Pro Edu. International Journal of Educational Sciences
No. 10, Year 6/2024
https://www.ifiasa.com/peijes e- ISSN 2668-5825, p-ISSN 2668-5817

https://doi.org/10.26520/peijes.2024.10.6.5-18

EXPLORING BOUNDARIES OF SOCIAL MEDIA: FREEDOM OF EXPRESSION HATE SPEECH, AND THE PURSUIT OF A SAFER DIGITAL SPHERE

Riadh MATMAT[1], Mouad LAHMER[2],

[1]University Center of Illizi, [2]University Center of Illizi,
ALGERIA.
[1]E-mail: riadh.matmat@cuillizi.dz, [2]E-mail: mouad.lahmer@cuillizi.dz

ABSTRACT

In the digital age, social media has become a pivotal platform for communication, expression, and information dissemination, fostering free expression and connection on a global scale. However, this freedom comes with a dark side: the proliferation of hate speech, raising significant ethical and legal challenges. This article delves into the complex boundaries of social media, examining the tension between between these two concepts. Real-life case studies vividly illustrate the tangible impact of hate speech on individuals and communities, underscoring the urgent need for intervention. By analyzing current policies, user behavior, and case studies, we explore how social media companies and regulators can balance these competing interests. Our study highlights the need for nuanced approaches to moderation and regulation that uphold the principles of free expression while mitigating the harms associated with hate speech, ultimately contributing to the pursuit of a safer and more inclusive digital sphere.

Keywords: *Social media; free expression; hate speech; content moderation; regulation*;

INTRODUCTION

Social media platforms have become ubiquitous in modern society, transforming the way we communicate, share information, and connect with others. This unprecedented level of connectivity has fostered an environment where freedom of expression can flourish, enabling individuals to share their thoughts, ideas, and experiences with a global audience. In a rapidly evolving world, the reach and impact of social media continue to grow, accompanied by a wide variety of online content created by numerous contributors, making it easily accessible for consumption and interaction (Sazzed, 2023). Remarkably, over 60% of the world's population is actively participating in social media. However, alongside democratization of expression, a darker reality has emerged: the rampant spread of hate speech, so social media platforms have become the main places for the dissemination and proliferation of hate speech (Ayele et al., 2023; Bran & Hulin, 2023; Mathew et al., 2021). Thus, with their wide reach and user anonymity, the immediacy and viral nature of these platforms allows hate speech to spread quickly and widely.

Recognizing hate speech can be tricky because it is a complex issue. There is no clear line between acceptable speech and what crosses the line into hateful language. This lack of well-defined boundaries makes it difficult to identify hate speech in text. Table 1 below presents a compilation of hate speech definitions from various sectors of society (Zhang &

Luo, 2019). The existence of hateful content on social media platforms is evident. This type of content fosters or incites violence, harassment, discrimination, or hostility against individuals or groups based on factors such as race, ethnicity, national origin, sexual orientation, gender, gender identity, religion, age, or disability, and its prevalence has led to increased research in the field, particularly in the areas of regulation, computational linguistics, and discourse analysis.

Source	Definition
The United Nations Strategy and Plan of Action on Hate Speech United Nations Office on Genocide Prevention and the Responsibility to Protect (2021)	Hate speech refers to any form of communication, spoken, written, or behavioral, that attacks or uses derogatory language to target individuals or groups based on their inherent identity factors, such as religion, ethnicity, nationality, race, or gender.
Council of Europe Committee of Ministers (Recommendation No. R (97) 20	Hate speech encompasses all forms of expression that spread, incite, promote, or justify hatred based on various forms of intolerance. This includes racial hatred, xenophobia, and anti-Semitism, as well as intolerance expressed through aggressive nationalism, ethnocentrism, discrimination, and hostility towards minorities, migrants, and people of immigrant origin.
United Nations General Assembly. (1966)	The law forbids promoting hatred based on nationality, race, or religion if it leads to people being discriminated against, feeling hostility, or resorting to violence.
Google. (n.d.). Policies for Content Posted by Users on Search	Any material that encourages violence or intense dislike against someone based on their background, beliefs, abilities, age, where they are from, military service, who they love, or how they identify themselves (including gender). This includes things that are often targeted by prejudice and unfair treatment.
Facebook (Facebook, community standards)	We consider hate speech to be any aggressive language targeting people directly, not ideas or organizations. This includes insults based on things like race, religion, disability, sexual orientation, or gender identity. Hate speech can take many forms, from violent threats to harmful stereotypes, putting people down, showing disgust, or calling for them to be excluded.
X's policy on hateful conduct (2023)	Treat everyone with respect, regardless of their background, beliefs, abilities, age, or who they are. This includes race, ethnicity, religion, and any health conditions.

