Mapping the Landscape of Content Moderation: A Bibliometric Perspective

Özlem Ozan[1]
Ali Rıza Sadıkzade[2]

Abstract

Rapid technological advancements have intensified user-content interactions, leading to complex regulation mechanisms such as A.I. filtering and user moderation. This study conducts a bibliometric analysis of 202 publications from 2016 to 2023, sourced from the Web of Science and Scopus databases, to explore contemporary topics in content moderation research. It identifies influential authors, institutions, countries, journals, funding agencies, keyword networks, and co-authorship patterns. The findings indicate that the Queensland University of Technology is the most influential institution, while the United States, England, and Australia are the most productive countries. The National Science Foundation and European Research Council are key funding bodies. New Media & Society, Social Media + Society, and Big Data & Society are leading journals in this field. Research primarily focuses on social media platforms like Facebook, Instagram, YouTube, and Twitter, with a thematic shift from transparency to hate speech and misinformation. Since 2016, there has been a steady increase in academic publications on content moderation, suggesting continued growth in this area across various disciplines.

Keywords: Content Moderation, Social Media, Social Media Regulations, Bibliometric Analysis, Bibliometric Data.

Öz

Hızla gelişen teknolojiler, kullanıcı ve içerik etkileşimlerini artırarak yapay zeka filtreleme ve kullanıcı moderasyonu gibi karmaşık düzenleme mekanizmalarının uygulanmasını gerekli kılmıştır. Bu çalışma, 2016-2023 yılları arasında Web of Science ve Scopus veritabanlarından elde edilen 202 yayının bibliyometrik analizini yaparak, içerik moderasyonu araştırmalarındaki güncel konuları incelemeyi hedeflemektedir. Araştırma, etkili yazarların, kurumların, ülkelerin, dergilerin, fon sağlayıcı kuruluşların, anahtar kelime ağlarının ve ortak yazarlık ilişkilerinin belirlenmesini amaçlamaktadır. Bulgulara

[1] Yaşar University, Communication Faculty, New Media and Communication Department
ozlem.ozan@yasar.edu.tr, https://orcid.org/0000-0002-4116-1551
[2] Yaşar University, PhD Candidate at Communication Programme
ali.riza.sadikzade@gmail.com, https://orcid.org/0000-0002-2824-2023

göre, Queensland Teknoloji Üniversitesi en etkili kurum, Amerika Birleşik Devletleri, İngiltere ve Avustralya ise en üretken ülkelerdir. National Science Foundation ve European Research Council, başlıca fon sağlayıcı kuruluşlar arasında yer almaktadır. New Media & Society, Social Media + Society ve Big Data & Society bu alandaki en etkili dergilerdir. Araştırmalar, genellikle Facebook, Instagram, YouTube ve Twitter gibi sosyal medya platformlarına odaklanmakta olup, araştırma temalarında şeffaflıktan nefret söylemi ve yanlış bilgilendirmeye doğru bir kayma görülmektedir. 2016 yılından bu yana akademik yayınlarda istikrarlı bir artış gözlenmekte, söz konusu ilginin farklı disiplinleri de kapsayan daha fazla sayıda araştırmaya yol açması beklenmektedir.

Anahtar Kelimeler: İçerik Moderasyonu, Sosyal Medya, Sosyal Medya Düzenlemeleri, Bibliyometrik Analiz, Bibliometric Veri.

1. Introduction: Internet, Social Media, and Content Moderation

The internet has become part of many individuals' daily lives. According to PEW research (Perrin & Atske, 2021), which focuses on Americans' internet usage, eight out of ten individuals use the internet daily, while three are almost always online. Its increasing salience in individuals' everyday experiences blurs the line between online and offline spheres.

Discussions about the implications of the internet have a wide range. For example, according to another PEW research (Wike et al., 2022), which focuses on the effects of social media on democracy, people accept the positive effects of social media, such as ease of access to information, increased social connectivity, exposure to diverse opinions, civilly approaching different identities and lifestyles, as well as the negative effects such as increased polarization and manipulative misinformation. In addition, there is a debate on whether social media platforms like YouTube create pipelines to extremist ideologies (Hosseinmardi et al., 2021; Tüfekçi, 2018). Such social media manipulations can influence the adoption of unreasonable decisions and exacerbate problems like real-life discrimination. They might also aggravate political conflicts, as observed in the Rohingya massacre (Amnesty International, 2022).

