IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11735

DETECTION OF CYBERBULLYING USING ADVANCED SECURITY

Abhishek R[1], K Sharath[2]

Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India[1]

Assistant Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India[2]

Abstract: Detection and Prevention of Cyberbullying is a natural language processing task, which aims to detect cyberbullying content in tweets which contain text and this text also contains emojis, and also detect the cyberbullying content in images. This has become increasingly important in recent times due to increase in social media activity and as the users increase the misusing of the content also increases. So, the cyberbullying content also has increased a lot in the recent times. To solve this problem, we propose a machine learning model which is trained on various social media content which has been marked manually by annotators as cyberbullying content. We aim to detect cyberbullying content and achieve state of the art performance on a variety of benchmark datasets.

I. INTRODUCTION

In an era dominated by digital communication and online interactions, the rise of cyberbullying has emerged as a formidable challenge, posing serious threats to the well-being and mental health of individuals. As technology continues to evolve, so too does the need for proactive measures to prevent and detect cyberbullying incidents. This essay explores the multifaceted landscape of cyberbullying prevention and detection, emphasizing the importance of education, technological innovations, and collaborative efforts to create a safer and more respectful online environment The pervasive nature of cyberbullying demands a proactive approach that extends beyond traditional methods of conflict resolution. Education stands at the forefront of this battle, as awareness and understanding are key components in fostering responsible digital citizenship. Schools, workplaces, and communities must implement comprehensive programs that educate individuals about the consequences of cyberbullying, promote empathy, and encourage responsible online behavior. By instilling these values early on, we can empower individuals to navigate the digital landscape with a sense of responsibility and respect for others.

Technological advancements, while contributing to the prevalence of cyberbullying, also offer innovative solutions for its prevention and detection. Monitoring tools equipped with sophisticated algorithms can identify patterns indicative of cyberbullying, enabling swift intervention. Social media platforms, being primary arenas for cyberbullying incidents, can implement robust monitoring systems to track and address instances of online harassment. Collaboration with technology companies becomes paramount, urging the development of features and tools that enhance cyberbullying detection, ultimately creating safer online spaces.

Crucially, prevention and detection efforts must extend beyond individual responsibility to encompass a collective and community-oriented approach. Establishing anonymous reporting systems empowers individuals to speak out against cyberbullying without fear of retribution. Community involvement, collaboration with law enforcement, and partnerships with relevant organizations can create a network of support that actively combats cyberbullying on multiple fronts. By fostering a sense of shared responsibility, communities can effectively address the root causes of cyberbullying and work together to create a culture of respect and understanding.

II. PROBLEM STATEMENT

Cyberbullying on social media platforms has emerged as a pervasive threat, causing severe psychological harm to victims. Traditional manual content moderation falls short in capturing the nuances of cyberbullying language and imagery.

To address this, we propose a machine learning approach trained on a diverse dataset, including text, images, Our model aims to automatically detect and classify cyberbullying content, including subtle indicators like hateful language, menacing imagery, and malicious emoji use. This approach offers a comprehensive solution to the challenge of cyberbullying across various media formats.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11735

BASELINE MODELS

- Logistic Regression: It is one of the well-known techniques introduced from the field of statistics by machine learning. Logistic regression is an algorithm that constructs a separate hyper-plane between two datasets utilizing the logistic function. The logistic regression algorithm takes features (inputs) and produces a forecast according to the probability of a class suitable for the input. For instance, if the likelihood is ≥0.5, the instance classification will be a positive class; otherwise, the prediction will be for the other class (negative class) , as given in Equation (1). In, logistic regression was used in the implementation of predictive cyberbullying models.