Instagram (Instagram help center)	Instagram, along with its parent company Meta (which also owns Facebook), defines hate speech as: violent or dehumanizing speech, this includes language that incites violence or promotes violence against people based on their protected characteristics, statements of inferiority, calls for exclusion or segregation and slurs.
TikTok (TikTok help center)	TikTok defines hate speech similarly to other social media platforms, focusing on content that attacks or diminishes individuals or groups based on their characteristics.

Table 1. *Definitions of Hate Speech by Various Sources and Platforms*

Many studies classify content as simply hateful or not hateful. This ignores the crucial role context plays in understanding hate speech. The same words or phrases can convey different meanings depending on the situation. We propose that hate speech exists on a continuum, shaped by its surrounding context and requiring a more context-sensitive classification system (Rawat et al, 2024; Clarke et al, 2023; Abebaw et al, 2022) .

This article delves into the complex and often contentious relationship between freedom of expression and hate speech in the digital realm.

We begin by examining the fundamental principles of freedom of expression, exploring its significance in a democratic society and its role in fostering open dialogue and debate. We then delve into the phenomenon of hate speech, analyzing its various forms, its detrimental effects on individuals and communities, and its potential to exacerbate social divisions and incite violence.

Having established the inherent tension between these two concepts, we turn to exploring potential solutions for navigating this challenging landscape. We examine the role of social media platforms in regulating content, the importance of user education and awareness campaigns, and the need to cultivate a culture of mutual respect and inclusivity online. We also consider the potential for legal frameworks and policy interventions to address the issue of hate speech while upholding the principles of free expression.

In conclusion, this article seeks to provide a nuanced and comprehensive understanding of the interplay between freedom of expression and hate speech in the context of social media. We recognize that this is a complex and multifaceted issue, with no easy solutions. However, by engaging in open and informed dialogue, we can strive to create a safer and more inclusive digital environment where freedom of expression can coexist with respect, tolerance, and the protection of all individuals.

1. FREEDOM OF EXPRESSION ON SOCIAL MEDIA

Freedom of expression is a foundational right that underpins democratic societies, fostering the exchange of ideas, promoting transparency, and enabling social progress.

This section examines the definition and importance of freedom of expression, the legal frameworks that support it, and case studies demonstrating its positive impact on social media.

1.1 Definition and importance of freedom of expression (Bonotti & Seglow, 2021; Alforova et al., 2022)

Freedom of expression encompasses the right to hold opinions without interference and to seek, receive, and impart information and ideas through any media. It is crucial for several reasons:

a) Democratic functioning

- It allows citizens to participate in public discourse, hold governments accountable, and make informed decisions.
- A vibrant public sphere where diverse viewpoints are expressed is essential for a healthy democracy.

b) Personal development

- It supports individual autonomy and personal growth by enabling people to express their identities, beliefs, and opinions.

c) Social progress

- It fosters innovation and progress by allowing the free exchange of ideas and information.
- Historical movements for civil rights, gender equality, and environmental protection have relied heavily on the ability to freely express and disseminate ideas.