In addition to the issues raised organically, political actors are motivated to use the internet and social media —overtly or covertly— to further their political agenda. Political actors can manipulate voters through coordinated campaigns, which involve spreading disinformation or mass postings, as seen in the 2016 U.S. Presidential Election (Ferrara et al., 2020). Furthermore, besides the political domain, communications on the internet also might raise problems within the personal and interpersonal domains. Users and content creators are now facing issues like data privacy, copyright, internet scams, and cyberbullying.

Because of the issues discussed above, online individuals and communities need safeguarding. Safeguarding is relatively new from a corporate and legislative standpoint.

On the legislative side, one of the early regulations is Section 230 of the Communications Decency Act of the U.S. Congress, which was passed in 1996. Arguably, Section 230 restricted the liabilities of online platforms for the content shared in their domains. However, tech companies attempted to review any content on their platforms that could damage their users. Nevertheless, it can be argued that tech companies sought to keep this process at a minimal level (Angwin & Grassegger, 2017; Caplan, 2018).

Although it could be challenging to pinpoint an exact time for it, emerging discussions propose an increased authority of institutions. Digital Services Act prepared by the European Union, Germany's NetzDG law, and controversial laws such as H.B. 20 of Texas (Levy, 2022), A.B. 587 of California (State of California Department of Justice, n.d), Anti Fake-News Act of Malaysia (Caplan, 2018) and Disinformation law of Turkey (Özbudun, 2022) can be seen as legislative examples.

More examples include legal actions taken or proposed against TikTok (Maheshwari, 2023), famous lawsuits against online personalities such as Alex Jones (Audureau, 2022), and the deplatforming of political figures (Fung, 2021) (most famously Donald J. Trump.[1]).

A recent decision by Meta has restricted the visibility of political content. The changes applied by Meta did cease recommendations of 'political content' from the content creators that the users do not follow. However, users may opt out of these changes by adjusting the content settings in their profile (De Guzman, 2024).

Although Meta's application was received with broad criticism by civil rights groups, press institutions, and Meta's user base and for its timing and the vagueness in Meta's description of 'political content,' the move can be seen as a continuation of META's attempt to depoliticize its platforms. The ironic part is that Meta has claimed that its move to depoliticize its platforms was made in accordance with the wishes of their user base (Treisman, 2024).

In Meta's case, the attempts of depoliticization created a rather bizarre situation, for wide-ranging pages (including prestigious newspapers) have posted content instructing their followers on how to opt out of the changes applied by Meta, therefore undermining the legitimacy of such changes.

On the platforms' side, there is an increased need to safeguard the online sphere they provide, and evidence of this is the drastic increase in the content moderation market and its estimated growth. The process of 'content moderation[2]' can ban, hide, reduce, or promote content and content creators. The process would involve algorithms, direct and indirect user cooperation, and active moderation teams following a guideline (MSNBC, 2021; Veglis, 2014). Another point about moderation processes and policies is that they could shape the platform's identity and community.

---

[1] Donald Trump's Twitter account has been reinstated but remains mostly inactive.

[2] Although the literature uses different terms, such as content governance, platform regulation, or self-regulation, we will use "content moderation" in this paper.

A study by the New York University Stern Center for Business and Human Rights found that the number and proportion of policies violated can change depending on the platforms (Barrett, 2020). Moreover, the moderation process is influential in constructing the public sphere provided by the platform, thus molding their identities.

Furthermore, although websites or platforms perform content moderation, online service providers could demand their clients follow ethical guidelines and decline to serve those who violate them, thus getting involved in the moderation process (Byman, 2022).

Especially on larger platforms, the guidelines and the performance of the content moderation have raised questions about the process, such as the adequacy of human moderators who understand the language of the content market (Debre & Akram, 2021) or the bias regarding politics, race, gender or sexuality on the moderation process (Angwin & Grassegger, 2017).

In the case of big platforms, it is observed that the moderation methods could change depending on the size or the financial structure of the platform, meaning that platforms with a relatively small user base could work with smaller teams and cases, focusing on harsher judgments, whereas "industrial" platforms could work with large moderation teams and cases with relatively minor judgments, (Caplan, 2018; Liu et al., 2022). In addition, some platforms, such as Wikimedia and Reddit, could encourage voluntary moderators in their communities (Caplan, 2018).

The content moderation process can rely on public and non-public documents. While moderation teams can refer to public documents (such as the list of criminal groups), they may also rely on documentation and training not available for public viewing (West, 2018).

For instance, when a filming crew was allowed to record the training session of a content moderation team for a dating site, they could capture the extent of detail that such training may specify. The recorded training shows how the guidelines may define vulgarity and nudity through determined elements such as the camera's focus or the exposure ratio (Field of Vision, 2017).