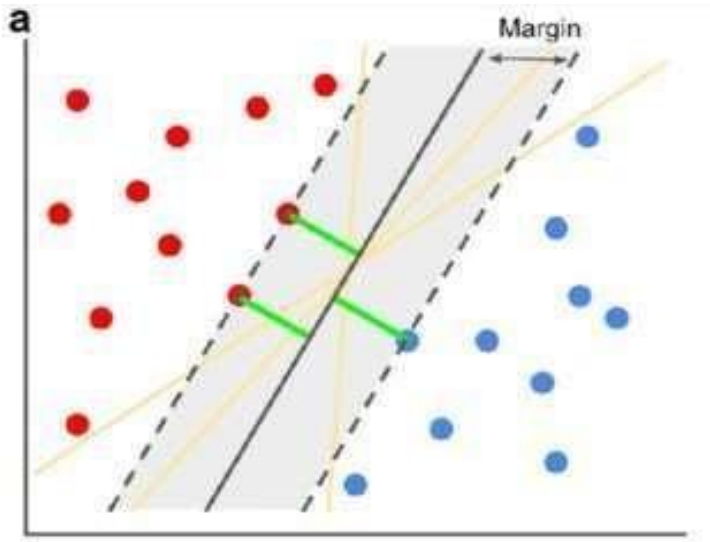
$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}},$$

if $h\theta(x) \geq 0.5$, $y = 1$ (Positive class)
and if $h\theta(x) \leq 0.5$, $y = 0$ (Negative class)

- Random Forest: Random Forest is a versatile ensemble learning method widely applied for both classification and regression tasks in machine learning. Comprising an ensemble of decision trees, the Random Forest algorithm leverages the strength of multiple learners to enhance predictive accuracy and generalization. During training, each tree is constructed on a random subset of the dataset, and their outputs are aggregated through a voting or averaging mechanism, depending on the task. This process introduces diversity among the trees, mitigating overfitting and improving the model's robustness to noise.

Cyberbullying on social media platforms has emerged as a pervasive threat, causing severe psychological harm to victims. Traditional manual content moderation falls short in capturing the nuances of cyberbullying language and imagery. To address this, we propose a machine learning approach trained on a diverse dataset, including text, images,. Our model aims to automatically detect and classify cyberbullying content, including subtle indicators like hateful language, menacing imagery, and malicious emoji use. This approach offers a comprehensive solution to the challenge of cyberbullying across various media formats.

SVC: Support Vector Classification (SVC) is a powerful machine learning model widely employed for binary and multiclass classification tasks. The core principle of SVC involves finding a hyperplane in the feature space that maximally separates instances of different classes. This hyperplane is determined by support vectors, which are the data points closest to the decision boundary. The model aims to maximize the margin between classes while minimizing classification errors. Mathematically, the decision function of an SVC can be represented as)f(x)=sign(w·x+b), where w is the weight vector, x is the input feature vector, b is the bias term, and sign(·)sign(·) is the sign function. SVC is particularly effective in high-dimensional spaces and exhibits robust performance in scenarios with complex decision boundaries. Its versatility and ability to handle non-linear transformations through kernel functions make it a valuable tool in various research domains.



IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11735

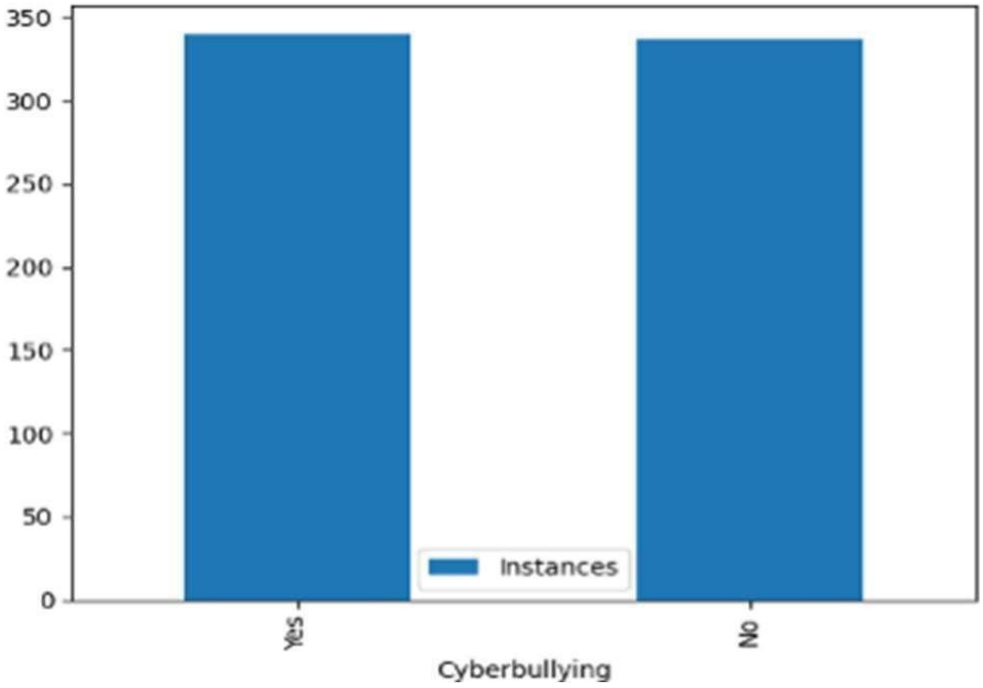
Naive Bayes: Naive Bayes is a probabilistic machine learning model commonly employed for classification tasks, known for its simplicity and efficiency. The model is grounded in Bayes' theorem, leveraging the assumption of independence among features, which simplifies computations and reduces the need for extensive training data.