1.2 Legal frameworks supporting freedom of expression (Geiger & Izyumenko, 2023)

Several legal instruments at both national and international levels enshrine the right to freedom of expression:

- **First amendment (united states):** The First Amendment to the U.S. Constitution protects the freedom of speech, press, assembly, and the right to petition the government. It prohibits the government from restricting individuals' speech based on its content, although certain limitations apply, such as speech that incites imminent violence or constitutes defamation.
- **Universal declaration of human rights (udhr):** Article 19 of the UDHR states that "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."
- **International covenant on civil and political rights (iccpr):** Article 19 of the ICCPR further elaborates on the right to freedom of expression, recognizing the right to hold opinions without interference and to seek, receive, and impart information and ideas of all kinds.
- **European convention on human rights (ECHR):** Article 10 of the ECHR guarantees the right to freedom of expression, including the freedom to hold opinions and to receive and impart information and ideas without interference by public authority.

1.3 Case studies highlighting the positive impact of free speech on social media

Social media has been a powerful tool for promoting freedom of expression, enabling grassroots movements, and facilitating global dialogue. Here are some notable examples:

- **Arab spring (2010-2011):** social media platforms like Facebook and Twitter played a crucial role in organizing protests and sharing information during the Arab Spring. Activists used these platforms to mobilize supporters, document government abuses, and attract international attention, leading to significant political changes in several countries in the Middle East and North Africa.

- **#Metoo movement (2017-Present):** the #MeToo movement, which started on social media, empowered individuals to share their experiences of sexual harassment and assault. The viral hashtag raised global awareness about the prevalence of sexual misconduct, leading to policy changes, corporate accountability, and a broader cultural shift towards addressing gender-based violence.
- **Black lives matter (2013-Present):** initiated as a hashtag, Black Lives Matter became a global movement advocating against police violence and systemic racism. Social media was instrumental in documenting incidents of police brutality, mobilizing protests, and fostering discussions on racial justice, ultimately influencing public policy and societal attitudes.
- **Climate activism:** young activists like Greta Thunberg have effectively used social media to raise awareness about climate change. Platforms like Twitter and Instagram have allowed climate activists to connect, organize events like Fridays for.

2. THE RISE OF HATE SPEECH

Hate speech has become a significant issue on social media platforms, posing challenges to the ideals of free expression and the safety of online communities. This section explores the examples of hate speech, its psychological and social impacts, and statistical data on its prevalence.

2.1 Examples of hate speech on social media

Here are some examples of hate speech on social media, categorized by the type of abuse:

a) Direct attacks
- Using slurs or derogatory terms to target individuals or groups based on their race, religion, ethnicity, sexual orientation, gender identity, disability, or other protected characteristics.
- Examples: "All [group] are criminals!" "Nobody wants to hear from [group] people."

b) Threats and incitement to violence
- Wishing harm or violence upon individuals or groups.
- Encouraging others to attack or harm specific groups.
- Examples: "[Group] should all be deported!" "We need to take action against these [group] people."

c) Dehumanization
- Language that portrays people or groups as subhuman, inferior, or a threat to society.
- Examples: "[Group] are like animals." "[Group] are trying to destroy our way of life."

d) Stereotypes and generalizations
- Perpetuating negative stereotypes about entire groups of people.
- Making assumptions about individuals based solely on their membership in a particular group.
- Examples: "[Group] are all lazy." "[Group] are only interested in causing trouble."

e) Hateful symbols and imagery
- Sharing symbols associated with hate groups or ideologies that promote violence or discrimination.
- Examples: Sharing images of swastikas or using Pepe the Frog memes in a hateful context.

f) Doxxing

- Sharing private or identifying information about someone online with the intent to harass or harm them.

Even seemingly lighthearted content like sarcasm, memes, and jokes can sometimes be used to deliver hateful messages. hate speech can be both obvious and cleverly hidden.

2.2 Psychological and social impacts of hate speech (Saha et al., 2019)

Hate speech can have profound and far-reaching effects on individuals and society as a whole.

a) Psychological impacts

- **mental health issues:** Exposure to hate speech can lead to anxiety, depression, and other mental health problems. Victims may experience increased stress, fear, and a sense of helplessness.
- **Emotional distress:** Hate speech can cause significant emotional pain, leading to feelings of worthlessness, humiliation, and isolation.
- **Impact on self-esteem:** Persistent exposure to hate speech can erode an individual's self-esteem and self-worth, particularly among young people and marginalized groups.

b) Social impacts

- **Polarization and Division:** Hate speech exacerbates social divides and fosters an environment of hostility and mistrust between different communities.
- **Marginalization:** It reinforces existing prejudices and stereotypes, contributing to the further marginalization and exclusion of targeted groups.
- **Violence and radicalization:** Hate speech can incite real-world violence and contribute to the radicalization of individuals, leading to hate crimes and extremist actions.
- **Undermining social cohesion:** The pervasive presence of hate speech undermines efforts to build inclusive and cohesive societies, making it difficult to achieve social harmony and mutual respect.