The process of content moderation may have negative effects on users, who may feel frustrated and lose faith in the process. Users who face moderation rarely get a proper explanation (Suzor et al., 2019). In line with this, the experience of their content being moderated can create a perception of censorship for the users, creating a sense that they can be 'targeted' for their political views (Suzor et al., 2019; West, 2018).

Moreover, users who face content moderation tend to devise strategies to evade it; users may avoid using certain words or hashtags, employ self-censorship on their content, or devise strategies to signal their intended content, such as employing significant aesthetics and/or coded words (Gerrard, 2018).

Therefore, transparency and accountability are the foundational topics within the discussion surrounding content moderation; on the algorithm side, transparency may allow ill-intentioned parties to rig the algorithms (Katzenbach & Ulbricht, 2019), yet the lack of

transparency tends to block platforms' accountability and harm the relationship between platforms and their userbase.

Although certain gains are made by being more 'transparent,' such as x, there seems to be a certain aspect that could trick outsiders. As Suzor et al. (2019) put it, transparency can be employed to evade higher accountability.

Elon Musk's takeover of X (Twitter) is a valuable example of what Suzor et al. (2019) mentions. Musk's statements claim that Twitter's operations do not realize its 'free speech' potential, and he intends to create a public sphere that is 'maximally trusted' and 'broadly inclusive' (Sato, 2022). In Musk's control, certain controversial figures that were once banned from the platform were reinstated or were offered to be; these figures include ex-President Donald Trump, Ye[3] (artist formerly known as Kanye West) and Andrew Anglin (an infamous neo-nazi); in addition to these decisions, X under Musk's ownership has noticeably reduced the number of content moderators; Musk's decisions within X have not only received criticisms from former Twitter workers and civil rights groups but the social media platform started to lose large advertisers (CBS News, 2022; Klepper & O'Brien, 2022)

In a way, Elon Musk's takeover is a different example of the political utilization of transparency. Musk's dramatic decrease in content moderation is, in a way, done to implement a politicized notion of free speech; upon his acquisition, Musk shared a Twitter post claiming that 'comedy is now legal,' signifying his free speech approach. However, following his takeover, the accounts that criticized or mocked Elon Musk were suspended, placing Musk's 'free speech absolutism' under question (Tangalakis-Lippert, 2022).

It appears that Musk's policies have incentivized hate speech; the Federal Anti-Discrimination Agency of Germany has left the platform, citing the rise in hate speech under Musk's rule (D.W. News, 2023). Research conducted by the Center for Countering Digital Hate (2023) has found that X has failed to remove 99% of the hateful content reported by the researchers[4], whereas X has enabled the spread of hate speech and misinformation that stoked the flames of U.K. race riots, it was Elon Musk himself sharing misinformation; during the riots, Musk co-tweeted a fake headline from the leader of the far-right Britain First, claiming that the rioters will be detained in the Falkland Islands (Freedland, 2024).

Moreover, X has recently been banned in Brazil, similar to X's position during the U.K. riots, X's lack of action against far-right content floating in its platform, following Musk's noncompliance with the demands of Brazilian authorities to assign a legal representative, the Brazilian Supreme Court has banned the platform (Phillips, 2024).

2. Methodology

This section covers research methodology adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Page et al., 2021). This framework mainly includes the key elements of the research: rationale and aim, search

[3] Ye was banned from Twitter by Elon Musk for posting a swastika image.
[4] X has sued the Center for Countering Digital Hate because the group's violation of the site's terms has influenced advertisers to cease their ties with X, the court has dismissed X's claims (CBS News, 2024)

strategy (database selection, search terms, inclusion, and exclusion criteria), study selection (screening process, data extraction), data analysis, results, discussion, and conclusion.

2.1. Rationale and Aim of the Research

This research conducts bibliometric analysis to assess the literature structure on content moderation. As defined by Broadus (1987), bibliometric analysis involves the quantitative study of published or bibliographic units. Zupic and Čater (2015) characterize bibliometrics as a quantitative method for describing, evaluating, and monitoring research.

In the contemporary context, the enhancements in computer programs, such as VOSviewer, GEPHI, Bibexcel, and CiteSpace II, along with the establishment of reputable databases like Web of Science and Scopus, have rendered bibliometric analysis more fruitful (Cobo et al., 2011; Zupic and Čater 2015). Therefore, bibliometric analysis effectively addresses the trends and issues in the literature on content moderation, which is essential for fostering a positive, safe, and constructive online environment.

