vol. 2021, Article ID 5594175, 12 pages, 2021.

11. Md Manowarul Islam; Md Ashraf Uddin; Linta Islam; Arnisha Akter; Selina Sharmin; Uzzal Kumar Acharjee Cyberbullying Detection on Social Networks Using Machine Learning Approaches.

12. J. Wang, R. J. Iannotti, and T. R. Nansel, "School bullying among US adolescents: Physical, verbal, relational and cyber," Journal of Adolescent Health, vol. 45, pp. 368--375, 2009.

13. Maral Dadvar and Kai Eckert Web-based Information Systems and Services, Stuttgart Media University Nebenstrasse 8, 70569 Stuttgart, Germany, Cyberbullying Detection in Social Networks Using Deep Learning Based Models.

14. Cyberbullying Detection on Social Networks Using Machine Learning Approaches Adya Bansal, Akash Baliyan, Akash Yadav, Aman Kamlesh, Hemant Kumar Baranwal Dept. of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, U.P. India.

15. RECENT ADVANCES ON SUPPORT VECTOR MACHINES RESEARCH Yingjie Tian1, Yong Shi2, Xiaohui Liu.

16. KNN Model-Based Approach in Classification, Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer.

17. Cyberbullying detection from text using ensemble classifier technique with base classifier decision trees, Shivam Sarawat.

18. Cyberbullying identification on Twitter using random forest classifier, Novalita Novalita, Anisa Herdiani, Diyas Puspandari.

19. Automated Detection of Cyberbullying Using Machine Learning and OCR, Niraj Nirmal, Pranil Sable, Prathamesh Patil, Prof. Satish Kuchiwale.