International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 371-373
Published Online June 2022 in IJEAST (http://www.ijeast.com)

ON THE ISSUE OF HATE SPEECH ON SOCIAL MEDIA PLATFORMS

Kavitha SG, Bhoomi KR
Department of IT
SSRV College of Engineering
Kerala, India

***Abstract*— In common language, "hate speech" loosely refer to offensive discourse targeting a group or an individual based on inherent characteristics - such as race, religion or gender - and that may threaten social peace. Under International Human Rights Law, there is no universal definition of hate speech as the concept is still widely disputed especially in regards to its relation to freedom of opinion and expression, non-discrimination and equality. With the aim to provide an unified framework for the UN system to address the issue globally, the United Nations Strategy and Plan of Action on Hate Speech defines hate speech as. "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.**

***Keywords*— hate speech; social media; toxicity detection; user security**

I. INTRODUCTION

While the above is not a legal definition and is broader than the notion of "incitement to discrimination hostility or violence" - prohibited under international human rights law - it highlights three important attributes:

- Hate speech can be conveyed through any form of expression [1], including images, cartoons, memes, objects, gestures and symbols and it can be disseminated offline or online.
- Hate speech is "discriminatory" - biased, bigoted, intolerant - or "pejorative" - in other words, prejudiced, contemptuous or demeaning [2] - of an individual or group
- Hate speech makes reference to real, purported or imputed "identity factors" of an individual or a group in a broad sense: "religion, ethnicity, nationality, race, colour, descent, gender," but also any other characteristics conveying identity, such as language, economic or social origin, disability, health status or sexual orientation, among many others.

It's important to note that hate speech can only be directed at individuals or groups of individuals; therefore, it does not include communication about entities such as States and their offices or symbols, public officials, nor religious leaders, or tenets of faith.

II. OVERVIEW

The proliferation of hateful content online coupled with easily shareable disinformation that digital communication enables has raised unprecedented challenges for our societies as governments struggle to enforce national legislation in the virtual world's scale and speed.

Unlike in traditional media, online hate speech [3] can be produced and distributed easily, at low cost and anonymously while having the potential to reach a global and diverse audience in real time. The relative permanence of online content is also problematic when hateful discourse can resurface and (re)gain popularity over time.

In such a context, understanding and monitoring the dynamic of hate speech across the diverse online communities and platforms are key for shaping new responses; but efforts are often stalled given the sheer scale and diversity of the phenomenon, current technological limitations of automated monitoring systems and the opacity of online companies [4].

Meanwhile, the growing weaponization of social media in order to disseminate hateful and divisive narratives - often promoted by online corporations proprietary algorithms bias - has exacerbated the stigmatization of vulnerable communities and exposed the fragility of our democracies worldwide [5]. This has prompted an increasing scrutiny on internet players and questions on their actual role and responsibility in real world harm. As a result, some States started to hold internet corporations accountable for moderating and removing content that they consider breaking the law, raising concerns about limitation of freedom of speech and censorship in return.

Despite these challenges, the United Nations and many others are exploring further ways of countering hate speech through initiatives that promote greater media and information literacy of online users while ensuring the protection of the right to freedom [6] of expression.

.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 371-373
Published Online June 2022 in IJEAST (http://www.ijeast.com)

II. EXPERIMENT AND RESULT

Upholding free speech is hugely important to open societies that respect human rights. Human Rights Treaties outlaw offensive speech when it poses a risk or threat to others. Speech that is simply offensive but poses no risk to others is generally NOT considered a human rights [7] violation.

Hate Speech becomes a human rights violation if it incites discrimination, hostility or violence towards a person or a group defined by their race, religion, ethnicity or other factors.

Hate Speech typically targets the 'other' in societies. This is manifested through the 'othering' of minority groups such as racial, ethnic, religious and cultural minorities, women and the LGBTQI+ community [8].

In 1997 the Council of Europe issued a recommendation on hate speech which defines it as 'all forms of expression which spread, incite, promote or justify racial hatred [9], xenophobia, anti-Semitism or other forms of hatred based on intolerance'.

The 2019 UN Strategy and Plan of Action on Hate Speech defines it as communication that 'attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factor' [10].

IV.CONCLUSION

The Holocaust and the Rwanadan genocide both illustrate how hate speech can fuel acts of genocide. In current and recent crises, such as the Anglophone Crisis in Cameroon and the treatment of Rohingya Muslims in Myanmar, hate speech has voiced deeply entrenched prejudices and discrimination. It has preceded and accompanied hate crimes [11] and mass atrocities.

Stanton's 10 Stages of Genocide recognise genocide as the outcome of a process beginning with the classification of groups of people, often by race, ethnicity or nationality. While this is not necessarily a linear process, his fourth stage identifies 'dehumanisation' as 'hate propaganda towards a victim group which depicts members as less than human. This can involve equating people with animals, insects or diseases'.

