Machine Learning Approaches for Cyberbullying Detection in Social Media

Ch. Swathi[1], Ch. Varun Teja[2], G. Gayatri Devi[3], P. Dharan Kumar[4], P. Yaswanth Kumar[5], G.Sateesh[6] [1-5] B. Tech Students, [6]Associate Professor, LIET

[1,2,3,4,5,6] Computer Science and Information Technology, Lendi Institute of Engineering and Technology, Vizianagaram

---***---

ABSTRACT:
Bullying has been an enduring issue throughout human history, with its manifestations evolving over time.The strategies have evolved in tandem with technological progress, shifting from conventional physical bullying to the contemporary issue of cyber bullying. The digital age, marked by the proliferation of user-generated web content, especially on social media platforms, has witnessed a disturbing surge in hate speech. This research project undertakes a systematic review of published studies focusing on approaches to detect cyberbullying and explores methods for identifying hate speech on social media. Additionally, the project conducts a comprehensive comparative study of various supervised machine learning algorithms, encompassing both standard and ensemble methods, to assess their effectiveness in tackling this pressing issue. The ultimate aim is to contribute to the mitigation of hate speech online and its ominous real-world consequences, including violence towards minorities, mass shootings, lynching's, and ethnic cleansing.

Key Words: Cyber bullying, Social media, Machine learning, Literature review, Open challenges

1. INTRODUCTION:

Through Cyberbullying, an individual or victim can be humiliated or hurt before whole network on the web [1]. Cyberbullying significantly impacts a person's mental and physical well-being, leading to a high incidence of depression and suicide. In recent times, the use of images or memes to bully individuals online has become widespread. pornographic images or pictures with oppressive, mean or defamatory remarks are being presented on one's profile in order to bully them. It is essential to implement safeguarding strategies to mitigate and control this issue. Thus, in our work we have made an automated framework which will, in addition to detection of bullying for text [2], also recognizes the bullying in the pictures Images may appear simple, without any text, or they might contain embedded text. Recently, with the rise of memes and GIFs dominating social media feeds, typographic and infographic visual content has emerged as a significant aspect of online data. Consequently, cyberbullying through diverse types of content has become increasingly common. Researchers globally have endeavored to explore innovative methods for identifying, addressing, and mitigating the impact of cyberbullying.Prevalence on social media. Sophisticated analytical techniques and computational models are crucial for the effective processing, analysis, and modeling required to identify bitter, taunting, abusive, or negative content in images, memes, or text messages.The limitations in detecting online bullying posts currently stem from the specificity of social media platforms, dependency on specific topics, and the diversity of manually created features [4].. State-of-the-art results are achieved by deep learning methods on some specific language problems by using the capabilities of hierarchical learning and generalization [5]. Relevant research highlights the application of deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), along with semantic image features, to detect bullying content. These studies focus on examining textual content, image-based elements, and user features.Much of the research into online cyber-aggression, harassment detection, and toxicity has focused primarily on text-based analytics. However, a limited number of studies have also explored visual analysis of images to identify instances of bullying.

2. METHODS AND MATERIAL:

In this project we are using various machine learning algorithms such as SVM, Random Forest, Naïve Bayes, Nearest Neighbors, and Decision Tree to predict child harasser's posts from social networks by employing a comprehensive set of algorithms, we plan to develop a model trained on both standard and harassing words and messages. This model will then be applied to new user posts to predict whether they contain similar inappropriate content. new post is normal or contain harasser's stuff

Support Vector Machine:
The Support Vector Machine (SVM) stands out as a widely embraced Supervised Learning algorithm, catering to both Classification and Regression tasks. Nevertheless, its predominant application resides in tackling Classification challenges within the realm of Machine Learning. The overarching objective of the SVM algorithm revolves around constructing an optimal line or decision boundary, often referred to as a hyperplane. This hyperplane effectively partitions the n-dimensional space into distinct classes, facilitating the seamless categorization of new data points in the future..

SVM (Support Vector Machine) operates by identifying the extreme points or vectors that are pivotal in forming the hyperplane. These critical points are known as support vectors, which is why the method is called Support Vector Machine. Imagine a scenario depicted in a diagram where two distinct categories are separated by a decision boundary or hyperplane. Switching focus to another technique, Random Forest is a well-regarded algorithm within machine learning, falling under the umbrella of supervised learning techniques. It is versatile, being applicable to both classification and regression problems in machine learning. Random Forest is grounded in the principle of ensemble learning, which involves combining several classifiers to tackle intricate problems and enhance the model's accuracy. Essentially, a Random Forest classifier comprises numerous decision trees on different subsets of the dataset and aggregates their outcomes to improve prediction accuracy and control over-fitting.

