Current Research in Law & Practice

Journal Homepage: www. brilliance-pub.com/crlp

Exploring the Impact of Content Moderation Policies on Freedom of Expression: A Case Study of Twitter in the United States from 2018 to Present

Zhian Qian[1]

[1]University of Arizona, United States

Correspondence: Zhian Qian, University of Arizona, United States.

ABSTRACT

This research explores the intricate dynamics of content moderation policies and their impact on freedom of expression on Twitter from 2018 to the present. The study encompasses the evolution of social media, Twitter's policy framework, legal dimensions in the United States, and a detailed case study. User perspectives and the nuanced relationship between policy enforcement and freedom of expression are critically analyzed. The findings contribute to a comprehensive understanding of how content moderation practices shape online discourse.

KEYWORDS

social media; Twitter; content moderation; freedom of expression; online discourse; legal framework; user perspectives

1. Background and Context

1.1 Evolution of Social Media

In the last two decades, social media has undergone a rapid and transformative evolution, redefining the landscape of modern communication. Social media platforms have become integral tools for individuals, organizations, and even governments to connect, share information, and engage with global audiences. This transformation has not only revolutionized the way people communicate but has also had profound implications for societal structures, information dissemination, and the exercise of fundamental rights.

1.1.1 Overview of the Transformative Role of Social Media

The transformative role of social media is evident in its ability to facilitate instant communication and information sharing on a global scale. Individuals can now connect with friends, family, and colleagues across borders, fostering a sense of global interconnectedness. Social media has become a primary source of news and information, influencing public discourse and shaping opinions on a wide range of issues.

Beyond personal connections and news dissemination, social media has empowered individuals to express their thoughts, opinions, and creativity freely. Platforms like Twitter have become virtual public squares where users can voice their perspectives on political, social, and cultural matters. This democratization of expression has contributed to a more inclusive and diverse public discourse.

1.1.2 Emergence of Concerns Regarding the Regulation of Content on These Platforms

However, alongside the positive transformations, concerns have emerged regarding the regulation of content on social media platforms. The open and unrestricted nature of these platforms has led to the rapid dissemination of information, including misinformation and harmful content. Issues such as hate speech, cyberbullying, and the spread of false information have prompted a reevaluation of the boundaries between free expression and responsible content management.

As social media platforms grapple with the challenge of maintaining a balance between facilitating open dialogue and preventing harm, content moderation policies have become central to the platform's functioning. Particularly, the policies set by platforms like Twitter play a pivotal role in shaping the online environment, raising questions about the impact of such policies on the fundamental right to freedom of expression.

In the following sections, we will delve into the specific content moderation policies employed by Twitter, examine the legal dimensions governing online expression in the United States, and conduct a case study to unravel the real-world implications of content moderation on Twitter from 2018 to the present.

2. Twitter's Content Moderation Landscape

2.1 Twitter's Policy Framework

Twitter, as a prominent social media platform, operates within a comprehensive policy framework designed to govern user behavior and content dissemination. The content moderation policies of Twitter are instrumental in maintaining a semblance of order, safety, and responsible expression within its virtual community.

2.1.1 In-Depth Exploration of Twitter's Content Moderation Policies

Twitter's content moderation philosophy is deeply ingrained in its mission to create an open and inclusive platform while mitigating potential harm. The platform strives to strike a delicate balance between preserving freedom of expression and protecting users from abusive, harmful, or misleading content.

Twitter's commitment to fostering a healthy online environment is manifested through its policies that address various facets of user interactions:

Hate Speech and Harassment: Twitter has instituted stringent measures to combat hate speech and harassment, defining unacceptable behavior within the platform's policies. This encompasses targeted attacks, discriminatory language, and malicious threats. The enforcement of these policies aims to create a space where users can express themselves without fear of intimidation or discrimination.

False Information: Acknowledging the importance of curbing the spread of false information, Twitter's approach involves comprehensive policies targeting misinformation, disinformation, and the promotion of false narratives. Fact-checking initiatives play a pivotal role in verifying and contextualizing information shared on the platform, contributing to a more informed and responsible digital discourse.

Graphic and Violent Content: Twitter maintains a clear stance on graphic and violent content, acknowledging the potential harm such material can inflict on users. The platform's guidelines meticulously outline the types of content considered sensitive and potentially harmful, providing a framework for moderation efforts that prioritize user safety while respecting freedom of expression.