2.3 Statistical data on the prevalence of hate speech online

Quantifying the prevalence of hate speech online is challenging due to its evolving nature and the varied definitions used by different platforms and studies. However, several reports and studies provide insights into its scope.

a) Global statistics

- A report by the Anti-Defamation League (ADL) found that 41% of Americans have experienced online harassment, with a significant portion of these incidents involving hate speech.
- A 2020 Pew Research Center survey reported that 35% of social media users in the U.S. have witnessed someone being harassed online due to their race or ethnicity.

b) Platform-specific data

- **Facebook:** In its Community Standards Enforcement Report for Q2 2021, Facebook reported taking action on 31.5 million pieces of hate speech content(the highest number), up from 6 million in Q1 2020, and down to 9.6 in Q3 2023.
- **Twitter:** Twitter's transparency report for the second half of 2020 indicated that the platform took action on 3.8 million accounts for violating its rules on hateful conduct.
- **YouTube:** YouTube's 2020 transparency report revealed that the platform removed over 200,000 videos for hate speech violations in Q2 2020 alone.

c) Regional variations

- The prevalence of hate speech varies by region, often reflecting broader societal tensions. For example, Europe has seen a rise in hate speech against immigrants and refugees, while in Asia, there has been an increase in religiously motivated hate speech.

3. REGULATORY AND PLATFORM RESPONSES

This section provides an overview of existing regulations, explores the policies and measures implemented by major social media platforms, and evaluates the effectiveness and limitations of these responses.

3.1 Overview of existing regulations addressing hate speech on social media

Various countries have enacted laws and regulations aimed at curbing hate speech online, reflecting diverse legal frameworks and cultural contexts.

a) United States

- **First amendment:** The U.S. protects free speech under the First Amendment, making it challenging to regulate hate speech unless it directly incites violence or constitutes a true threat. However, platforms can establish their own community guidelines.

b) European Union

- **General Data Protection Regulation (GDPR):** While primarily focused on data protection, GDPR also mandates transparency in how personal data, including content moderation data, is handled.
- **Digital Services Act (DSA):** Proposed legislation aimed at increasing the responsibilities of online platforms to tackle illegal content, including hate speech. It emphasizes transparency, accountability, and enhanced user rights.
- **Eu code of conduct on countering illegal hate speech online:** A voluntary agreement with major tech companies to remove illegal hate speech within 24 hours of notification.

c) Germany

- **Network enforcement act (NetzDG):** Requires social media platforms with over 2 million users to remove "obviously illegal" content within 24 hours and other illegal content within seven days. Non-compliance can result in significant fines.

d) United Kingdom:

- **Online safety bill:** A proposed regulation that imposes a duty of care on platforms to protect users from harmful content, including hate speech. It includes provisions for fines and sanctions for non-compliance.

e) Other countries

- Various nations, including Australia, India, and Canada, have introduced or are considering regulations to address online hate speech, each reflecting their unique legal and social landscapes.

3.2 Policies and measures implemented by major social media platforms

Major social media platforms have developed extensive policies and measures to combat hate speech, leveraging both human moderators and advanced technologies.

a) Facebook

- **Community standards:** Facebook's guidelines explicitly prohibit hate speech and outline the types of content that are not allowed.

