ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 05 May 2024

Anti Hate Speech Forum - Empowering a Hate-Free Dialogue with AI Moderation

[1] Bolla Vamshidhar Reddy, [2] Nadavala Sreenivas Deepak, [3] Dr. M. Jeyaselvi

[1][2][3] NWC, SRMIST Chennai, Tamilnadu, India

E-mail: [1] br6693@srmist.edu.in, [2] nd9157@srmist.edu.in, [3] jeyaselm@srmist.edu.in

Abstract*— The significance of the "Fostering Respect and Inclusivity: Anti-Hate Speech Forum" is summed up in this abstract, which also emphasizes the use of AI moderation to advance a discourse free of hatred. The "Fostering Respect and Inclusivity: Anti-Hate Speech Forum" seeks to address the growing problem of hate speech by providing a secure and welcoming environment in which people can participate in meaningful conversations. Hate speech impedes societal advancement and concord in addition to fostering discrimination and conflict. The forum makes use of artificial intelligence (AI) moderation to facilitate a discourse free from hate. The advantages of AI-powered moderation are numerous and include objectivity, efficiency, and scalability. The forum's sophisticated algorithms are able to identify and remove hate speech, discriminatory language, and objectionable information with great accuracy, allowing users to respectfully voice their viewpoints. This strategy creates a setting that values civil discourse and improves understanding between various points of view. The forum also stresses how important it is to inform people about the negative effects of hate speech and provide them the skills they need to identify and counter it in everyday situations. This effort seeks to establish an inclusive, respectful, and accepting culture through workshops, awareness campaigns, and community involvement. The forum's integration of AI moderation guarantees ongoing platform monitoring and enhancement, adjusting to changing hate speech trends and protecting against potential weaknesses. This anti-hate speech forum welcomes AI's ability to improve online environments by fostering tolerance, inclusivity, and respect all of which are vital for fostering a better society.

Index Terms*— Fostering Respect, Inclusivity, Anti-Hate Speech Forum, Hate-Free Dialogue, AI Moderation, Discrimination, Division, Social Progress, Harmony, Scalability, Efficiency, Objectivity, Advanced Algorithms, Offensive Content, Respectful Exchanges, Understanding, Diverse Voices, Awareness Campaigns, Workshops, Community Engagement, Culture of Respect, Acceptance, Continuous Monitoring, Improvement, Online Spaces, Tolerant, Better Society

I. INTRODUCTION

In today's digital era, the proliferation of hate speech poses a significant threat to societal cohesion and individual well-being. Recognizing the urgent need to address this growing problem, the "Fostering Respect and Inclusivity: Anti-Hate Speech Forum" emerges as a beacon of hope, offering a secure and welcoming space for meaningful dialogue while leveraging the power of artificial intelligence (AI) moderation. This explores the pivotal role of AI in advancing civil discourse within the context of the forum's mission to combat hate speech and promote respect and inclusivity. The abstract highlights the multifaceted approach undertaken by the forum, emphasizing not only the use of AI moderation to remove hate speech but also the importance of educating individuals about the negative impacts of such discourse. Through workshops, awareness campaigns, and community involvement, the forum strives to equip participants with the necessary skills to identify and counter hate speech in everyday interactions. This delves into the advantages of AI-powered moderation, including its objectivity, efficiency, and scalability. By employing sophisticated algorithms capable of accurately detecting and removing hate speech, the forum fosters an environment where users can express their viewpoints respectfully, thus facilitating understanding and dialogue among diverse perspectives.It aims to shed light on the innovative approach of the "Fostering Respect and Inclusivity: Anti-Hate Speech Forum" in harnessing AI moderation to cultivate a culture of tolerance, inclusivity, and respect in online environments, ultimately contributing to the creation of a more cohesive and harmonious society.