Privacy and Personal Information: In safeguarding user privacy, Twitter has implemented policies restricting the unauthorized sharing of personal information. This includes stringent guidelines against doxxing and the dissemination of private details without user consent. The platform endeavors to create an atmosphere where individuals can engage without compromising their personal security.

Twitter employs a combination of automated systems and human moderation to enforce its content moderation policies:

Automated Systems: The role of automated systems, including algorithms and artificial intelligence, is integral to identifying and flagging content that violates Twitter's policies. These systems operate at scale, assisting in the swift identification and removal of content that contravenes established guidelines.

Human Moderation: Human moderators play a crucial role in the content review process. Trained to interpret context and nuance, they navigate the complexities of user interactions, making judgment calls in situations where automated systems may fall short. The involvement of human moderators underscores Twitter's commitment to nuanced and context-aware content moderation.

Twitter's content moderation efforts are not without challenges and criticisms:

Algorithmic Bias: Concerns about algorithmic bias have been raised, questioning the fairness and neutrality of automated systems. Twitter acknowledges these concerns and strives to address biases, ensuring that content moderation decisions are equitable across diverse user demographics.

Transparency and Accountability: Ensuring transparency in content moderation processes and mechanisms for accountability remains an ongoing effort. Twitter faces the challenge of balancing transparency with user privacy while continually refining its processes to enhance public understanding and trust.

This in-depth exploration provides a holistic understanding of Twitter's content moderation policies, emphasizing the platform's commitment to fostering a safe and inclusive digital space. The subsequent sections will delve into the legal dimensions of online expression in the United States and scrutinize the real-world implications of Twitter's content moderation efforts through a case study spanning from 2018 to the present.

2.1.2 Identification of Key Principles and Guidelines Shaping User Behavior

The principles and guidelines embedded in Twitter's content moderation policies are instrumental in shaping user behavior and fostering a digital environment that aligns with the platform's commitment to openness, inclusivity, and safety.

Transparency and Clarity: Twitter prioritizes transparency by clearly articulating its principles and guidelines. Users are provided with accessible and comprehensible information outlining the acceptable norms for engagement. This transparency sets the foundation for a community where users understand the rules governing their interactions.

Non-Discrimination and Inclusivity: A core principle of Twitter's content moderation is the promotion of non-discrimination and inclusivity. The guidelines explicitly prohibit content that promotes discrimination based on factors such as race, ethnicity, gender, sexual orientation, and religion. By doing so, Twitter aims to create an inclusive digital space that respects and values diversity.

Freedom of Expression with Responsibility: While upholding the value of freedom of expression, Twitter emphasizes responsible expression. Users are encouraged to share diverse viewpoints, but the guidelines draw a clear line against content that incites violence, promotes hate speech, or engages in harassment. This approach seeks to strike a balance between fostering open dialogue and preventing harm.

Protection Against Harmful Conduct: Key guidelines revolve around protecting users from harmful conduct. Policies address various forms of harm, including targeted harassment, threats of violence, and the dissemination of graphic or violent content. By establishing clear boundaries, Twitter seeks to ensure a safe and secure digital environment for all users.

Promotion of Authenticity and Integrity: To maintain the authenticity and integrity of interactions, Twitter's guidelines discourage the creation and dissemination of spam, fake accounts, and artificially amplified content. By curbing deceptive practices, the platform aims to enhance the credibility of user interactions and the information shared on the platform.

Privacy and Respect for Personal Information: Respecting user privacy is a fundamental principle in Twitter's content moderation approach. Guidelines explicitly forbid the sharing of private and confidential information without consent. This commitment to protecting user privacy contributes to building trust among the platform's user base.

Community Collaboration and Reporting Mechanisms: Twitter actively involves the community in shaping its content moderation policies. Users are encouraged to report violations, and the platform provides reporting mechanisms to address potential breaches promptly. This collaborative approach fosters a sense of shared responsibility among users.

By identifying and adhering to these key principles and guidelines, Twitter seeks to shape user behavior in a manner that upholds a digital space characterized by respectful discourse, safety, and the responsible exercise of freedom of expression. The subsequent sections will explore the legal dimensions of online expression in the United States and delve into the real-world implications of Twitter's content moderation efforts through a case study spanning from 2018 to the present.