- **Content moderation:** Utilizes a combination of artificial intelligence and human moderators to detect and remove hate speech. Reports indicate billions of pieces of content being acted upon annually.
- **Algorithm adjustments:** Algorithms are continually refined to better identify and limit the reach of hate speech, while also prioritizing content from trusted sources.

b) Twitter

- **Hateful conduct policy:** Twitter's policy bans the promotion of violence, threats, and harassment against people based on their identity.
- **Content moderation tools:** Includes automated systems to detect and flag hate speech, as well as manual review processes. Users can also report hate speech, which is then reviewed by Twitter's moderation team.
- **Labeling and deprioritization:** Tweets that violate policies but do not warrant removal are often labeled and deprioritized in users' feeds.

c) YouTube

- **Community guidelines:** Prohibits content that promotes hatred against individuals or groups based on various protected characteristics.
- **Machine learning and human review:** Uses machine learning algorithms to detect hate speech, with flagged content reviewed by human moderators.
- **Educational resources**: Provides resources to educate users on hate speech and encourages positive engagement.

d) Instagram

- **Community guidelines:** Similar to Facebook, Instagram's policies ban hate speech and provide detailed criteria for what constitutes such content.
- **Moderation and AI:** Uses AI to detect hate speech in comments, posts, and messages, supplemented by human review.
- **User controls**: Allows users to filter out offensive comments and report hate speech easily.

3.3 Effectiveness and limitations of current regulatory and platform responses

While significant efforts have been made to address hate speech, challenges and limitations remain.

a) Effectiveness

- **Rapid response:** Some regulations, like Germany's NetzDG, have improved the speed at which platforms remove illegal content.
- **Increased accountability:** Laws like the EU's proposed DSA aim to increase transparency and accountability, encouraging platforms to take their responsibilities seriously.
- **Technological advancements:** Platforms' use of AI and machine learning has improved the detection and removal of hate speech at scale.

b) Limitations

- **Over-Censorship:** There is a risk of over-censorship, where legitimate content is mistakenly removed, impacting free expression. Automated systems are not always accurate and can flag non-offensive content.
- **Evasion techniques:** Perpetrators of hate speech often find ways to evade detection, such as using coded language or private groups.
- **Inconsistent enforcement:** Enforcement of policies can be inconsistent, with some high-profile or controversial cases slipping through the cracks.

- **Jurisdictional conflicts:** The global nature of social media means that a platform might be compliant with regulations in one country but face legal challenges in another.
- **Resource constraints:** Smaller platforms may lack the resources to implement robust content moderation and compliance mechanisms, leading to inconsistent application of policies.

4. CASE STUDIES

Here is a table summarizing the key points related to the dynamics of hate speech on social media and the impact on various groups, along with the responses and strategies to counteract it:

Category	Specific Examples	Impacts	Responses & Strategies
Hateful Rhetoric Against Public Figures	Politicians, journalists, women facing online abuse	Emotional harm, threats of violence, suppression of free speech	Support networks, stricter platform guidelines
Coordinated Hate Campaigns	Organized groups spreading hate speech and misinformation	Manipulation of public opinion, targeted harassment	Enhanced moderation, cross-platform collaboration
Gamification of Hate	Normalization of hate speech in online games and forums	Desensitization to hate speech, blurred lines between online and offline behavior	Monitoring gaming communities, promoting positive behavior
Algorithmic Amplification	Social media algorithms amplifying hateful content	Creation of echo chambers, reinforcement of biases	Algorithm adjustments, transparency in content recommendations
Dogpiling and Harassment	Coordinated harassment campaigns, doxing	Emotional distress, fear, suppression of free speech	Improved reporting systems, support for victims
Exploiting Tragedies	Blaming entire communities for mass shootings or other tragedies	Increased marginalization, incitement of violence	Prompt content moderation, factual counter-narratives
Hate Speech Disguised as "Free Speech"	Users defending hateful rhetoric under free speech	Difficulty in constructive conversations, hostile climate for marginalized voices	Clear definitions of hate speech, education on free speech limits
Hate Speech Targeting People With	Mockery, exclusion, misinformation about disabilities	Emotional harm, social exclusion, spread of harmful	Education on disabilities, stricter content moderation

