International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 13 Issue 7 July 2024 || PP 80-87

A Multimodal Approach to Cyberbullying Detection: Integrating Text and Video in Social Media

M. Sathya Devi
Vasavi College of Engineering, Hyderabad.
sathyamaranganti@staff.vce.ac.in

Dr.B.Indira
Chaitanya Bharathi Institute of Technology, Hyderabad.
bindira_mca@cbit.ac.in

Mothe Siddhi Vinayak
Vasavi College of Engineering, Hyderabad,
siddhumothe7@gmail.com

Bhimana Sai Vinay
Vasavi College of Engineering, Hyderabad,
bsaivinay613@gmail.com

Abstract— *In this digital age, protecting oneself online is crucial due to growing worries about abusive and harassing information. This study uses cutting-edge machine learning approaches to tackle this problem from two angles: video content classification and Twitter harassment detection. By utilizing a Twitter dataset, a strong algorithm is created to detect and reduce online harassment, encouraging a safer socialmedia landscape. In order to classify movies as abusive ornormal, the project also makes use of deep learning models, suchas ResNet101 and DenseNet169. This helps to promote a proactive approach to content management. Combining these approaches results in a complete approach to improving digital safety. This project intends to promote pleasant online interactions and add to the existing discussion on ethical machine learning applications through responsible AI practices.*

Keywords— *Online Safety, Machine Learning, Twitter Harassment Detection, Video Content Classification, Deep Learning, ResNet101, DenseNet169, Ethical AI, Social Media, Content Moderation.*

I. INTRODUCTION

The rapid development of the digital landscape in recent years has fundamentally changed communication and information. With the widespread adoption of online platforms, the convenience of quick connection and information dissemination has become a hallmark of today'ssociety. However, this shift to digital connectivity has also exposed significant challenges, particularly related to online harassment and offensive content. The negative impact of such behavior on individuals and communities underscores the urgent need to find effective solutions to mitigate these problems.

To address these pressing issues, this paper presentsa comprehensive approach that focuses on the application ofadvanced machine learning techniques and deep learning models. The main objectivesare twofold: first, to develop a robust algorithm adapted to detect and treat cyberbullying using the rich Twitter dataset, and second, to use advanced deep learning models such as ResNet101 and DenseNet169 to accurately classify videos asViolent or normal. By bringing together these cutting-edge approaches, this paper aims to significantly advance the creation of a safer and more positive online environment.

At the core of the meaning of this work is an unwavering commitment to responsible AI practices. As technology increasingly shapes and influences online communication, implementing machine learning solutions that meet ethical standards is essential. This paper is a key researcher into the potential of artificial intelligence to develop a digital space where users can chat, collaborate and share content without fear of harassment or exposure tooffensive content.

Using machine learning techniques, we strive to equip online platforms with proactive tools that can quickly identify and resolve incidents of online harassment. Using the complete Twitter dataset allows the development of an algorithm that detects subtle patterns of abuse, thus enablingplatform administrators to take quick and targeted action.

This paper aims to use advanced machine learningmodels for accurate detection of text harassment. The modelswill be rigorously trained on vast datasets to identify patternsrelated to abusive language and behavior. Such models enable platforms to respond promptly to harassment andsafeguard users from harm. Due to the growing number of abusive or harmful video content posted online, this paper endeavors to classify the videos.

The goal of using deep learning models like ResNet101 and DenseNet169 is to determine if the videos areabusive or normal. By taking the lead, content moderation becomes feasible and safeguards users from potentially harmful content. The use of proactive moderation is a way toenhance online safety. Ads can be categorized into normal orabusive videos, which allows sites to filter out harmful content before it is published. The use of this technique can result in less harmful content for users, leading to a more positive online community. Combining text harassment detection and video content classification methods is a comprehensive approach to increase digital safety. By utilizing advanced machine learning techniques, the paper seeks to provide a comprehensive solution to onlineharassment from various media sources.