3. Legal Dimensions of Online Expression in the United States

3.1 Legal Framework Analysis

The legal framework governing online expression in the United States plays a pivotal role in shaping the boundaries of freedom of expression on platforms like Twitter. A comprehensive analysis of relevant U.S. laws provides insight into the legal landscape that Twitter navigates to ensure compliance while safeguarding the right to free speech.

3.1.1 Examination of Relevant U.S. Laws Influencing Freedom of Expression Online

First Amendment Protections: The cornerstone of freedom of expression in the United States is the First Amendment of the Constitution. This constitutional provision protects individuals' rights to freedom of speech, assembly, and the petition of grievances. Understanding the nuances of how the First Amendment applies to online expression is crucial in evaluating the legal parameters that platforms like Twitter must consider.

Section 230 of the Communications Decency Act: Section 230, a crucial component of the Communications Decency Act, provides immunity to online platforms from liability for content posted by users. This legal provision shields platforms like Twitter from being held responsible for the vast array of user-generated content while empowering them to moderate and remove content deemed objectionable. An examination of Section 230 sheds light on the balance between platform immunity and content moderation authority.

Hate Speech and Incitement to Violence: While the First Amendment protects free speech, certain forms of expression are subject to limitations, particularly when they involve hate speech or incitement to violence. Analyzing relevant U.S. laws addressing hate speech and the incitement of violence elucidates the legal boundaries within which platforms like Twitter must operate to ensure a safe online environment.

Privacy Laws and Personal Information Protection: U.S. privacy laws, such as the Electronic Communications Privacy Act and the Children's Online Privacy Protection Act, impact online expression by regulating the collection and dissemination of personal information. An examination of these privacy laws provides insights into the legal considerations surrounding the protection of user data and the prevention of unauthorized disclosures.

Digital Millennium Copyright Act (DMCA): The DMCA addresses intellectual property concerns related to online content. Understanding how the DMCA operates in the context of user-generated content on platforms like Twitter is essential to navigating legal challenges related to copyright infringement and intellectual property rights.

Recent Legislative Developments: Reviewing recent legislative developments, including proposed changes to Section 230 and new initiatives addressing online misinformation, provides a forward-looking perspective on potential shifts in the legal landscape. This analysis considers how evolving laws may impact Twitter's content moderation practices.

By conducting a thorough examination of these relevant U.S. laws, this section aims to illuminate the legal context within which Twitter operates and exercises its content moderation policies. The subsequent sections will delve into a case study of Twitter's content moderation from 2018 to the present, providing a real-world application of the legal dimensions explored.

3.1.2 Changes and Updates in Legislation Since 2018

Since 2018, the legal landscape governing online expression in the United States has witnessed notable changes and updates. These modifications have implications for platforms like Twitter, influencing how they navigate content moderation while upholding the principles of freedom of expression.

Amendments to Section 230: One of the most significant developments has been the discourse and proposed amendments surrounding Section 230 of the Communications Decency Act. Policymakers have debated potential changes to this foundational provision, aiming to refine the balance between platform immunity and accountability for user-generated content. Understanding the evolving discussions and potential amendments to Section 230 is crucial for assessing Twitter's legal obligations and responsibilities.

State-Level Legislation on Online Expression: Several U.S. states have introduced or passed legislation addressing online expression and content moderation. These state-level initiatives vary in scope and intent, with some seeking to enhance transparency in content moderation decisions and others aiming to address issues like political bias. Examining these state-level legislative changes provides insights into the diverse approaches being taken to regulate online expression.

Focus on Privacy Legislation: In the wake of increased concerns about user data privacy, several states have enacted or proposed privacy legislation that could impact online platforms. Evaluating these legislative changes helps in understanding the evolving regulatory landscape around the protection of user information and the implications for platforms like Twitter.

Initiatives Targeting Online Misinformation: Given the rising concerns about online misinformation and disinformation, policymakers have explored initiatives to address these issues. Proposals for legislation focusing on combating misinformation, particularly in the context of elections and public health crises, reflect an evolving legal response to the challenges posed by false information online.

Intellectual Property and Copyright Reforms: Changes in intellectual property and copyright laws can influence how platforms handle user-generated content. Examining any legislative updates in this domain provides insights into the legal considerations for platforms like Twitter in addressing copyright infringement and protecting intellectual property rights.

International Data Protection Influences: As global discussions on data protection intensify, changes in international laws and regulations may indirectly impact U.S.-based platforms. Evaluating developments in international data protection laws helps in understanding the broader legal context within which platforms operate.