The Naive Bayes classifier calculates the probability of a data point belonging to a particular class based on the conditional probabilities of its features given the class. The simplicity of its probabilistic approach and independence assumption facilitates quick training and prediction. The fundamental formula for Naive Bayes is expressed as

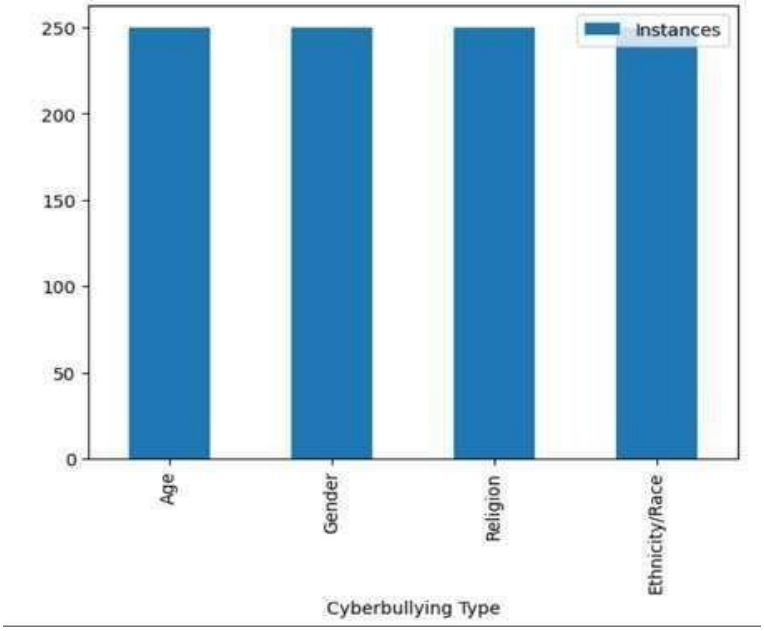
$P(y|x1,x2,...,xn)=P(x1)\cdot P(x2)\cdot...\cdot P(xn)P(y)\cdot P(x1\ |y)\cdot P(x2|y)\cdot...\cdot P(xn|y)$, where y represents the class label, and 1,2,...,$x1$,$x2$,...,xn denote the feature variables. Despite its "naive" assumptions, Naive Bayes often delivers competitive performance and is particularly effective in text classification and spam filtering applications.

DATASET

We have utilized a dataset of over 650 tweets labelled as “Cyberbullying” or “Non-cyberbullying”. To extend the detection we have also used a dataset of about 1000 tweets indicating the type of cyberbullying used. For the detection of cyberbullying in images we have used a dataset of over 750 images which have labelled as “Offensive” or “Non-offensive”. Using the datasets we were able to build to a robust machine learning model.



IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11735



III. METHODOLOGY

Text Classification: With the features in place, the next step involves model training. Various models, including Support Vector Classification (SVC), Logistic Regression, Random Forest Classifier, K Nearest Neighbors, Naïve Bayes, and Extra Tree Classifier, are employed. Following the training process, the model with the highest accuracy is selected for further use—in this scenario, the SVC model proves to be the most accurate.

Type Classification: Building on the extracted features, the data is funneled through various models. The model exhibiting the highest accuracy, identified through extensive training, is the chosen one for type classification—in this instance, the SVC model stands out for its superior performance Image Classification: For image classification, the VGG16 model, a pre-trained model renowned for its efficacy, is employed. Simultaneously, Stacked-LSTM, featuring two layers of LSTM, is utilized for text within the images. The amalgamation of these models forms a robust framework. The combined model is subsequently trained with inputs encompassing both image and text data. The training process involves iterative passes through the dataset, with the epoch yielding the highest accuracy being identified. The model at this pinnacle epoch is then saved for future detection an utilization.