Disabilities		stereotypes	
Hate Speech Coded in Symbols and Emojis	Use of coded language to bypass filters	Difficulty in detection, persistence of hateful content	Adaptation of moderation strategies, ongoing vigilance
Global Landscape of Hate Speech	Manifestation across cultures and languages	Diverse impacts depending on regional context	International collaboration, localized moderation efforts
Hate Speech Disguised as "Patriotism"	Nationalistic fervor justifying hate against immigrants	Suppression of nuanced discussions, increase in xenophobia	Promoting inclusive narratives, education on national identity issues
Hate Speech Targeting Religious Symbols	Attacks on specific religions or religious symbols	Increased religious tensions, potential for violence	Interfaith dialogue promotion, quick response to inflammatory content
Hate Speech and Deepfakes	Manipulated media spreading hate	Convincing false narratives, difficulty in identifying sources	Advanced detection tools, collaboration with tech companies
Hate Speech Leveraging Current Events	Exploitation of news events or social movements for hate	Increased targeting of specific populations, spread of misinformation	Timely fact-checking, counter-narratives
Hate Speech Targeting Physical Appearance	Body shaming and other forms of appearance-based hate	Negative impact on self-esteem, mental health issues	Promoting body positivity, swift action on hateful content
Hate Speech Disguised as Academic Discourse	Hateful ideologies masked as intellectual debate	Difficulty in challenging hate, spread of sophisticated rhetoric	Critical discourse analysis, promoting inclusive academic discussions
Hate Speech Targeting Indigenous Communities	Harassment and hate speech against indigenous cultures and rights	Impact on social cohesion, cultural preservation efforts	Support for indigenous voices, promotion of cultural understanding
Hate Speech Leveraging AI Tools	AI chatbots or deepfakes spreading hate	Automated and seemingly authentic hate speech, difficulty in identifying sources	Development of advanced detection tools, cooperation with AI experts
Hate Speech in Virtual Reality	Challenges in moderating VR experiences	Increased emotional impact, immersive exposure to hate	Development of VR-specific moderation tools, user education on VR

Platforms		speech	safety
Rise of "Hate Speech as a Service" Platforms	Platforms catering to hate groups with minimal moderation	Safe havens for hate speech, planning of real-world violence	International cooperation, law enforcement interventions
Hate Speech Disguised as Social Commentary	Hateful messages masked as critiques of political correctness	Difficulty in distinguishing legitimate criticism from hate	Contextual analysis, promoting constructive dialogue
Hate Speech in Fictional Content	Hate in video games, movies, fan fiction	Normalization of hate speech, negative influence on younger audiences	Content guidelines for creators, parental controls
Spread of Misinformation Fueling Hate Speech	Hate speech campaigns manipulating public opinion	Distorted perceptions, real-world consequences	Fact-checking initiatives, public education on media literacy
Hate Speech in International Sporting Events	Nationalistic hate targeting athletes	Hostile environment, increased tensions between fans	Promotion of sportsmanship, strict moderation during events
Rise of "Hate Speech Influencers"	Social media personalities spreading hate	Mainstreaming of hate speech, increased influence of harmful ideologies	Accountability for influencers, promotion of positive role models
Fabricated Stories and Conspiracy Theories	Misinformation portraying refugees as threats	Manipulation of public opinion, increase in xenophobia	Fact-checking, promotion of accurate information about refugees
Exploitation of Cultural Differences	Hate speech targeting cultural practices and traditions	Marginalization of specific cultural groups, increase in cultural tensions	Promotion of cultural understanding, swift action against hate speech targeting cultural differences
Normalization of Hate Speech in Comedy	Hateful jokes in stand-up and online content	Desensitization to hate speech, spread of harmful stereotypes	Promotion of inclusive humor, education on the impact of harmful jokes
Hate Speech Against Children and Teens	Bullying and harassment of young people online	Severe emotional distress, potential for self-harm or suicide	Education on online safety, robust moderation policies for content targeting young users