By analyzing these changes and updates in legislation since 2018, this section aims to provide a nuanced understanding of the evolving legal constraints and opportunities shaping Twitter's approach to content moderation. The subsequent sections will delve into a case study of Twitter's content moderation from 2018 to the present, offering practical insights into how legal dimensions manifest in real-world scenarios.

4. Case Study: Content Moderation on Twitter (2018-Present)

4.1 Examining Policy Implementation

4.1.1 Investigation into How Twitter's Content Moderation Policies Were Applied in Practice

Overview: This section delves into a detailed examination of how Twitter's content moderation policies have been implemented in practice from 2018 to the present. The case study seeks to unravel the real-world application of Twitter's guidelines, shedding light on the challenges, successes, and user experiences within the dynamic landscape of online expression.

Implementation of Hate Speech and Harassment Policies: Investigating the application of Twitter's hate speech and harassment policies involves analyzing specific instances where these policies were invoked. This includes identifying patterns of enforcement, evaluating the effectiveness of automated systems and human moderation, and assessing Twitter's responsiveness to evolving forms of online abuse.

Handling False Information and Misinformation: A critical aspect of this case study is an exploration of how Twitter has dealt with false information and misinformation. Examining notable instances, fact-checking initiatives, and the platform's response to the dissemination of inaccurate information provides insights into the practical challenges of mitigating misinformation within the Twitter ecosystem.

Management of Graphic and Violent Content: The application of policies related to graphic and violent content requires an examination of Twitter's efforts to strike a balance between allowing freedom of expression and safeguarding users from potentially harmful material. This includes assessing the effectiveness of content warnings, automated content detection mechanisms, and user feedback in shaping policy implementation.

Privacy Protections and Personal Information Handling: An investigation into how Twitter has implemented privacy protections involves evaluating the platform's response to unauthorized sharing of personal information, doxxing incidents, and breaches of user privacy. Understanding the measures taken to uphold user confidentiality and respond to privacy-related concerns is integral to assessing Twitter's commitment to user protection.

Effectiveness of Automated Systems and Human Moderation: An analysis of the interplay between automated systems and human moderation is essential. This includes assessing the strengths and limitations of automated content detection algorithms, the role of human moderators in contextual decision-making, and the challenges faced in maintaining a nuanced approach to policy enforcement.

User Reactions and Impact on Digital Discourse: Examining user reactions to content moderation decisions provides valuable insights into the impact of policies on digital discourse. This involves assessing how users perceive and respond to moderation actions, understanding potential tensions between free expression and policy enforcement, and exploring the broader implications for the Twitter community.

Challenges and Iterative Changes: Identifying challenges faced by Twitter in implementing content moderation policies and understanding how the platform has iteratively adapted its policies based on user feedback and emerging issues is crucial. This involves exploring policy updates, clarifications, and adjustments made in response to the evolving nature of online expression.

Through this investigation, the case study aims to provide a nuanced understanding of how Twitter's content moderation policies have been translated into real-world practices, highlighting the complexities and dynamics of managing online expression on a global platform. The subsequent sections will further analyze the impacts on freedom of expression and draw conclusions based on the findings.

4.1.2 Uncovering Notable Instances of Policy Enforcement and User Reactions

This section delves into specific instances where Twitter's content moderation policies were enforced and examines the corresponding user reactions, providing a nuanced perspective on the real-world impact of these policies.

Hate Speech and Harassment Policy Enforcement: Identifying notable instances where Twitter's hate speech and harassment policies were enforced allows for a closer examination of the platform's response to specific cases. This includes instances of targeted attacks, discriminatory language, or malicious threats. The section explores the effectiveness of enforcement measures and analyzes user reactions to gauge community sentiment and perceptions of the policies in action.

Handling False Information and Misinformation: Examining instances where Twitter addressed false information and misinformation sheds light on the challenges faced by the platform in countering misleading content. Notable cases, fact-checking initiatives, and responses to viral misinformation provide insights into the practical considerations and user reactions. Understanding how Twitter navigated these situations contributes to a broader understanding of its approach to information integrity.

Management of Graphic and Violent Content: Uncovering notable instances of policy enforcement related to graphic and violent content involves analyzing how Twitter responded to specific cases of potentially harmful material. This examination includes the application of content warnings, automated content detection mechanisms, and user reactions to assess the platform's effectiveness in maintaining a balance between freedom of expression and user safety.