In 2014, the UN produced a Framework [12] for Analysis for Atrocity Crimes which outlined that atrocity crimes are 'not spontaneous or isolated events; they are processes, with histories, precursors and triggering factors'. The framework places emphasis on the prevention of atrocity crimes by identifying a number of risk factors. These include 'enabling circumstances', which involve 'inflammatory rhetoric, propaganda campaigns or hate speech', as well as 'triggering factors', comprising partly of 'acts of incitement or hate propaganda targeting particular groups or individuals'~\cite{oak2019malware}.

Similarly, the Anti-Defamation League [13] models the process of mass atrocities through a Pyramid of Hate, illustrating that genocidal acts cannot occur without being upheld by the lower stages that act as a base for mass atrocities. In the Pyramid, Biased Attitudes, such as stereotypes, misinformation and micro-aggressions, form the bedrock that enables escalation of hate and discrimination. It shows a progression towards Acts of Bias, including dehumanisation and slurs, to Discrimination, Violence and, eventually [14], Genocide.

To date, hate speech is neither wholly defined nor specifically protected against in international human rights law. However, a number of international institutions include provisions which protect against other types of expression, such as incitement to discrimination and dissemination of racist ideas.

Advocacy or promotion of hate

Several international treaties, namely the 1965 International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) and the 1966 International Covenant on Civil and Political Rights (ICCPR), prohibit the advocacy of hate, discrimination, hostility or violence. This is also reflected in the 1969 American Convention on Human Rights (ACHR) [15].

Advocacy, or promotion, implies the speaker intends to encourage these ideas. Crucially, this means that a speaker who uses offensive language with other intentions, for example, for satire, would not be recognised as advocating hate. A speaker that is merely offensive without seeking to encourage hate in others is also not generally recognised as a human rights violation without other aggravating factors. Therefore, there is a cut-off point between speech informed by bias that is acceptable and hate speech that violates human rights. A six point test or checklist has been developed to help analyse the context and determine when offensive speech becomes unlawful.

V. REFERENCE

[1] R. Oak, "A Study of Digital Image Segmentation Techniques," International Journal Of Engineering And Computer Science, 2016.

[2] Oak, R. (2018). A literature survey on authentication using Behavioural biometric techniques. Intelligent Computing and Information and Communication, 173-181.

[3] Oak, R., Du, M., Yan, D., Takawale, H., & Amit, I. (2019, November). Malware detection on highly imbalanced data through sequence modeling. In Proceedings of the 12th ACM Workshop on artificial intelligence and security (pp. 37-48).

[4] Oak, R., & Khare, M. (2017, September). A novel architecture for continuous authentication using behavioural biometrics. In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC) (pp. 767-771). IEEE..

[5] Khare, M., & Oak, R. (2020). Real-Time distributed denial-of-service (DDoS) attack detection using decision trees for server performance maintenance. In

Performance Management of Integrated Systems and its Applications in Software Engineering (pp. 1-9). Springer, Singapore..

[6] Sehwag, V., Oak, R., Chiang, M., & Mittal, P. (2020). Time for a background check! uncovering the impact of background features on deep neural networks. arXiv preprint arXiv:2006.14077.

[7] Jhala, K. S., Oak, R., & Khare, M. (2018, June). Smart collaboration mechanism using blockchain technology. In 2018 5th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2018 4th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom) (pp. 117-121). IEEE..

[8] Oak, R. (2021). The Fault in the Stars: Understanding the Underground Market of Amazon Reviews. arXiv preprint arXiv:2102.04217.

[9] Oak, R. (2019, November). Poster: Adversarial Examples for Hate Speech Classifiers. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (pp. 2621-2623).

[10] Oak, R., Rahalkar, C., & Gujar, D. (2019, November). Poster: Using generative adversarial networks for secure pseudorandom number generation. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (pp. 2597-2599)..

[11] Jain, H., Oak, R., & Bansal, J. (2019, January). Towards Developing a Secure and Robust Solution for E-Voting using Blockchain. In 2019 International Conference on Nascent Technologies in Engineering (ICNTE) (pp. 1-6). IEEE..

[12] Oak, R., Khare, M., Gogate, A., & Vipra, G. (2018, April). Dynamic Forms UI: Flexible and Portable Tool for easy UI Design. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1926-1931). IEEE..

[13] Newman, J. C., & Oak, R. (2020). Artificial Intelligence: Ethics in Practice. login Usenix Mag., 45(1).

[14] C. Hsu and J. Wu, "Multi-resolution Watermarking for Digital Images", IEEE Transactions on Circuits and Systems- II, Vol. 45, No. 8, pp. 1097-1101, August 1998.

[15] R. Mehul, "Discrete Wavelet Transform Based Multiple Watermarking Scheme", in Proceedings of the 2003 IEEE TENCON, pp. 935-938, 2003.