IV. RESULTS

The of different types of models and different models used produces different results and the different f-scores and best of them is Cyberbullying Text Classification.

The Support Vector Classifier (SVC) achieves the highest accuracy (84.56%) and F1 score (83.97%) for cyberbullying text classification, outperforming other models.

Cyberbullying Type Classification
The SVC also demonstrates superior performance in cyberbullying type classification, achieving high accuracy and F1 scores across multiple categories.

Multimodal Cyberbullying Detection
The combination of Stacked Long Short-Term Memory (LSTM) and VGG16 yields a moderate accuracy (69%) for image-based cyberbullying detection. Further optimization and fine-tuning could enhance the model's performance.

Overall Evaluation
The project successfully tackles cyberbullying detection from both text and image perspectives. The SVC emerges as a robust model for text classification and cyberbullying type classification. The combination of Stacked LSTM and VGG16 presents a promising approach for image-based detection.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11735

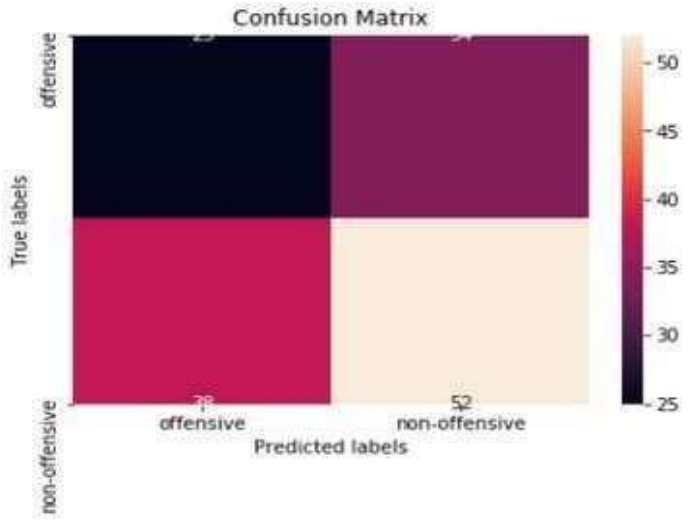


Fig 18 Confusion Matrix of the predictions using the model

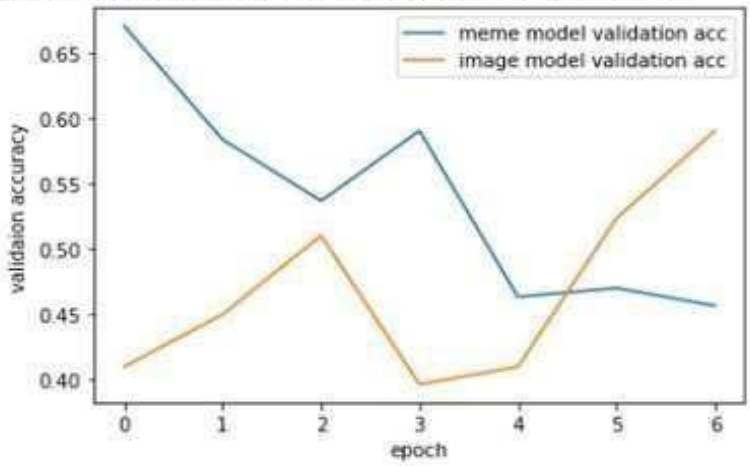


Fig 19 Comparing the combined model with only image model

V. CONCLUSION

In conclusion, the critical importance of continuous advancements in cyberbullying prevention and detection strategies. The examination of existing methodologies reveals progress in addressing online harassment, yet challenges persist. The multidimensional approach, incorporating education, technology, and community engagement, emerges as a key to mitigating cyberbullying's impact. The potential future developments discussed, such as artificial intelligence integration and global collaboration, signify promising avenues for enhanced effectiveness. As technology evolves, so must our strategies to ensure the well-being of individuals in the digital realm.