Privacy Protections and Personal Information Handling: Investigating instances where Twitter addressed privacy concerns and unauthorized sharing of personal information provides insights into the platform's commitment to user confidentiality. Notable cases of doxxing or privacy breaches are examined, along with user reactions, to understand how Twitter navigated these challenges and maintained a secure online environment.

User Reactions to Policy Enforcement: Understanding user reactions to content moderation decisions is crucial for gauging the impact of policies on the Twitter community. Analyzing how users perceive and respond to enforcement actions, including their feedback on the platform, provides valuable insights into the broader implications of policy decisions on digital discourse and user satisfaction.

By uncovering specific instances of policy enforcement and user reactions, this section aims to provide a granular understanding of the practical outcomes and challenges associated with Twitter's content moderation efforts. The subsequent sections will further analyze the collective impact on freedom of expression and draw comprehensive conclusions based on the findings.

5. Impacts on Freedom of Expression: A Critical Analysis

5.1 User Perspectives

5.1.1 Compilation and Analysis of User Experiences in the Context of Content Moderation

Compilation of User Experiences: This section aims to compile a diverse range of user experiences related to Twitter's content moderation policies. Users from various backgrounds and perspectives will be sampled to provide a comprehensive understanding of how these policies impact different segments of the Twitter community. The compilation will encompass instances of content removal, account suspension, or any other interactions with the moderation system.

Analysis of User Experiences: Each user experience will be analyzed in-depth to discern common themes, challenges, and patterns. This analysis will explore the following key aspects:

1) Perceived Impact on Freedom of Expression: Understanding how users perceive the impact of content moderation on their freedom of expression. This involves assessing whether users feel empowered to express their views or if they perceive limitations due to the moderation policies.
2) Effectiveness of Policy Communication: Evaluating how well Twitter communicates its content moderation policies to users. This includes analyzing user awareness of the rules, clarity in policy communication, and the impact of understanding or misunderstanding the guidelines.

3) Community Dynamics: Examining the impact of content moderation on the dynamics within the Twitter community. This involves assessing changes in user interactions, the emergence of self-censorship, or alterations in the tone and nature of discussions.
4) User Satisfaction with Resolution Processes: Investigating users' satisfaction with how Twitter handles appeals, provides explanations for moderation actions, and addresses user concerns. Understanding the resolution process is crucial in gauging the platform's responsiveness to user feedback.
5) Balancing Freedom and Safety: Analyzing user perspectives on the delicate balance between freedom of expression and the need for a safe and respectful online environment. This involves assessing whether users believe the content moderation policies strike an appropriate balance.

Identifying Patterns and Trends: Through the compilation and analysis of user experiences, patterns and trends will be identified. This includes common challenges faced by users, recurring themes in user feedback, and any evolving dynamics in user interactions on the platform. Identifying patterns provides valuable insights into the collective impact of content moderation on freedom of expression within the Twitter community.

Informing Policy Recommendations: The critical analysis of user experiences will contribute to informing potential policy recommendations. By understanding user perspectives, the aim is to identify areas where content moderation policies can be improved to better align with user expectations, enhance transparency, and mitigate unintended consequences on freedom of expression.

Through this comprehensive examination, the section seeks to offer a nuanced understanding of how Twitter's content moderation practices influence the freedom of expression from the users' standpoint. The subsequent section will draw conclusions based on the critical analysis and provide recommendations for enhancing the balance between content moderation and freedom of expression on the platform.

5.1.2 Understanding the Nuanced Relationship between Policy Enforcement and Freedom of Expression

The intricate interplay between Twitter's content moderation policies and the overarching concept of freedom of expression necessitates a nuanced examination. Striking a delicate balance between policy enforcement and the preservation of freedom of expression is a complex endeavor for the platform.

Analyzing this relationship involves a consideration of how Twitter manages the inherent tension between maintaining a respectful online environment and allowing users the freedom to express diverse perspectives. The platform's commitment to striking this balance is evident in its continuous efforts to refine content moderation policies.

One crucial aspect of this nuanced relationship is the impact on diverse voices. By exploring whether policy enforcement disproportionately affects certain groups or individuals, we can discern whether Twitter's content moderation inadvertently introduces barriers to the participation of marginalized communities.

Another facet is the potential chilling effects and self-censorship induced by policy enforcement. Examining whether users alter their behavior or withhold certain expressions due to apprehensions about moderation actions reveals the subtle ways in which policies may influence user interactions.

