Page 2171 - 2176

***DOI:** https://doij.org/10.10000/IJLMH.116017*

INTERNATIONAL JOURNAL OF LAW MANAGEMENT & HUMANITIES

[ISSN 2581-5369]

Volume 6 | Issue 5

2023

Hate Speech in Social Media

TRIVENI SINGAL[1]

ABSTRACT

Once confined to the barriers of the physical world, hate speech has now spread across the Internet and has become increasingly visible on mainstream social media platforms. Fearing that this harmful rhetoric will incite violence and drive extremism, governments worldwide are passing laws and regulations and pressuring social media companies to implement policies to stop the spread of online hate speech. However, despite its widespread, hate speech does not have a single legal definition. In this article, I discuss the various definitions of hate speech, and its detection and juxtaposition with the freedom of expression citing cases for better understanding.

Keywords: *hate speech, freedom of expression, EU law, social media.*

I. INTRODUCTION

The United Nations has defined hate speech as any form of communication (oral, written, body language, gestures, behavior, etc) that attacks or uses discriminatory or pejorative language regarding a person or a group on the basis of their ethnicity, nationality, gender, age, color, descent, religion, or other such identifying factors.

Under the EU law[2], Member States are obligated to ensure that the following intentional conduct directed towards a group of persons/its member due to their race, color, religion, descent, or national or ethnic origin, is punished -

1) Publicly inciting violence or hatred against the group or its members
2) Publicly condoning, denying, or grossly trivializing the group or its members
3) Crimes of genocide, against humanity, and war crimes against the group or the members'

Social media platforms have also defined hate speech under their terms and conditions for the purpose of moderating user-generated content on their platforms. For example, YouTube's hate speech policy[3] prohibits content that promotes violence or hatred against individuals or groups

[1] Author is a student at Jagiellonian University, India.

[2] REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law (COM/2014/027 final)

[3] *Hate speech policy - youtube help* (no date) *Google*. Available at: https://support.google.com/youtube/answer/2801939?hl=en&sjid=4935284197137386107-EU (Accessed: 24 October 2023).

based on age, caste, ethnicity, race, nationality, etc. Similarly, Twitter's policy on hateful content[4] specifies that "You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease". Twitter is one of the most misused social media platforms used to spread spiteful content that often causes irreparable harm to the subjects. The company exists to promote free expression as a fundamental human right and expressly prohibits hateful conduct. Despite its commitment to combat abuse motivated by hatred, many users still overlook the rules and use their accounts to incite violence and make hurtful comments about others.

Facebook's definition of hate speech[5] does not contain the phrase "incitement to violence", instead identifies hate speech as "content that directly attacks people based on their race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, or gender identity; or serious disabilities or diseases".

In some cases, the content posted on social media is clearly very offensive, while in other cases it is using associational terms that qualify as hate speech.

Hate speech propels the ongoing segregation and radicalization in different societies. It encourages violence against vulnerable or marginalized groups, which has been repeatedly documented in hate crimes across the world. Social media users can express their hate, give their opinions a public dimension, receive applause from friends and followers, and feel somehow validated. The online hate speech can be produced and spread at low cost, does not go through an editing process like other written work, experiences vastly different levels of exposure depending on the popularity of the post, and can be posted cross-nationally, as platform servers and headquarters do not need to be in the same country as the user and their intended audience. Hate speech online can also be available for longer and go through waves of popularity, connect with new networks, or reappear, as well as being anonymous.

(A) Viral nature of online hate speech

People can express their hatred intentionally or unintentionally in various forms. A user of any social media platform may post something that is perceived differently by their followers and hence generate unexpected responses, while in most cases the hateful tone is set by the owner. The viral nature of online hate speech is attributed to three main factors. First, hateful content

[4] *X's policy on hateful conduct | X help* (no date) *Twitter*. Available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (Accessed: 24 October 2023).

[5] *Hate speech* (no date) *Transparency Center*. Available at: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ (Accessed: 24 October 2023).

grows larger, lives longer, and has more structural vitality. Second, hateful content posted by verified users (like influencers, athletes, actors, etc) has more tendency to go viral. Third, the use of hashtags or direct mentions leads to limited virality due to which such hateful content is often posted without any additional text that can stop other users from resharing or reposting it. Unlike positive emotions, negative emotions are known to drive online behaviors such as interactions and sharing, which makes hate speech "infectious". Thus, laws and regulations that promote online literacy are very important solutions to curb the effects and spread of online hate speech. However, it remains a challenge due to the vagueness of metrics used to measure what amounts to hate in posted content.

(B) Detecting Online Hate Speech

The different forms of content that can be posted on the various social media platforms present challenges for developing a foolproof method for detecting hatred sentiments. Whereas screening the text-based microblogging sites may have fairly straightforward criteria, others based on different forms of media, for example, videos, may be challenging to detect. Detection of hate speech online can be done by using modern technologies or by relying on human content moderators who review the content. The latter method is time-consuming, laborious, expensive, and has practical limitations due to the enormous content generated online, and thus, more reliance is placed on technological methods like machine learning, natural language processing, keyword-based approach, distributional semantics, sentiment analysis, deep learning, and so on.

Social media companies have largely shifted from reacting to posts flagged by users as hate speech to proactively detecting and addressing such content through their automated systems before users have seen it. Excessive content removal could create chilling effects and undermine free speech.