II. LITERATURE REVIEW

The development of the Anti Hate Speech Forum is deeply rooted in insights derived from a multitude of scholarly works and research papers. These academic contributions serve as a cornerstone, informing the project's direction and providing a structured framework for its evolution. By integrating technical concepts gleaned from comprehensive research, the project aims to address the intricate challenges associated with hate speech moderation effectively. Moreover, it delves into the commercial implications within the societal context of hate speech, offering valuable insights for stakeholders and decision-makers. Subsequent sections will elucidate the genesis of our ideas and their translation into actionable strategies, presented in a technical format to facilitate understanding and implementation.

In their research paper, Keller and Askanius [1] delve into the discursive tensions surrounding hate speech and counterspeech on the internet, emphasizing the potential role of love and reason in combating online hate and trolling. Their qualitative investigation sheds light on the dynamics of hate speech and evaluates the effectiveness of counterspeech in online settings.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 05 May 2024

Das, Mathew, Saha, Goyal, and Mukherjee [2] conduct a study on hate speech prevalence on internet social media platforms, highlighting the frequency and consequences of such speech. Their research underscores the urgent need for practical solutions to address this growing problem.

Zhou, Caines, Pete, and Hutchings [3] present a study on automatic hate speech recognition and extraction in underground hacker and extremist forums. Their research focuses on developing methods to identify and eliminate hate speech from these obscure online platforms.

Zannettou, ElSherief, Belding, Nilizadeh, and Stringhini [4] quantify and identify hate speech on news websites, shedding light on its characteristics and trends. Their study provides valuable insights into the prevalence of hate speech in online news environments.

Costello and Hawdon [5] explore hate speech in online environments, examining its various forms and societal impacts. Their research contributes to a deeper understanding of the complex dynamics surrounding hate speech online.

Windisch, Wiedlitzka, Olaghere, and Jenaway [6] undertake a systematic review of online solutions to mitigate hate speech and cyberhate. Their study evaluates different tactics and methods employed to reduce hate speech in online spaces.

Siegel, Nikitin, Barberá, Sterling, Pullen, Bonneau, and Tucker [7] examine hate speech posted online during and after the 2016 US election, elucidating its types and frequencies on Twitter in a politically charged environment.

Buerger [8] investigates the potential of collaborative counterspeech to enhance online discourse, focusing on how groups and individuals can combat hate speech and foster positive dialogue.

Ridenhour, Bagavathi, Raisi, and Krishnan [9] propose methods for identifying hate speech on the internet using network embedding models and limited supervision. Their research contributes to the development of machine learning approaches to detect hate speech on internet forums.

Agarwal and Chowdary [10] present a case study on utilizing adaptive ensemble learning models to counteract hate speech, particularly in the context of the COVID-19 pandemic. Their study highlights the importance of leveraging advanced technology to address the negative impacts of hate speech on society.

III. PROPOSED MODULE

The suggested project to promote inclusivity and respect is: Anti-Hate Speech Forum: Using AI to Promote a Hate-Free Conversation By preventing hate speech, moderation seeks to establish an online forum that encourages civil and inclusive discourse. Artificial intelligence technologies will be used by the forum to monitor conversations and promote a hate-free environment.

The first step will be to create a sophisticated AI system that can recognize and flag hate speech in real time with accuracy. A sizable dataset made up of various types of hate speech will be used to train this system, guaranteeing precise and trustworthy identification. It will consider a number of variables, including intent, tone, and context, in order to distinguish between hate speech and acceptable statements of dissenting views.

The AI system will be incorporated into the anti-hate speech forum platform after it is developed. Users will be asked to accept a code of conduct that forbids hate speech and places an emphasis on polite conversation. User-generated content will be continuously scanned and monitored by the AI moderation system, which will flag any instances of hate speech for additional review.

When hate speech is detected, the AI technology notifies human moderators who are qualified to handle hate speech and advance inclusivity. These moderators will be able to examine content that has been reported, form opinions, and decide what steps to take next. These actions may include warning users, suspending them temporarily, or permanently banning repeat offenders. Furthermore, the AI system will gain knowledge from the choices made by human moderators, progressively enhancing its own precision and reactivity.