Hate Speech During Crisis Situations	Exploitation of pandemics or natural disasters for spreading hate	Increased fear and anxiety, targeting of specific groups	Timely and accurate information dissemination, countering misinformation
Hate Speech Targeting Healthcare Workers	Harassment and misinformation during health crises	Increased stress and burnout among healthcare workers, hindered public health efforts	Support for healthcare workers, countering misinformation campaigns
Rise of Hate Speech Chatbots	AI-powered bots spreading hate speech	Automated and wide-reaching dissemination of hate speech	Development of AI ethics guidelines, improved detection and moderation of chatbot interactions
Hate Speech Targeting Marginalized Communities During Elections	Hate speech campaigns targeting specific groups to influence elections	Disenfranchisement, increased political polarization	Election monitoring, strict enforcement of hate speech laws during election periods
Hate Speech in User-Generated Memes	Memes spreading hate and misinformation	Rapid and wide dissemination of hate speech, difficulty in moderation	Monitoring and moderation of meme content, public education on meme culture

Table 2. *Comprehensive Analysis of Hate Speech: Examples, Impacts, and Responses*

The table paints a concerning picture of the multifaceted nature of hate speech on social media. While platforms offer a space for free expression, they also enable the spread of hateful content targeting a wide range of groups. This hate speech can have severe consequences, causing emotional distress, inciting violence, and marginalizing entire communities.

The tactics used to spread hate speech are diverse and constantly evolving. From coordinated harassment campaigns to weaponized reporting systems and hate speech disguised as humor or academic discourse, bad actors exploit the anonymity and reach of social media to spread their messages. Algorithmic amplification further intensifies the problem, creating echo chambers and radicalizing users.

The fight against online hate speech requires a multi-pronged approach. Social media platforms need to develop stricter content moderation policies and implement effective detection tools, including those that can identify hate speech disguised as symbols or memes. Collaboration with NGOs and experts on regional hate speech nuances is crucial. Additionally, promoting media literacy and educating users on the impact of hate speech can help counter its normalization.

The table also highlights the tension between free speech and protecting vulnerable groups. Defining hate speech clearly and establishing transparent guidelines is essential. However, content moderation needs to be balanced to avoid stifling legitimate criticism or dissent.

5. LESSONS LEARNED AND POLICY IMPLICATIONS

- **Need for comprehensive strategies:** A single solution won't work. Addressing online hate speech requires a combination of platform regulation, user education, legal intervention, and fostering counter-narratives.
- **International cooperation:** Hate speech transcends borders. Collaborative efforts with international organizations and cross-platform communication are vital to tackle coordinated hate campaigns and hate speech disguised as cultural practices.
- **Empowering bystanders:** The table emphasizes the psychological impact of witnessing hate speech. Encouraging bystanders to report hate speech and supporting victims are crucial steps.
- **Promoting media literacy:** Equipping users with skills to critically evaluate information online can help them distinguish between legitimate content and hate speech disguised as jokes, news, or social commentary.
- **Accountability for all:** Social media platforms, influencers, and even politicians need to be held accountable for spreading hate speech.
- **Protecting freedom of expression:** Clear frameworks defining hate speech and protecting legitimate criticism are essential to uphold freedom of expression while creating a safer online space for everyone.

CONCLUSION

By acknowledging the complexities of online hate speech and implementing these lessons, we can work towards a healthier online environment that fosters free expression without allowing hate to flourish.

In the face of the complex interplay between freedom of expression and hate speech on social media, our exploration reveals the urgent need for multifaceted strategies to foster a safer digital sphere. From understanding the foundational principles of freedom of expression to recognizing the insidious impacts of hate speech, our analysis underscores the critical importance of proactive intervention.

Drawing upon real-life case studies and global statistical data, we have highlighted the pervasive nature of hate speech and its detrimental effects on individuals and society at large. However, amidst these challenges, there is room for optimism and actionable solutions.

Through collaboration between social media platforms, policymakers, civil society, and users, we can develop comprehensive approaches to mitigate the harms of hate speech while upholding the principles of free expression. From refining content moderation policies to enhancing user education and fostering cross-border cooperation, there are various avenues for progress.

Importantly, our journey has emphasized the need for nuanced approaches that recognize the contextual complexities of hate speech and its ever-evolving manifestations. By promoting media literacy, empowering bystanders, and holding all stakeholders accountable, we can cultivate a digital environment where inclusivity, respect, and dialogue thrive.