Ethical considerations and responsible technology usage are necessary to address online safety concerns. This work aims to reinforce these values and contribute to the ongoing dialogue on responsible AI. It aims to promote the use of open methods and ethical principles in creating AI systems that are designed to ensure user safety. Beyonddetection and moderation, the paper seeks to promote a culture of digital respect and integrity. Online platforms can promote safer interactions and reduce instances of abuse. A cultural shift towards digital respect is necessary to create a more inclusive and supportive online community. The creation and implementation of machine learning models foronline safety require transparency and accountability. It is important to clarify the measures taken to detect and respondto online harassment, in order to build trust and confidence inthe platform's moderation capabilities.

Through this paper the authors seek to solve the 'multiple problem of cyber security' through innovative machine learning solutions. Its focus is on text harassment detection, and categorizing video to make the Internet safer and more enjoyable for everyone. By focusing on collaboration, ethics,and user empowerment, the project intends to contribute to the ongoing efforts to promote digital safety and integrity.

II. RELATED WORK

The research environment for automated cyberbullying detection systems has made significant progress, reflecting aconcerted effort to address the complex issues arising from cyberbullying and abuse. Balakrishnan et al. [1] investigatedan innovative approach that uses psychological characteristics of Twitter users, such as personality, emotions, and feelings, to develop a robust cyberbullying detection system. The study used models such as the Big Fiveand the Dark Triad. This comprehensive analysis, based on adataset tagged with the #Gamergate hashtag, highlights the potential to incorporate psychological insights into cyberbullying detection.

Cheng et al. [2] presented a hierarchical way to alert, based on network adapted for cyberbullying detection which includes features that mimic the structure of social media sessions. This model uses attentional mechanisms at the word and comment level, taking into account nuanced context analysis.The study highlights the importance of considering differentaspects of cyberbullying dynamics in social media interactions. Visual features complement textual analysis in cyberbullying detection, as Singh et al. [3], emphasizing the need for multimodal approaches to improve forecast accuracy.

Van Hee et al. [4] contributed to the field by curating and annotating a cyberbullying corpus in English and Dutch,allowing automatic detection to be analyzed using binary classification tests. The use of Linear Support Vector Machines (SVM) has shown successful detection of cyberbullying from significant data sources. The study underlines the importance of robust datasets and effective feature selection for improving detection performance.

Despite progress, challenges remain, especially combating cyberbullying in different linguistic contexts. Haidar et al. [5] highlighted cyberbullying detection strategies but found a gap in the study of Arabic content. The integration of machine learning (ML) and natural language processing (NLP) technology offers opportunities for developing recognition skills in different languages.

Agrawal et al. [6] emphasized the importance of a diversity of data sources for comprehensive detection ofcyberbullying. Their study evaluated deep learning (DL)models and learning transfer across multiple social media platforms, providing insight into how well the models perform in different thematic contexts.

Gencoglu et al. [7] emphasized fairness and bias reduction in model training techniques and presented methods to mitigate unwanted biases without compromising model quality. The study highlights the ethical need to develop transparent and unbiased ML solutions for cyber- social health.

Balakrishnan et al. [8] further extended the link between user psychology and cyberbullying by integrating personality traits into detection algorithms. Incorporating RFand core methods into the classification of cyberbullying behaviors highlights the importance of multidisciplinary approaches to effectively address cyberbullying.

Iwendi et al. [9] evaluated the effectiveness of DL models in detecting insults in social comments, using

techniques such as text cleaning and tagging to improve model accuracy. Their findings highlight the potential of bidirectional long-term memory (BLSTM) models in detecting cyberbullying.

Yao et al. [10] introduced CONCISE, a framework for timely detection of cyberbullying in Instagram media sessions. Their sequential hypothesis testing framework minimizes feature requirements while maintaining classification accuracy, illustrating a practical approach to real-time detection.

Raisi et al. [11] focused on extracting seed vocabulary and consistency of user roles to identify cyberbullying across different social media platforms. Their Participant and Vocabulary Continuity (PVC) approach demonstrated the effectiveness of quantitative and qualitativeidentification of cyberbullying behavior.