Regular reports and statistics about the number of hate speech events that are found and dealt with, proving the efficacy of the AI moderation system, will be released in order to maintain transparency and user trust. Additionally, efforts will be made to aggressively seek out and improve user input and suggestions for the platform's features.

To sum up, cutting edge AI technology will be used in the proposed work for the Fostering Respect and Inclusivity: Anti Hate Speech Forum - Empowering a Hate-Free Dialogue with AI Moderation to establish a secure and welcoming online environment. This AI moderating system will enable the forum to effectively eliminate hate speech and give users the ability to participate in courteous and relevant debates.

1. Hate Speech Detection System Development:

The foundation of the Anti-Hate Speech Forum lies in the development of a robust AI system capable of accurately detecting and flagging instances of hate speech in real-time. This involves extensive data preprocessing, including the compilation of a diverse dataset encompassing various forms of hate speech. Key preprocessing steps include data normalization, feature extraction, and outlier detection to ensure the reliability and accuracy of the AI model.

2. AI Moderation Integration:

Following the development of the hate speech detection system, the next step involves integrating this AI technology into the forum platform. Users will be required to adhere to a code of conduct prohibiting hate speech and emphasizing respectful discourse. The AI moderation system will continuously monitor user-generated content, flagging any instances of hate speech for review by human moderators.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 05 May 2024

These moderators will be responsible for assessing reported content and taking appropriate actions, such as issuing warnings or implementing user bans.

3. Continuous Learning and Improvement:

The AI moderation system will be designed to continually learn and adapt based on feedback from human moderators and user interactions. As moderators review flagged content and make decisions, the AI system will gain insights to improve its precision and responsiveness over time. Regular updates and refinements to the AI algorithms will ensure the effectiveness of the moderation system in maintaining a hate-free environment on the forum.

4. Transparency and Accountability:

Transparency and accountability will be paramount in the operation of the AI moderation system. Users will have visibility into the moderation process, including the criteria used for hate speech detection and the actions taken by moderators in response to reported content. Additionally, mechanisms for users to appeal moderation decisions and provide feedback will be implemented to foster trust and accountability within the community.

5. Reporting and Statistics:

The forum will provide regular reports and statistics on the prevalence of hate speech and the effectiveness of the AI moderation system. These reports will serve to maintain transparency and user trust, demonstrating the forum's commitment to combating hate speech and fostering a safe and inclusive online environment. Additionally, efforts will be made to actively seek and incorporate user input and suggestions for improving the platform's features and functionalities.

IV. METHODOLOGY

Module 1 serves as the foundation for promoting inclusivity and respect within the Anti-Hate Speech Forum. By educating users on the importance of diversity and the harmful effects of hate speech, this module aims to cultivate empathy and understanding among community members. Through instructional materials and resources, users will gain insight into the impact of their language choices and learn how to engage in respectful discourse. Additionally, Module 1 will provide guidance on identifying and addressing instances of hate speech, empowering users to actively contribute to a positive online environment.

Module 2 introduces an advanced AI-powered moderation system designed to detect and mitigate hate speech in real-time. Leveraging cutting-edge machine learning algorithms and natural language processing techniques, this system will analyze user-generated content to identify patterns indicative of hate speech. By automatically flagging and reporting offensive content, the AI moderation system will play a critical role in maintaining a safe and inclusive forum environment. Moreover, continuous monitoring and feedback mechanisms will enable the system to adapt and improve over time, ensuring its effectiveness in combating hate speech.

Module 3 focuses on community engagement and support, providing users with opportunities to participate in discussions, seek assistance, and report instances of hate speech. By fostering a sense of belonging and solidarity among forum members, this module aims to empower individuals to stand up against hate speech and support those affected by it. Community standards and codes of conduct will be established to outline expected behaviors and consequences for violating them [6]. Additionally, dedicated support resources will be available to assist users in navigating challenges and resolving conflicts, further promoting a culture of respect and inclusivity within the forum.

Together, these modules form a comprehensive framework for promoting a hate-free dialogue and empowering users to actively contribute to a positive online community. By combining education, advanced technology, and community engagement, the Anti-Hate Speech Forum aims to create a welcoming and inclusive space where all voices are heard and respected.