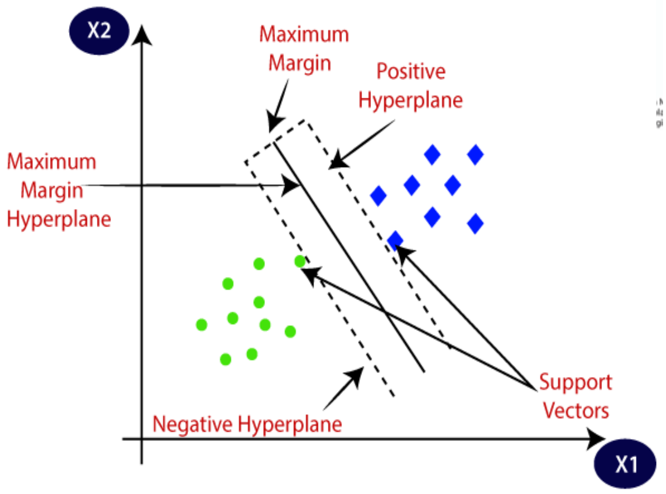
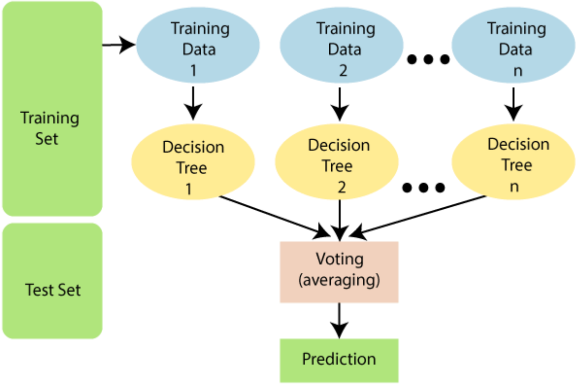


Fig -1: Figure

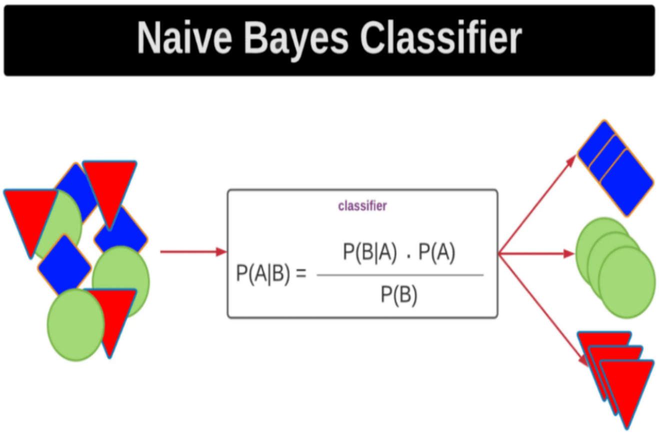
2.1 Random Forest:

Random Forest stands out as a widely used algorithm in machine learning, categorized under supervised learning methods. Suitable for tackling both classification and regression tasks, this approach leverages ensemble learning. This strategy entails integrating various classifiers to address complicated issues effectively and boost the model's accuracy. The essence of Random Forest lies in its assembly of numerous decision trees, each trained on different data subsets from the provided dataset. By averaging the predictions of these trees, Random Forest aims to enhance the overall predictive accuracy and mitigate overfitting.Instead of depending on the prediction of a single decision tree, a random forest aggregates predictions from multiple trees. The final prediction is determined by the majority vote of these individual predictions, contributing to enhanced predictive accuracy for the dataset and it predicts the final output. Increasing the quantity of trees within the forest generally results in improved accuracy and helps to avoid the issue of overfitting, as it allows for a more comprehensive and nuanced understanding of the data. Increasing the quantity of trees within the forest generally results in improved accuracy and helps to avoid the issue of overfitting, as it allows for a more comprehensive and nuanced understanding of the data.



2.2 Naive Bayes classifier:

The Naïve Bayes classifier is a supervised machine learning algorithm primarily employed in classification tasks, with a common application in tasks like text classification.. It is also considered among the generative learning algorithms, aiming to model the distribution of inputs belonging to a specific class or category.



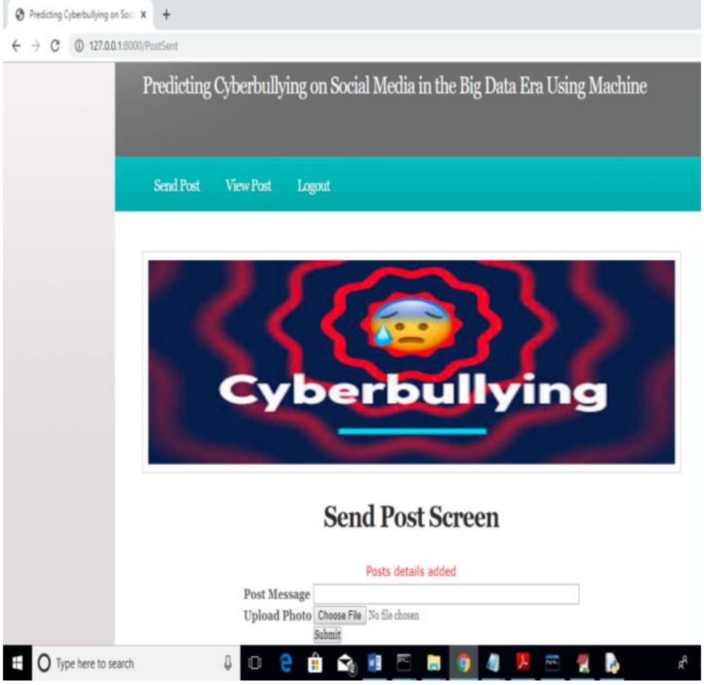
2.3 IMPLEMENTATION:

Modules:

User module: Using this module users can create an account. Using account details they can login to application and then send and view posts.

Admin Module: Admin can view all registered user account and then accept or reject new user account. Admin responsible to add new harasser/non-harasser messages to machine learning train dataset. Admin has to run all or at-least one SVM algorithm to perform harasser's message detection from user side. Admin can view or monitor all posts send by all users.

3. RESULTS:



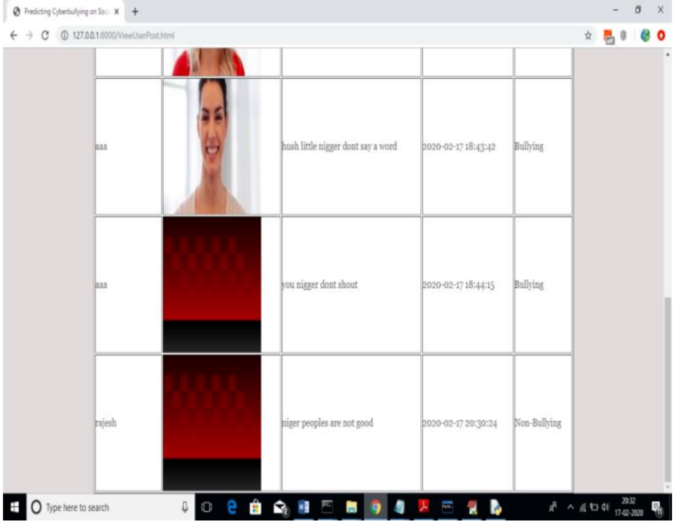
Predicting Cyberbullying on Social Media in the Big Data Era Using Machine

Send Post Screen

Posts details added

Post Message
Upload Photo No file chosen

In the scenario described, messages were posted and a photo was also uploaded.


Predicting Cyberbullying on Soc
127.0.0.1:8000/ViewUserPost.html
Type here to search
20:32
17-02-2020

aaa		hush little nigger dont say a word	2020-02-17 18:43:42	Bullying
aaa		you nigger dont shout	2020-02-17 18:44:15	Bullying
rajesh		niger peoples are not good	2020-02-17 20:30:24	Non-Bullying

In above screen we are seeing posts from all users and post predicted as Non-Bullying. Here based on words given in dataset will get prediction as bullying on non-bullying.

4. CONCLUSION:

Social media and the internet have created avenues for both empowerment and oppression. Platforms intended for meaningful engagement have, at times, turned into spaces where individuals become susceptible to online ridicule, making them vulnerable targets in the digital real.The creation of a Detecting cyberbullying in online content necessitates a crucial predictive model. This study has introduced a prototype model designed specifically for this objective. The uniqueness of the proposed hybrid deep learning model, DLCNN is that it deals with different modalities of content, namely, textual, and info-graphic (text with image).The outcomes have undergone evaluation and comparison against various baselines, revealing that the suggested model exhibits superior performance accuracy. The model's limitations stem from the nature of real-time social data, which is inherently "high-dimensional," meaning it has a vast number of variables, "imbalanced" or skewed, indicating a disproportionate distribution of classes, "heterogeneous," referring to the diverse types of data, and "cross-lingual," highlighting the challenges posed by multiple languages. The growing use of micro-text (wordplay, creative spellings, slangs) and emblematic markers (punctuations and emoticons) further increase the complexity of real-time Cyberbullying detection.

REFERENCES:

1. Hinduja, S., & Patchin, J. W. (2015). Cyberbullying: An update and synthesis of the research. In International handbook of adolescent risk behavior (pp. 107-128). Springer, Cham

2. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E. Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter In the proceedings of the 2017 ACM Web Science Conference, pages 13-22.

3.Raghavendra, B., & Nandakumar, R. (2019). Cyberbullying: Detection and mitigation. In Cyber Threat Intelligence (pp. 225-248). Springer, Cham.

4.Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Cyberbullying Among Youth in the Digital Age: A Comprehensive Review and Meta-Analysis of Researc. Psychological bulletin, 140(4), 1073.

5. Bai, Z.S., Malempati, S. (2023).Addressing Cyberbullying through an Ensemble Approach: Analysis of Text Messages and Images in the Artificial Intelligence Review, 37(1): 179-184.