In summary, recent research highlights the multidisciplinary nature of cyberbullying identification by integrating psychological insights, multimodal analyzes and ethical considerations learning frameworks. These effortshighlight the evolving landscape of responsible AI applications aimed at promoting a safer and more inclusive digital environment.

III. PROPOSED METHOD

Cyberbullying and harassment are two instances of negative online behavior that is negatively affecting people's life. Because of this phenomenon, data-driven and automatedmethods for evaluating and identifying such behaviors are necessary. In this study, we propose a comprehensive methodology utilizing ensemble learning techniques and deepconvolutional neural networks to address online harassment detection in text data and classify videos into abusive or normal categories. The aim is to contribute towards fosteringa safer and more positive online environment by leveraging advanced machine learning algorithms.

A. Text Data Classification:

- *Twitter Text Dataset Preprocessing:*

The text is tokenized to break it down into individual words or tokens. Text cleaning techniques are applied to remove stop words, special characters, and perform stemming or lemmatization to standardize the text format.

- *Feature Engineering:*

Relevant features are extracted from the cleaned text data, such as word frequencies,sentiment scores, and linguistic patterns associated with cyberbullying behavior.

- *AdaBoost Algorithm for Online HarassmentDetection:*

AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm designed to improve the performance of weak classifiers by combining them into a strong classifier. The fundamental principle behind AdaBoost is to train aseries of weak learners sequentially, typically decision trees,each focusing on misclassified instances predicted by the previous model in the sequence.

Here's a step-by-step explanation of how AdaBoostworks:

1. Initialization: Each training instance is initially assigned an equal weight.

2. Training Iterations: AdaBoost iteratively trains asequence of weak learners (e.g., decision trees).In each iteration, the algorithm focuses on the instances that were misclassified by the previous weak learner. The goal is to learn from the mistakes of theprevious model and place more emphasis on challenging instances.

3. Weight Update: After each iteration, the weightsof misclassified instances are increased. This adjustment ensures that subsequent weak learners pay more attention to the instances that were difficult to classify correctly in previous rounds.

4. Combining Weak Learners: The final AdaBoost model is a weighted sum of all the weak learners. Each weak learner's contribution to the final model'sprediction is determined by its accuracy. Models that perform well in detecting online harassment areassigned higher weights, indicating their importancein the ensemble.

5. Final Model Prediction: To make predictions on new data, AdaBoost aggregates the predictions fromall the weak learners, with more weight given to thepredictions of stronger classifiers. The final model's prediction is typically a weighted vote or a combination of the predictions made by the individual weak learners.

In summary, AdaBoost is a powerful ensemble learning technique that trains weak learners sequentially to collectively build a strong classifier. Its ability to learn from misclassified instances andassign higher weights to effective classifiers makesit well-suited for tasks like detecting online harassment in text data. The final AdaBoost model is a weighted combination of these weak learners, leveraging their individual strengths to achieve superior predictive performance.

- *XGBoost Algorithm for Online HarassmentDetection:*

XGBoost is used as another ensemble learning technique known for its efficiency and performance. It constructs a series of decision trees and combines them to enhance predictive accuracy. Regularization techniques and parallelprocessing are employed to improve efficiency andprevent overfitting of the model. Here are the core

concepts of XGBoost:

1. Gradient Boosting Framework:

XGBoost builds an ensemble of weak learners, typically decision trees, in sequential manner. Each new tree is trained to correct the errors (residuals) of the ensemble built so far, thereby reducing the overall prediction error.

2. Decision Tree Construction:

XGBoost constructs a series of decision trees, where each tree learns to capture different patterns in the data. Trees are added sequentially to the ensemble, with each subsequent tree aiming to improve upon the predictions of the previous ones.

3. Regularization Techniques:

XGBoost integrates regularization techniques to control model complexity and prevent overfitting. Common regularization methods include L1 (Lasso) and L2 (Ridge) regularization, which penalize the complexity of the model by adding a regularization term to the objective function.