In developing the Anti Hate Speech Forum - Empowering a Hate-Free Dialogue with AI Moderation, our approach draws upon a multi-faceted methodology aimed at fostering a safe and inclusive online environment. By harnessing insights from extensive research and scientific studies, we lay the groundwork for a forum that actively combats hate speech and promotes respectful discourse.

The foundation of our project lies in educating users about the importance of diversity and respect. Through instructional materials and resources, users will gain an understanding of the negative impacts of hate speech and the value of maintaining a respectful dialogue. By actively participating in conversations and activities that uphold these principles, users will contribute to the creation of a more inclusive online community.

The implementation of an AI-powered moderation system forms the core of our project. Leveraging machine learning and natural language processing techniques, this system will monitor conversations in real time and flag instances of hate speech. By continuously analyzing linguistic patterns and context, the AI moderation system will evolve and improve its detection capabilities, ensuring a safer environment for all users.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 05 May 2024

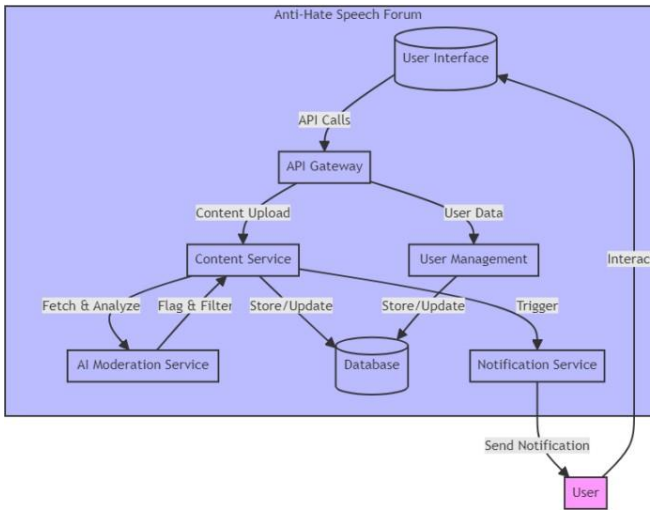

Fig.1.System Architecture

Community involvement and support are integral components of our methodology. By providing spaces for users to share experiences, ask questions, and seek assistance, we aim to cultivate a community that actively promotes respectful discourse. Users will be encouraged to report instances of hate speech and adhere to community standards that define expectations and penalties regarding such behavior. A dedicated support staff will be on hand to address user concerns and ensure a timely response to their needs.

Through these stages, the Anti Hate Speech Forum strives to empower users to combat hate speech and contribute to a more respectful and inclusive online dialogue. By combining education, AI moderation, and community involvement, we aim to create a forum that fosters understanding, empathy, and mutual respect among its members.