Softwares like Hatemeter that detects anti-Muslim hate speech using machine learning and natural language processing techniques, COSMOS that collects and analyses data from Twitter in real-time by keyword specification, using sentiment analysis and natural language processing, and many more have been developed.

(C) Freedom of expression and hate speech

Every person enjoys the freedom of opinion and expression enshrined under Article 19[6] of the Universal Declaration of Human Rights. Constitutions of all countries have incorporated this

[6] "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."

principle in their legal systems. Article 11 of the EU Charter grants a clear right to freedom of expression and Article 10 of the European Convention on Human Rights. However, this freedom is not absolute and is subjected to limitations including public safety, prevention of disorder, prevention of crime, protection of morals, protection of reputation, protection of the rights of others, preventing the disclosure of information received in confidence, etc.

In one of its judgments, the European Court of Human Rights has stated that a person must not resort to unjustified offensive and inappropriate language or go beyond a generally acceptable degree of exaggeration[7].

- **Nix v. Germany[8]**

In this case, it was affirmed that the Internet is a means of communication and the publication of photographs online forms part of the right to freedom of expression. The case concerned Nix's conviction for posting a picture on his blog in 2014 of the former SS chief, Heinrich Himmler, in SS uniform wearing a swastika armband. The Court found that, due to the particular historical context of Germany, the banning of symbols and imagery related to the Nazis was legitimate. The interference was therefore proportionate to the legitimate aim pursued and was thus "necessary in a democratic society". The Court noted that the applicant did not intend to spread totalitarian propaganda, incite violence, or utter hate speech and that his expression had not resulted in intimidation, but, instead, was used for purposes of the blog posts which did possibly contribute to public debate. However, the Court concluded that this was a 'gratuitous use of symbols' that the national law sought to prohibit.

- **Milisavljević v. Serbia [9]**

In this case, it was stated that a fair balance must be struck between the right to freedom of expression and the right to respect for private life while assessing the need to interfere for the protection of the reputation or rights of others. Further, several criteria must be examined to determine this balance including, contribution to the debate on the public interest in the article, popularity of the victim, the subject matter of the publication, the behavior of the person concerned before the publication of the article, the way in which the information is obtained and its truthfulness, the content, form, and consequences of the publication, and the severity of the sanction imposed.

[7] Judgment of the European Court of Human Rights of 5 July 2016, 26115/10

[8] ECtHR decision of 13 March 2018, application No 35285/16

[9] Judgment of the European Court of Human Rights of 4 April 2017, 50123/06

II. Conclusion

Hate speech on social media is a widespread problem that has huge societal significance. In 2021, around 40 % of the U.S. society has personally experienced online hate speech[10]. Hate speech is known to have a negative impact on the mental as well as physical well-being of online users. In particular, young adults tend to suffer from the psychological consequences. Online hate speech further reinforces hateful attitudes (i.e., radicalization) and motivates violent acts including hate crimes. For example, online hate speech has played a crucial role in the 2018 Pittsburgh synagogue shooting, the 2017 Rohingya genocide in Myanmar, and anti-Muslim mob violence in Sri Lanka.

Though social media is a very important platform for connecting with family, friends, and other individuals, it also carries with itself a massive proliferation of hate speech, that causes alarming consequences for the marginalized and vulnerable groups of society. Because it is a huge space, many people believe that the internet is a lawless space where they are allowed to act in the way they want, without being faced with the consequences. That's why it's still common to see intolerant comments on social media. People are careful not to express prejudice and aggressive opinions in real life, away from the screen of computers and smartphones, mainly out of fear of consequences, in the virtual world these hateful behaviors seem to be released. In addition, social networks, online games, forums, and the internet as a whole also need to be active in the fight against this crime.

Hate speech is a type of verbal violence, and its basis is the non-acceptance of differences and intolerance. However, many people claim that freedom of expression gives them the right to express themselves in the way that best suits them which is incorrect. Social media platforms seem to be the perfect place for disseminating hate speech due to virality and anonymity on the internet.

*****

[10] Emily A. Vogels. 2021. The state of online harassment. Pew Research Center (2021). https://www.pewresearch.org/internet/2021/01/13/the-state-of-onlineharassment/

III. References

- Handbook freedom of expression final revised. (n.d.). https://cjc.eui.eu/wp-content/uploads/2020/05/eNACT_Handbook_Freedom-of-expression-compresso.pdf

- Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a67955

- Maarouf, A., Pröllochs, N., & Feuerriegel, S. (2022, October 25). *The virality of hate speech on social media*. arXiv.org. https://arxiv.org/abs/2210.13770

- AlZarouni, E. (n.d.). *Detection of hateful comments on social media*. RIT Scholar Works. https://scholarworks.rit.edu/theses/11176

- Siegel, A. (2020). *Online Hate Speech.* In N. Persily & J. Tucker (Eds.), Social Media and Democracy: The State of the Field, Prospects for Reform (SSRC Anxieties of Democracy, pp. 56-88). Cambridge: Cambridge University Press.

- Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). *How social media users perceive different forms of online hate speech: A qualitative multi-method study.* New Media & Society, 0(0). https://doi.org/10.1177/14614448221091185

- Nazmine, & Manan, Khan & Tareen, Hannan Khan & Noreen, Sidra & Tariq, Muhammad. (2021). *Hate Speech and Social Media: A Systematic Review.* Turkish Online Journal of Qualitative Inquiry. 12. 5285-5294.

*****