4. Parallel Processing:

XGBoost is optimized for parallel computation, enabling efficient training on large datasets. It leverages parallel processing capabilities to distribute computation tasks across multiple CPU cores or machines, significantly reducing training time.

B. Abuse/Normal Video Classification using ConvolutionalNeural Networks:

➢ *Data Preprocessing for Video Classification:*

Video frames are extracted to createindividual image datasets. Image resizing ensures consistent dimensions for model input.Data augmentation techniques like rotation and flipping introduce variations in the dataset, enhancing model robustness.

➢ *ResNet101 (Residual Network):*

ResNet, short for Residual Network, is a Deep Convolutional Neural Network (CNN) architecture designed to overcome the challenges of training very deep neural networks, especially the vanishing gradient problem. This architecture introduced the concept of residual learning, which greatly improves the training and convergence of multilayer deep networks.

1. Feature Extraction: ResNet101 is employed invideo frame analysis tasks to extract meaningfulfeatures from individual frames. Each frame is processed through the ResNet101 network, which learns to detect and represent relevant visual patterns and structures.

2. Hierarchical Feature Learning: The deep layersof ResNet101 facilitate hierarchical feature learning, where lower layers capture low-level features (e.g., edges, textures) and higher layerscapture complex and abstract features (e.g., object parts, semantic information).

3. Residual Function Learning: The residual connections in ResNet101 enable the model to focus on learning residual functions, allowing for effective gradient propagation and enhanced model training.

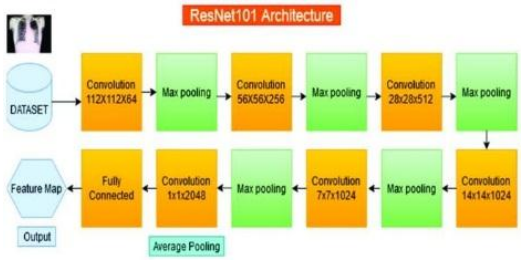

Fig.1. Architecture of Resnet101

➢ *DenseNet169 (Densely Connected ConvolutionalNetworks):*

Densely Connected Convolutional Networks (DenseNet), is a Deep Convolutional Neural Network (CNN) architecture known for its unique connectivity pattern and powerful feature reuse capabilities. DenseNet increases data flow andpromotes feature reuse by connecting each layer directly to every other layer in a dense block. This architecture enables efficient gradient flow, alleviatesthe vanishing gradient problem, and allows efficient learning of hierarchical representations of the input data.

1. *Hierarchical Feature Learning:* DenseNet169, a variant with 169 layers, is employed for video content analysis tasks such as video classification oraction recognition. The deep architecture of DenseNet169 enables hierarchical feature learning, where lower layers capture low-level visual features(e.g., edges, textures) and higher layers capture complex and abstract features (e.g., object parts, semantic information).

2. *Feature Reuse and Gradient Flow:* Denseconnectivity within DenseNet facilitates efficientfeature reuse, allowing the network to learn diverserepresentations of video content across different layers. This connectivity pattern enhances gradient flow during training, leading to more stable and effective learning of hierarchical

representations.

3. *Effective Parameter Utilization:* DenseNet's dense connectivity reduces the number of parameters compared to traditional CNN architectures, makingit more parameter-efficient and enabling effective training on limited computational resources.

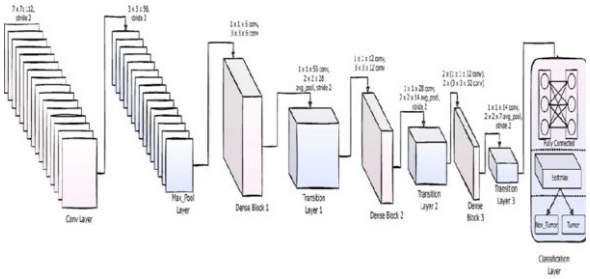

Fig.2. Architecture of Densenet169

➢ *Model Training and Evaluation:*

The preprocessed image data is fed into ResNet101 and DenseNet169 models for training. The models learn to distinguish between abusive andnormal video content based on extracted features. The following evaluation metrics are used to assess the performance of the trained models, Accuracy, Precision, Recall, and F1 score.

