Team members name: Nikitha Sadananda, Rakshith Venkatesh Murthy Gowda, Yashwanth Rama Krishna Reddy, Dhiraj

Literature Survey:

[https://guof.people.clemson.edu/papers/chi23ai.pdf](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fguof.people.clemson.edu%2Fpapers%2Fchi23ai.pdf&data=05%7C01%7Cdgunashe%40buffalo.edu%7Cf000f62cf0d54c0d41ec08db6aee1b58%7C96464a8af8ed40b199e25f6b50a20250%7C0%7C0%7C638221343119474715%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=wgC2Q3Y3c%2F7dII45Y%2F2ZV43s3tuvRnbnwtXQh3MQIH8%3D&reserved=0)

[https://dl.acm.org/doi/pdf/10.1145/3579630](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdl.acm.org%2Fdoi%2Fpdf%2F10.1145%2F3579630&data=05%7C01%7Cdgunashe%40buffalo.edu%7Cf000f62cf0d54c0d41ec08db6aee1b58%7C96464a8af8ed40b199e25f6b50a20250%7C0%7C0%7C638221343119474715%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=FCukNwieA%2FqLLrvh0aBGpqRrjdEDmpobGFDDUdNrPkc%3D&reserved=0)

<https://cse.buffalo.edu/~hongxinh/papers/NDSS2021A.pdf>

[https://openaccess.thecvf.com/content/CVPR2023/papers/Narayan\_DF-Platter\_Multi-Face\_Heterogeneous\_Deepfake\_Dataset\_CVPR\_2023\_paper.pdf](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fopenaccess.thecvf.com%2Fcontent%2FCVPR2023%2Fpapers%2FNarayan_DF-Platter_Multi-Face_Heterogeneous_Deepfake_Dataset_CVPR_2023_paper.pdf&data=05%7C01%7Cdgunashe%40buffalo.edu%7Cf000f62cf0d54c0d41ec08db6aee1b58%7C96464a8af8ed40b199e25f6b50a20250%7C0%7C0%7C638221343119318484%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=chFL5SsKGphYFJR0OqDf4aQGs4XMH1rDRhWmP4mVKkw%3D&reserved=0)

1) Title: AI-based Detection of Harassment in Virtual Reality (VR)

Problem Statement: Harassment in VR poses a more severe problem compared to text, image, or video-based harassment, as it can leave individuals with mental scars and physical injuries. The current approach to addressing harassment in VR relies on human-based moderation, which can have negative effects on the mental health of moderators. However, this method is limited by the fact that humans cannot be present everywhere simultaneously in the vast VR space. Moreover, human moderation introduces inherent biases, and the process is often slow, relying on user reports and subsequent human review. Additionally, there is no existing automated solution for detecting harassment in VR. Furthermore, the majority vote-based system in VR spaces for removing individuals can lead to potential discrimination against minorities. As a result, there is a critical need to develop an AI-based solution to automatically detect and prevent harassment in VR.

a) Background and Objectives: The background of the problem is derived from a study on the detection of cyberbullying in images, which outlined future work on detecting harassment in videos and VR. Choosing VR as the focus of this research is justified by the fact that harassment in VR can cause significant mental and physical harm. Furthermore, the absence of current automated detection methods and the limitations of human moderation make it essential to explore AI-based solutions. By addressing these challenges, the objective of this project is to develop an AI system capable of detecting and preventing harassment in VR. This solution aims to alleviate the mental health impact on human moderators, prevent all forms of harassment, eliminate bias in handling harassment cases, and proactively intervene to prevent incidents before they occur. These contributions are crucial to building trust in VR platforms, ensuring user safety, and maintaining platform revenue.

b) Potential Contribution: The project holds immense potential to contribute to the problem domain in the following ways:

1. Improved Mental Health for Moderators: AI-based detection of harassment would alleviate the mental health issues faced by human moderators by reducing their exposure to disturbing content and traumatic incidents.

2. Comprehensive Harassment Prevention: By leveraging AI, VR companies can implement proactive measures to prevent all forms of harassment, ensuring a safe and inclusive environment for users.

3. Elimination of Bias: AI systems are capable of handling harassment cases without any inherent bias, ensuring fair treatment and reducing discriminatory outcomes.

4. Preemptive Intervention: With AI monitoring the entire VR space, harassment incidents can be detected and prevented before they occur, mitigating the harm inflicted on individuals.

5. Addressing the Limitations of Human Moderation: AI operates 24/7, providing continuous monitoring and intervention, unlike human moderators who have limitations and cannot work round the clock.

6. Overcoming Discrimination: The automatic detection of discriminatory majority votes allows for the removal of biased features from VR spaces, preventing discrimination against minority groups.

7. Preservation of User Trust: By effectively detecting and preventing harassment, AI-based solutions instill confidence in VR platforms, ensuring users continue to utilize the platform, thus preserving platform revenue.

This contribution is crucial due to the following reasons:

1) Prevention of Mental and Physical Harm: AI-based detection can prevent individuals from suffering mental scars and physical injuries caused by harassment in VR.

2) Addressing the Existing Gap: The absence of automated detection of harassment in VR necessitates the development of AI solutions to tackle this significant problem effectively.

3) Mitigating Negative Impact on Moderators: By reducing the reliance on human moderation, AI can prevent the mental health issues experienced by moderators when exposed to distressing content.

4) Extensive Monitoring in a Vast Space: While humans cannot be present everywhere simultaneously in the expansive VR space, AI can actively monitor the entire environment, ensuring comprehensive coverage.

5) Eliminating Bias: AI systems do not possess inherent biases, thereby providing an impartial approach to handling harassment cases.

6) Swift Detection and Intervention: Unlike the slow process of user reporting and human review, AI can promptly detect and prevent harassment in real-time, minimizing the duration of exposure to harmful situations.

7) 24/7 Monitoring: AI operates continuously, without the limitations of human availability, ensuring round-the-clock surveillance of the VR space.

8) Combating Discrimination: AI's automatic detection of discriminatory majority votes allows for the removal of biased features, preventing discrimination against minorities and promoting inclusivity.

9) Preserving User Trust and Platform Revenue: By effectively addressing harassment and ensuring user safety, AI-based solutions uphold user trust, preventing potential loss of trust in VR platforms and maintaining platform revenue.

2) We queried ~ 100 combination of broader VR terms (i.e., VR, virtual reality, metaverse, social VR), names of mainstream social VR applications (i.e., VRchat, Rec room, Altspace vr, and Horizon Worlds), and synonyms of safety risk (i.e., danger, hatred, issue, risk, poor conduct, safety, harm, privacy, sick, harassment, toxic). We scraped ~ 5000 youtube links. We downloaded 83 videos.

A screen shot of a computer program

Description automatically generated with low confidence

3) Out of the 83 downloaded videos, we manually went through each to remove 74 videos and left with 9.