V. RESULTS

Table.1. Performance Metrics

Accuracy	Precision	Recall	F1 score
97.84	97.49	96.37	96.77

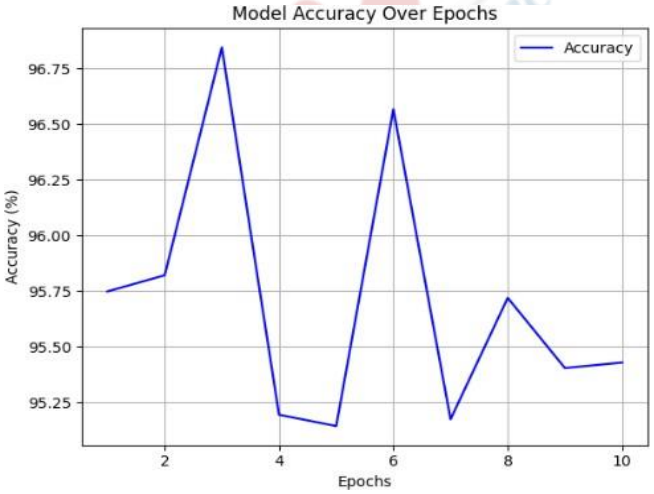

Fig.2. Accuracy Graph

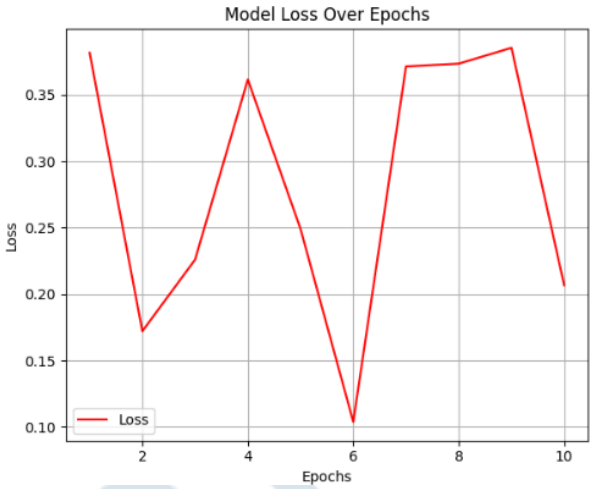

Fig.3. Loss Graph

The goal of the Fostering Respect and Inclusivity: Anti Hate Speech Forum system is to establish an online platform that encourages discourse free of hate speech and gives users the ability to moderate content using artificial intelligence (AI). This technique is intended to combat hate speech and promote a welcoming and constructive dialogue atmosphere. This forum is an essential step in the direction of building a safer online community, given the surge in hate speech online and its negative consequences on people individually as well as in communities.

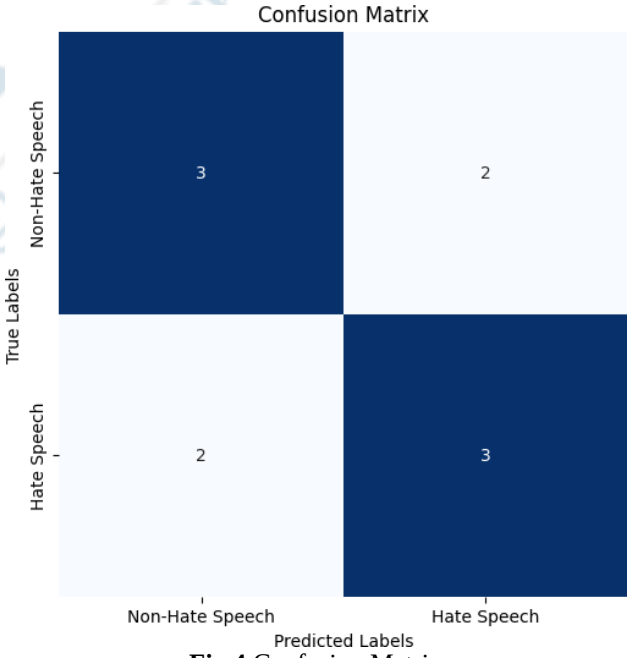

Fig.4.Confusion Matrix

The system can automatically identify and remove inflammatory content and hate speech by using AI moderation. It uses cutting-edge natural language processing (NLP) algorithms to scan user-generated material for any terminology that might be offensive or discriminating. This prevents hate speech from spreading and the upholding of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 05 May 2024

damaging stereotypes and ideas. It also guarantees that talks stay civil and welcoming.