IV. RESULTS AND DISCUSSION

The paper concludes with insights on model performance and recommendations for improving cyberbullying detection using machine learning approaches. Actionable steps may include further model tuning, exploring additional features, or integrating with real-time monitoring systems toenhance online safety. The common metrics used for classification of the text are as follows:

➢ **Precision**: Precision measures the accuracy of positive predictions made by the model. For each sentiment class (Negative, Class 1, Class 2, Class 3, Class 4), precision indicates theproportion of true positive predictions (correctly identified instances of that class) among all instances predicted as positive.

➢ **Recall**: Recall (also known as sensitivity or truepositive rate) measures the ability of the model to correctly identify all instances of a particularclass. It is the proportion of true positive predictions among all instances that are actuallyof that class.

➢ **F1-Score:** The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It combinesboth precision and recall into a single metric, especially useful when dealing with imbalanced datasets.

➢ **Support**: Support indicates the number of instances (tweets in this case) for each sentiment class in the test dataset.

➢ **Accuracy:** Overall accuracy represents the proportion of correctly predicted instances (all sentiment classes combined) out of all instancesin the test dataset.

➢ **Macro Average**: Macro average calculates the average of precision, recall, and F1-score across all sentiment classes. It treats each class equally, regardless of class imbalance.

➢ **Weighted Average:** Weighted average calculates the average of precision, recall, and F1-score, weighted by the number of instances for each sentiment class. It provides a more representative measure of overall model performance, considering class imbalances in the dataset.

➢ ***Text Classification using Adaboost Classifier:***

Table.1. Classification Report of Adaboost Classifier

Sentiment	Precision	Recall	F1-Score
Not Cyberbullying	0.71	0.90	0.79
Class 1	0.95	0.77	0.85
Class 2	0.96	0.97	0.97
Class 3	0.96	0.92	0.94
Class 4	0.97	0.93	0.95

The table provided summarizes the performance metrics of amachine learning model (AdaBoost) for cyberbullying detection on a dataset consisting of different sentiment classes. The analysis offers a comprehensive evaluation of the AdaBoost model's effectiveness in cyberbullying detection across multiple sentiment classes. It helps assess the model's ability to correctly classify different types of online content and provides insights into areas of strength and potential improvement. Achieved an overall accuracy of 89.2% in identifying cyberbullying instances within the Textdatasets using AdaBoost Classifier.

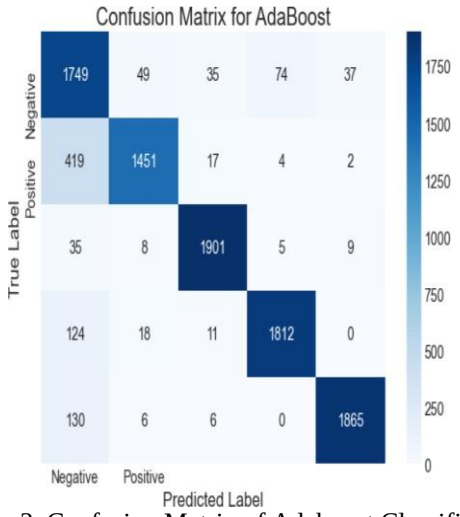

Fig.3. Confusion Matrix of Adaboost Classifier

➢ ***Text Classification using XGBoost Classifier****:*

The table provided summarizes the performance metrics of a machine learning model (AdaBoost) for cyberbullying detection on a dataset consisting of different sentiment classes.. The high precision, recall, and F1-scores demonstrate the model's robustness in identifying and mitigating instances of cyberbullying in text data. Achieved an overall accuracy of 93.2% in identifying cyberbullying instances within the Text datasets using XGBoost Classifier.