Understanding the contextual considerations in policy enforcement is vital. Twitter grapples with the challenge of interpreting content in various contexts, and the analysis delves into how the platform navigates this complexity to avoid unintended restrictions on legitimate expression.

Assessing user empowerment and agency within the framework of content moderation policies sheds light on whether users feel encouraged to express themselves or perceive limitations on their freedom of expression. It offers insights into the user experience and how policies impact their sense of participation.

The iterative adaptations of content moderation policies play a pivotal role in shaping the nuanced relationship. By exploring how Twitter responds to user feedback and emerging challenges, we gain insights into the platform's commitment to refining the delicate balance between policy enforcement and freedom of expression.

Considering the global and cultural sensitivities inherent in Twitter's diverse user base, the analysis explores how the platform addresses these nuances. It examines whether policies are adapted to respect diverse norms and whether the platform acknowledges and accommodates cultural differences in its moderation practices.

Furthermore, evaluating the level of user education and awareness regarding content moderation policies enhances our understanding. Examining whether users are adequately informed about the rules and consequences provides insights into how well users comprehend the parameters within which they can express themselves on the platform.

In essence, unraveling the nuanced relationship between policy enforcement and freedom of expression on Twitter requires a comprehensive exploration of these multifaceted dimensions. The findings from this analysis contribute to a holistic understanding of how content moderation practices influence the diverse and open discourse that defines the platform.

6. Conclusion

The exploration of the impact of content moderation policies on freedom of expression on Twitter from 2018 to the present reveals a complex and evolving landscape. This comprehensive analysis has delved into the evolution of social media, Twitter's content moderation framework, the legal dimensions influencing online expression in the United States, and a detailed case study of policy implementation. The critical analysis of user experiences and the nuanced relationship between policy enforcement and freedom of expression provide valuable insights into the dynamics of online discourse.

The evolution of social media, particularly Twitter, has transformed modern communication, making it a powerful platform for expression, activism, and information dissemination. However, concerns regarding the regulation of content on these platforms have become increasingly prominent, necessitating a nuanced approach to balancing freedom of expression and user safety.

Twitter's content moderation landscape is intricate, with policies addressing hate speech, misinformation, graphic content, and privacy considerations. The implementation of these policies involves a delicate interplay between automated systems and human moderation, with challenges such as algorithmic bias and the need for transparency and accountability.

The legal analysis highlights the foundational role of the First Amendment, the significance of Section 230, and the evolving legislative landscape. Changes and updates in legislation since 2018, including proposed amendments to Section 230 and state-level initiatives, contribute to the dynamic legal framework shaping Twitter's content moderation practices.

The case study of Twitter's content moderation from 2018 to the present unveils the practical application of policies. Examining instances of policy enforcement, user reactions, and the iterative changes made by the platform provides a real-world perspective on the challenges and successes of content moderation.

The critical analysis of user perspectives illuminates the impact on freedom of expression. Compiling and analyzing user experiences reveal patterns and trends, informing potential policy recommendations. Understanding the nuanced relationship between policy enforcement and freedom of expression involves assessing the balance, impact on diverse voices, potential chilling effects, contextual considerations, and user empowerment.

In conclusion, Twitter's content moderation journey is a dynamic and iterative process, shaped by legal frameworks, user experiences, and ongoing societal dialogues. The platform's commitment to fostering a space for open discourse while addressing challenges reflects the evolving nature of online expression. As Twitter continues to navigate the complexities of content moderation, the findings from this analysis contribute to a nuanced understanding of the platform's role in shaping the digital public sphere.

7. References

[1] Sander, B., "Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation," Fordham Int'l LJ, 43, p. 939, 2019.

[2] Llansó, E., van Hoboken, J., Leerssen, P., & Harambam, J., "Content Moderation, and Freedom of Expression," Algorithms, 2020.

[3] Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E., "Online harassment and content moderation: The case of blocklists," ACM Transactions on Computer-Human Interaction (TOCHI), 25 (2), pp. 1-33, 2018.

[4] Tworek, H., Fathaigh, R. Ó., Bruggeman, L., & Tenove, C., "Dispute Resolution and Content Moderation: Fair, Accountable, Independent," Algorithms, 2020.

[5] Wu, Y., "Privacy, Free Speech and Content Moderation: A Literature Review and Constitutional Framework Analysis," Innovation in Science and Technology, 1 (4), pp. 30-39, 2022.