In conclusion, the pursuit of a safer digital sphere demands collective action and unwavering commitment to fundamental rights. By leveraging the lessons learned and policy implications outlined in this article, we can strive towards a future where social media serves as a catalyst for positive change, rather than a breeding ground for hatred and division.

BIBLIOGRAPHY

[1.] Abebaw, Z., Rauber, A., & Atnafu, S. (2022). Design and Implementation of a Multichannel Convolutional Neural Network for Hate Speech Detection in Social Networks. *Revue d'Intelligence Artificielle*, *36*(2), 175–183.

[2.] Alforova, T. M., Koba, M. M., Lehka, O. V., & Kuchuk, A. M. (2022). Right to Freedom of Expression v. Reputation Protection (Based on ECtHR Practice Materials). *The Age of Human Rights Journal*, *18*, 311–330.

[3.] Ayele, A. A., Yimam, S. M., Belay, T. D., Asfaw, T., & Biemann, C. (2023). Exploring Amharic Hate Speech Data Collection and Classification Approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing* (pp. 49–59). Varna, Bulgaria: INCOMA Ltd., Shoumen.

[4.] Bonotti, M., & Seglow, J. (2021). Freedom of Expression. *Philosophy Compass*, *16*(7), 1-13.

[5.] Council of Europe Committee of Ministers. (1997). *Recommendation No. R (97) 20 of the Committee of Ministers to member states on "hate speech"*. Retrieved from https://www.coe.int/en/web/freedom-expression/committee-of-ministers-adopted-texts/-/asset_publisher/aDXmrol0vvsU/content/recommendation-no-r-97-20-of-the-committee-of-ministers-to-member-states-on-hate-speech-.

[6.] Clarke, C., Hall, M., Mittal, G., Yu, Y., Sajeev, S., Mars, J., & Chen, M. (2023). Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 364–376. Toronto, Canada: Association for Computational Linguistics.

[7.] Geiger, C., & Izyumenko, E. (2023). Designing a Freedom of Expression-Compliant Framework for Moral Rights in the EU: Challenges and Proposals. In Y. Gendreau (Ed.), *Research Handbook on Intellectual Property and Moral Rights* (pp. 292–314). Cheltenham, UK/Northampton, MA: Edward Elgar.

[8.] Google. (n.d.). *Policies for Content Posted by Users on Search*. Retrieved from https://www.google.com/intl/en-US/search/policies/usercontent/.

[9.] Hulin, A., & Brant, J. (2023). *Social Media 4 Peace: Local lessons for global practices*. United Nations Educational, Scientific and Cultural Organization (UNESCO).

[10.] Instagram. (n.d.). *Help center* [Instagram profile]. Retrieved from https://help.instagram.com/581066165581870/?helpref=hc_fnav.

[11.] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). *HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection*. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 14867-14875.

[12.] Meta. (n.d.). *Community standards* [Facebook page]. Retrieved from https://transparency.meta.com/fr-fr/policies/community-standards.

[13.] Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*, 16(2), e1648.

[14.] Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and Psychological Effects of Hateful Speech in Online College Communities. In *Proceedings of the 11th ACM Conference on Web Science*.

[15.] Sazzed, S. (2023). Discourse Mode Categorization of Bengali Social Media Health Text. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 52–57). Toronto, Canada: Association for Computational Linguistics.

[16.] TikTok. (n.d.). *Safety Center* [TikTok website]. Retrieved from https://www.tiktok.com/safety/en/countering-hate?sc_version=2024.

[17.] United Nations General Assembly. (1966). *International Covenant on Civil and Political Rights*. Resolution 2200A (XXI) of 16 December 1966. Retrieved from https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

[18.] United Nations Office on Genocide Prevention and the Responsibility to Protect. (2021). *United Nations Strategy and Plan of Action on Hate Speech*. Retrieved from https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml.

[19.] X. (n.d.). Policy on hateful conduct [X Help Center]. Retrieved from https://help.x.com/en/rules-and-policies/hateful-conduct-policy.

[20.] Zhang, Z., & Luo, L. (2019). *Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter*. *Semantic Web*, *10*(5), 925-945.