Table.2. Classification Report of XGBoost Classifier

Sentiment	Precision	Recall	F1-Score
Not Cyberbullying	0.79	0.92	0.85
Class 1	0.96	0.83	0.89
Class 2	0.99	0.99	0.99
Class 3	0.97	0.94	0.95
Class 4	0.99	0.98	0.98

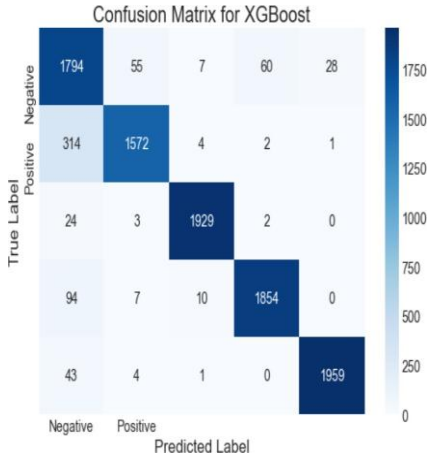

Fig.4. Confusion Matrix of XGBoost Classifier

```
text_input="I'm tired of all the blase fat jokes, rape jokes, gay jokes, etc. NOT FUNNY. #mylife"
ada_prediction, xgb_prediction = predict_sentiment(text_input)
print("AdaBoost Prediction:", ada_prediction)
print("XGBoost Prediction:", xgb_prediction)

AdaBoost Prediction: gender
XGBoost Prediction: gender
```


```
text_input="I'm tired of all the blase fat jokes, rape jokes, gay jokes, etc. NOT FUNNY. #mylife"
ada_prediction, xgb_prediction = predict_sentiment(text_input)
print("AdaBoost Prediction:", ada_prediction)
print("XGBoost Prediction:", xgb_prediction)

AdaBoost Prediction: gender
XGBoost Prediction: gender
```


Fig.5. Classification of Text

Video Classification using Densenet169:

Table.3. Classification Report of Densenet169

Sentiment	**Precision**	**Recall**	**F1-Score**
Abusive	0.94	0.83	0.88
Normal	0.85	0.95	0.90

This table provides a detailed analysis of the model's performance in classifying videos into abusive (Class 0) andnormal (Class 1) categories. It indicates that the model performs well in classifying both abusive and normal videos,with high precision and recall values for each class.

Fig.5. Total Loss vs Total Validation Loss (Densenet)

The weighted average metrics confirm the model's effectiveness in accurately distinguishing between abusive and normal video content, contributing to a safer online environment through proactive content moderation. Achieved an overall accuracy of 93.2% in identifying cyberbullying instances within the Video dataset using Densenet169 architecture.

- ***Video Classification using Resnet101:***

Sentiment	Precision	Recall	F1-Score
Abusive	0.69	0.46	0.55
Normal	0.60	0.79	0.68

Table.4. Classification Report of Resnet101

Table 4 provides a detailed analysis of the model'sperformance in classifying videos into abusive (Class 0) andnormal (Class 1) categories. It indicates that the modelperforms well in classifying both abusive and normal videos,with high precision and recall values for each class. The weighted average metrics confirm the model's effectiveness in accurately distinguishing between abusive and normalvideo content, contributing to a safer online environment through proactive content moderation. Achieved an overall accuracy of 63.2% in identifying cyberbullying instanceswithin the Video dataset using Resnet101 architecture.

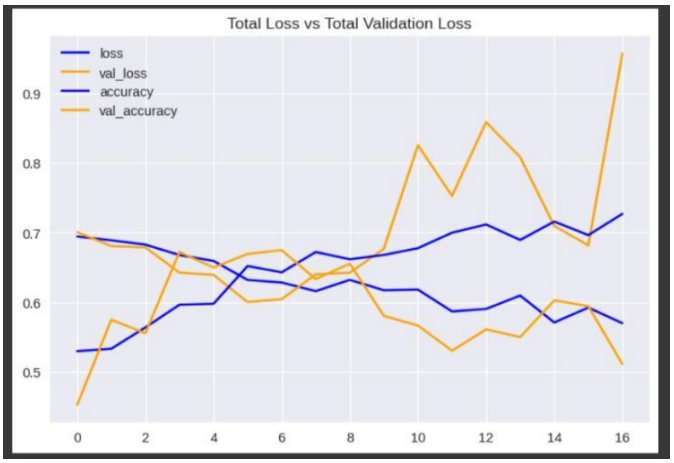

Fig.8. Total loss vs Total validation Loss (Resnet)