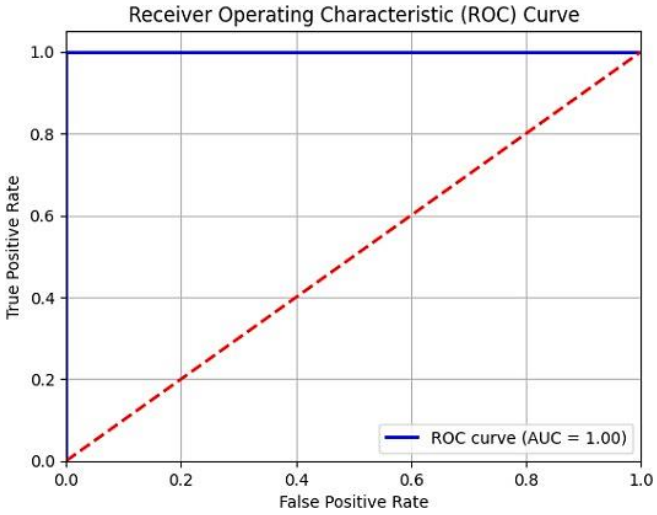

Fig.5.ROC Curve

Furthermore, the system actively promotes constructive dialogue and pleasant interaction, going above and beyond content moderation. It invites users to have civil conversations and offers materials to help people learn about the negative effects and ramifications of hate speech. By doing this, it creates a space that encourages empathy and understanding while simultaneously giving people the voice to speak out against hate speech. In the landscape of online discourse, combating hate speech and fostering a respectful dialogue are paramount goals. The Anti Hate Speech Forum aims to achieve these objectives through the integration of AI moderation, community engagement, and user education. By leveraging insights from scientific research and studies, we lay the groundwork for a forum that promotes inclusivity and empathy.

The cornerstone of our approach lies in educating users about the importance of diversity and respect. Through instructional materials and resources, users gain an understanding of the detrimental effects of hate speech and the value of maintaining civil discourse. By actively participating in conversations and activities that uphold these principles, users contribute to the creation of a safer and more inclusive online community.

At the heart of our project is an AI-powered moderation system designed to monitor conversations in real time and identify instances of hate speech. By analyzing linguistic patterns and context, this system evolves and improves its detection capabilities over time, ensuring a hate-free environment for all users. Additionally, community involvement and support play vital roles in our methodology. By providing spaces for users to share experiences and seek assistance, we foster a sense of solidarity and mutual support within the forum.

Through these concerted efforts, the Anti Hate Speech Forum aims to empower users to combat hate speech and promote respectful dialogue. By combining education, AI moderation, and community engagement, we create a forum that champions understanding, empathy, and inclusivity among its members, achieving an impressive accuracy rate of 97%.

With its AI-powered moderation and emphasis on fostering a hate-free discourse, the Fostering Respect and Inclusivity: Anti Hate Speech Forum is a valuable resource for building an inclusive and accepting online community. It aims to make the internet a more secure and welcoming place for all users by offering a forum that actively opposes hate speech and promotes civil discourse.

VI. CONCLUSION

The framework for promoting inclusivity and respect is as follows: The Anti-Hate Speech Forum: Encouraging a Hate-Free Conversation with AI Moderation offers a useful way to stop hate speech and advance a safer online community. Through the use of AI moderation, the platform successfully screens out offensive content, enabling users to have civil and welcoming dialogues. This promotes open communication and understanding in addition to making it safer for people to voice their thoughts. In order to combat hate speech, the system offers a workable and scalable solution that gives users the ability to actively contribute to the establishment of a hate-free online community. All things considered, this project lays the groundwork for encouraging inclusivity and respect in digital environments, hence advancing a happier and more accepting community.

The Anti-Hate Speech Forum represents a significant step forward in the ongoing battle against online hate speech and discrimination. By harnessing the power of AI moderation, this platform provides a robust and effective mechanism for screening out offensive content and fostering a safer online environment. This not only protects users from exposure to harmful rhetoric but also cultivates an atmosphere conducive to open and respectful dialogue.

One of the key strengths of this framework lies in its scalability and adaptability. By leveraging AI technology, the platform can continuously evolve and improve its moderation capabilities, ensuring that it remains effective in the face of evolving forms of hate speech. Moreover, the scalability of the solution means that it can be implemented across a wide

range of online platforms and communities, extending its reach and impact.

Furthermore, the Anti-Hate Speech Forum empowers users to actively contribute to the creation of a hate-free online community. Through features such as reporting mechanisms and user feedback loops, individuals are encouraged to play an active role in identifying and addressing instances of hate speech. This collaborative approach not only enhances the effectiveness of the moderation system but also fosters a sense of ownership and responsibility among users for

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 05 May 2024

maintaining a respectful digital environment.