Ultimately, this research advocates for ongoing research, adaptation, and collaboration among stakeholders to create a safer online environment. By staying attuned to emerging trends and embracing innovative solutions, society can foster a culture of respect, empathy, and responsible digital citizenship, thereby combating the pervasive issue of cyberbullying. The project successfully delves into the multifaceted realm of cyberbullying detection by employing a multi-modal approach that encompasses both text and image analysis.

The SVC model consistently proves to be a robust choice for text classification and cyberbullying type classification, achieving high accuracy and F1 scores. The combination of Stacked LSTM and VGG16 presents a promising avenue for image-based cyberbullying detection, with further optimization and fine-tuning potentially enhancing its performance.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11735

VI. FUTURE SCOPE

The evolving landscape of cyberbullying prevention and detection strategies, exploring current methodologies, technological interventions, and educational initiatives. The study provides an in-depth analysis of existing prevention measures and detection technologies, highlighting their strengths and limitations. Additionally, it discusses the impact of cyberbullying on individuals and communities, emphasizing the importance of a multidimensional approach to address this pervasive issue. The paper also sheds light on the potential future developments in the field, including the integration of artificial intelligence, predictive analysis, and global collaboration. By examining the current state of cyberbullying prevention and detection and proposing future perspectives, this research aims to contribute valuable insights to the ongoing efforts to create a safer and more respectful online environment.

ACKNOWLEDGEMENT

For This paper we would like to express their gratitude to the following individuals and organizations for their invaluable contributions to this research: The Acharya Institute Of Technology and the staff for providing the resources and support necessary to conduct this research.

The project's research guide **Mr. M Gowtham Raj** for their guidance and mentorship throughout the research process. The project's Team members and our fellow students and mentors for their valuable insights and contributions to the project's development.

We would also like to thank the anonymous reviewers for their constructive feedback, which helped to improve the quality of this paper.

REFERENCES

[1]. Aldhyani TH, Al-Adhaileh MH, Alsubari SN. Cyberbullying identification system based deep learning algorithms. Electronics. 2022 Oct 12;11(20):3273.
[2]. Al-Ajlan MA, Ykhlef M. Deep learning algorithm for cyberbullying detection. International Journal of Advanced ComputerScience and Applications. 2018;9(9).
[3]. Aggarwal, K., Bamdev, P., Mahata, D., Shah, R.R. and Kumaraguru, P., 2020. Trawling for trolling: A dataset. arXiv preprint arXiv:2008.00525.
[4]. Roy, P.K. and Mali, F.U., 2022. Cyberbullying detection using deep transfer learning. Complex & Intelligent Systems, 8(6), pp.5449-5467.
[5]. Muneer, A. and Fati, S.M., 2020. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet, 12(11), p.187.
[6]. Mozafari, M., Farahbakhsh, R. and Crespi, N., 2020. A BERT-based transfer learning approach for hate speech detection in online social media. In Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8 (pp. 928-940). Springer
[7]. InternationalPublishing.
[8]. Aluru, S.S., Mathew, B., Saha, P. and Mukherjee, A., 2020. Deep learning models for multilingual hate speech detection. arXiv preprint arXiv:2004.06465.
[9]. Rabbimov, I., Kobilov, S. and Mporas, I., 2021. Opinion classification via word and emoji embedding models with LSTM. In Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings
[10]. 23 (pp. 589- 601). Springer International Publishing.
[11]. Sabat, B.O., Ferrer, C.C. and Giro-i-Nieto, X., 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. arXiv preprint arXiv:1910.02334.
[12]. Vadivukarassi, M., Puviarasan, N. and Aruna, P., 2017. Sentimental analysis of tweets using Naive Bayes algorithm. World Applied Sciences Journal, 35(1), pp.54-59.
[13]. Wu, C.S. and Bhandary, U., 2020, December. Detection of hate speech in videos using machine learning. In 2020 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 585- 590). IEEE.
[14]. Atif, M. and Franzoni, V., 2022. Tell Me More: Automating Emojis Classification for Better Accessibility and Emotional Context Recognition. Future Internet, 14(5), p.142.
[15]. Atif, M. and Franzoni, V., 2022. Tell Me More: Automating Emojis Classification for Better Accessibility and Emotional Context Recognition. Future Internet, 14(5), p.142.