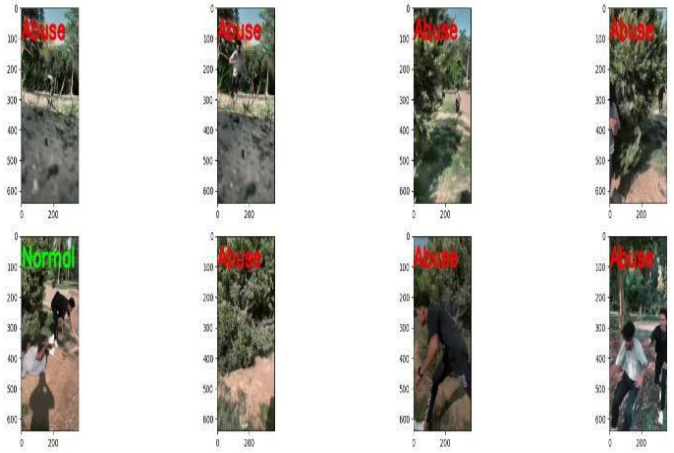

Fig.9. Classification of Video

V. FUTURE SCOPE

In expanding the scope of the paper to include the detection of cyberbullying in audio files, where the model will be detecting and classifying the audio file as bullying ornot. This work can also be extended by working on combined data of text, video and audio where a video with voice and subtitles is taken as input and sentiment of the video will be detected and the videos are classified.

REFERENCES

[1]. [1] Nowland, R., Necka, E. A., & Cacioppo, J. T. (2018). Loneliness and social internet use: pathways to reconnection ina digital world?. Perspectives on Psychological Science, 13(1),70-87.

[2]. [2] Cao, X., Khan, A. N., Ali, A., & Khan, N. A. (2020). Consequences of cyberbullying and social overload while using SNSs: A study of users' discontinuous usage behavior inSNSs. Information Systems Frontiers, 22(6), 1343-1356.

[3]. Kubiszewski, V., Fontaine, R., Potard, C., & Auzoult, L. (2015). Does cyberbullying overlap with school bullying whentaking modality of involvement into account?. Computers in Human Behavior, 43, 49-57

[4]. Watts, L. K., Wagner, J., Velasquez, B., & Behrens, P. I. (2017). Cyberbullying in higher education: A literature review.

[5]. Computers in Human Behavior, 69, 268-274

[6]. Novitasari, N. F., & Hia, N. I. A. (2021). CYBERBULLYING IN MOVIE CYBERBULLY: AN ANALYSIS FROM THE PSYCHOLOGICAL PERSPECTIVE. Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics, 8(1), 44-64.

[7]. Al-Rahmi, W. M., Yahaya, N., Alamri, M. M., Aljarboa, N. A.,Kamin, Y. B., & Saud, M. S. B. (2019). How cyber stalking and cyber bullying affect students' open learning. Ieee Access,7, 20199-20210.

[8]. Cheng, L., Guo, R., Silva, Y., Hall, D., & Liu, H. (2019, May).Hierarchical attention networks for cyberbullying detection on the instagram social network. In Proceedings of the 2019 SIAMinternational conference on data mining (pp. 235-243). Societyfor Industrial and Applied Mathematics.

[9]. Singh, V. K., Ghosh, S., & Jose, C. (2017, May). Toward multimodal cyberbullying detection. In Proceedings of the2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2090-2099).

[10]. Haidar, B., Chamoun, M., & Serhrouchni, A. (2017). A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Advances inScience, Technology and Engineering Systems Journal, 2(6), 275-284.

[11]. Agrawal, S., & Awekar, A. (2018, March). Deep learning for detecting cyberbullying across multiple social mediaplatforms. In European conference on information retrieval (pp. 141-153). Springer, Cham