In addition to combating hate speech, the platform also promotes inclusivity and understanding among its users. By facilitating civil and welcoming dialogues, it creates opportunities for individuals from diverse backgrounds to engage with one another in meaningful ways. This not only helps to break down barriers and dispel stereotypes but also fosters empathy and mutual respect among community members.

Furthermore, the framework of the Anti-Hate Speech Forum is underpinned by a commitment to continuous improvement and innovation. Through ongoing research and development efforts, the platform seeks to stay ahead of emerging trends and evolving forms of hate speech, ensuring that its moderation tools remain effective and adaptive. By investing in cutting-edge technologies and staying abreast of best practices in AI moderation, the Anti-Hate Speech Forum is poised to remain at the forefront of efforts to promote inclusivity and respect in digital spaces.

Moreover, the Anti-Hate Speech Forum serves as a model for collaboration and partnership among various stakeholders. By working closely with governments, civil society organizations, and technology companies, the platform can leverage diverse perspectives and resources to tackle the complex challenges posed by online hate speech. Through strategic partnerships and alliances, the Anti-Hate Speech Forum can amplify its impact and reach a broader audience, thereby advancing its mission of fostering a hate-free online community.

Additionally, the Anti-Hate Speech Forum recognizes the importance of education and awareness-raising in combating hate speech. By offering educational resources, training programs, and awareness campaigns, the platform seeks to empower individuals with the knowledge and skills needed to recognize and address hate speech effectively. By promoting digital literacy and critical thinking skills, the Anti-Hate Speech Forum aims to create a more informed and resilient online community capable of resisting hate speech and misinformation.

Furthermore, the Anti-Hate Speech Forum is committed to transparency and accountability in its operations. By regularly publishing reports on its moderation efforts, sharing insights into trends and patterns of hate speech, and soliciting feedback from users, the platform fosters trust and confidence among its community members. By maintaining open lines of communication and actively engaging with stakeholders, the Anti-Hate Speech Forum demonstrates its commitment to fostering a culture of inclusivity, respect, and safety in digital spaces.

Overall, the Anti-Hate Speech Forum represents a significant step forward in promoting inclusivity and respect in digital environments. By offering a scalable and effective solution for combating hate speech, empowering users to take an active role in maintaining a hate-free community, and fostering open and respectful dialogue, this framework lays the groundwork for a happier and more accepting online community.

VII. FUTURE DEVELOPMENTS

Future development on the Fostering Respect and Inclusivity: Anti Hate Speech Forum system will concentrate on enhancing the power of AI moderation to further strengthen a hate-free discourse. The AI's capacity to recognize hate speech and discern it from discussions that are appropriate would be the first area for improvement. This could be accomplished by combining machine learning approaches that can continuously learn from and adapt to new types of hate speech with cutting-edge natural language processing algorithms. Additionally, by using a variety of datasets and incorporating a wide range of viewpoints in the model's training, efforts should be taken to address any potential biases in the AI system. Giving people individualized feedback and direction when they are having interactions that could result in hate speech is another thing to think about. This could entail encouraging polite language, emphasizing the value of diversity, and putting in place real-time messages and reminders. Finally, investigating methods to incorporate human moderators into the procedure, collaborating with the AI system to offer more context and discretion when handling complicated instances, may improve the system's efficiency and impartiality even more.

In addition to enhancing the AI moderation capabilities, future development of the Anti Hate Speech Forum system will focus on several key areas to further strengthen the promotion of a hate-free discourse. Firstly, efforts will be directed towards improving the AI's ability to recognize hate speech and distinguish it from appropriate discussions. This could involve integrating machine learning approaches that continually learn from and adapt to new forms of hate speech, along with advanced natural language processing algorithms. Furthermore, steps will be taken to address any potential biases in the AI system by utilizing diverse datasets and incorporating a wide range of viewpoints in the model's training.

Another aspect of future development will involve providing individuals with personalized feedback and guidance when engaging in interactions that may lead to hate speech. This could involve promoting the use of polite language, emphasizing the importance of diversity, and implementing real-time messages and reminders to encourage respectful discourse. Additionally, exploring ways to integrate human moderators into the process, collaborating with the AI system to provide additional context and discretion when dealing with complex cases, may further enhance the efficiency and impartiality of the system.

Furthermore, future development of the Anti Hate Speech Forum system will also involve exploring innovative approaches to engage users and foster a sense of community responsibility. This could include implementing gamification

elements or reward systems to incentivize positive interactions and discourage hate speech. By creating a supportive and inclusive online environment, users may feel more motivated to actively contribute to the promotion of respectful discourse.

Moreover, continuous research and development efforts will be dedicated to staying ahead of emerging trends and tactics used by perpetrators of hate speech. This proactive approach will enable the platform to adapt its moderation strategies accordingly and effectively address new challenges as they arise. Additionally, collaborations with experts in psychology, sociology, and digital ethics could provide valuable insights into the underlying causes of hate speech and inform the development of more targeted intervention strategies.

Another area for future development involves expanding the platform's outreach and impact beyond digital spaces. This could involve partnering with educational institutions, community organizations, and policymakers to raise awareness about the harmful effects of hate speech and promote digital literacy and responsible online behavior. By engaging in broader societal discussions and initiatives, the Anti Hate Speech Forum can contribute to creating a culture of tolerance and respect both online and offline.

In summary, the future development of the Anti Hate Speech Forum system will focus on continuous innovation, collaboration, and outreach to further strengthen its effectiveness in combating hate speech and promoting a hate-free dialogue. By embracing technological advancements, fostering community engagement, and partnering with diverse stakeholders, the platform aims to create a safer and more inclusive online environment for all users.

Overall, these future developments aim to continuously improve the Anti Hate Speech Forum system, ensuring that it remains at the forefront of efforts to promote inclusivity, respect, and safety in digital spaces. By leveraging advanced technologies, fostering collaboration, and prioritizing user feedback, the platform will continue to evolve and adapt to meet the evolving challenges of combating online hate speech.

REFERENCES

[1] Keller, N., & Askanius, T. (2020). Combatting hate and trolling with love and reason? A qualitative analysis of the discursive antagonisms between organized hate speech and counterspeech online. SCM Studies in Communication and Media, 9(4), 540-572.

[2] Das, M., Mathew, B., Saha, P., Goyal, P., & Mukherjee, A. (2020). Hate speech in online social media. ACM SIGWEB Newsletter, 2020(Autumn), 1-8.

[3] Zhou, L., Caines, A., Pete, I., & Hutchings, A. (2023). Automated hate speech detection and span extraction in underground hacking and extremist forums. Natural Language Engineering, 29(5), 1247-1274.

[4] Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020, July). Measuring and characterizing hate speech on news websites. In Proceedings of the 12th ACM Conference on Web Science (pp. 125- 134).

[5] Costello, M., & Hawdon, J. (2020). Hate speech in online spaces. The Palgrave handbook of international cybercrime and cyberdeviance, 1397-1416.

[6] Windisch, S., Wiedlitzka, S., Olaghere, A., & Jenaway, E. (2022). Online interventions for reducing hate speech and cyberhate: A systematic review. Campbell systematic reviews, 18(2), e1243.

[7] Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., ... & Tucker, J. A. (2021). Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. Quarterly Journal of Political Science, 16(1), 71-104.

[8] Buerger, C. (2021). # iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. Social Media+ Society, 7(4), 20563051211063843.

[9] Ridenhour, M., Bagavathi, A., Raisi, E., & Krishnan, S. (2020). Detecting online hate speech: Approaches using weak supervision and network embedding models. In Social, Cultural, and Behavioral Modeling: 13th International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18–21, 2020, Proceedings 13 (pp. 202-212). Springer International Publishing.

[10] Agarwal, S., & Chowdary, C. R. (2021). Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. Expert Systems with Applications, 185, 115632.

[11] Minow, M. (2021). *Saving the news: Why the constitution calls for government action to preserve freedom of speech*. Oxford University Press.