In this context, the research aims to answer the following questions:

- RQ1: What is the growth pattern of publications in content moderation? Which years have seen significant increases or decreases in publication outputs?
- RQ2: Which authors had the highest publication productivity and citations in content moderation?
- RQ3: What is the co-authorship network structure among authors in content moderation? Can we identify central authors, communities, or specific collaboration patterns?
- RQ4: Which journals hosted the most research on content moderation?
- RQ5: Which institutions and countries have contributed the most to the literature on content moderation?
- RQ6: Which articles on content moderation have received the highest number of citations?
- RQ7: What are the most frequently used keywords in articles on content moderation, and how have these keywords evolved? What does the co-occurrence network of keywords reveal about thematic clusters and relationships?
- RQ8: How have research topics related to content moderation evolved?

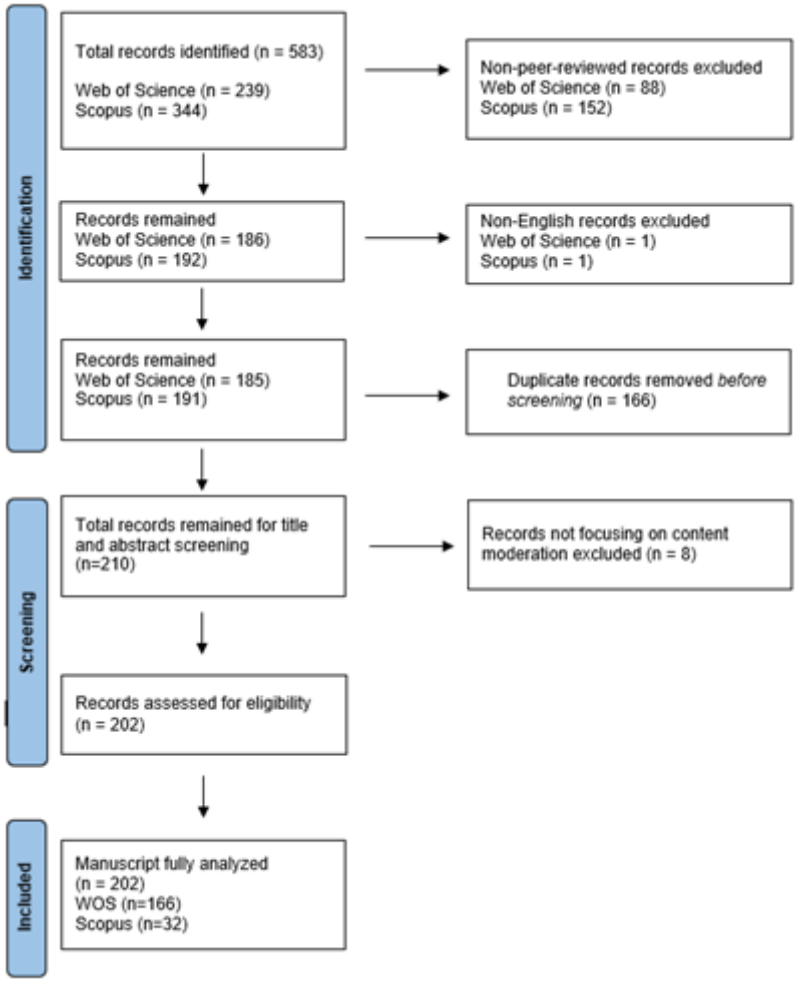
2.2. Search Strategy

Database Selection: We systematically searched for relevant articles in the Web of Science (WOS) and Scopus databases, comprehensively covering academic literature across various disciplines.

Search Terms: Choosing an effective search strategy was crucial for identifying relevant research on content moderation, given the diverse terminology and lack of standardized keywords within the field. We initially experimented with combining controlled vocabulary terms and keywords, but this approach included irrelevant studies.

We adopted an iterative approach to optimize results, ultimately selecting a strategy focused on "content moderation" within the Author Keywords field of both WOS and Scopus. This strategy allowed us to capture a broader range of research while minimizing false positives, as studies explicitly mentioning content moderation in their author keywords are more likely to be relevant to our analysis.

Figure 1: PRISMA Framework of the Study



Inclusion and Exclusion Criteria: We included the studies if they meet the following criteria: (1) Published in peer-reviewed journals, (2) Address content moderation as a primary focus or a significant component, (3) Available in English, and (4) Accessible as Full-text. Studies that did not meet these criteria were excluded.

2.3. Study Selection

Screening Process: The authors independently screened titles and abstracts to identify potentially eligible studies.

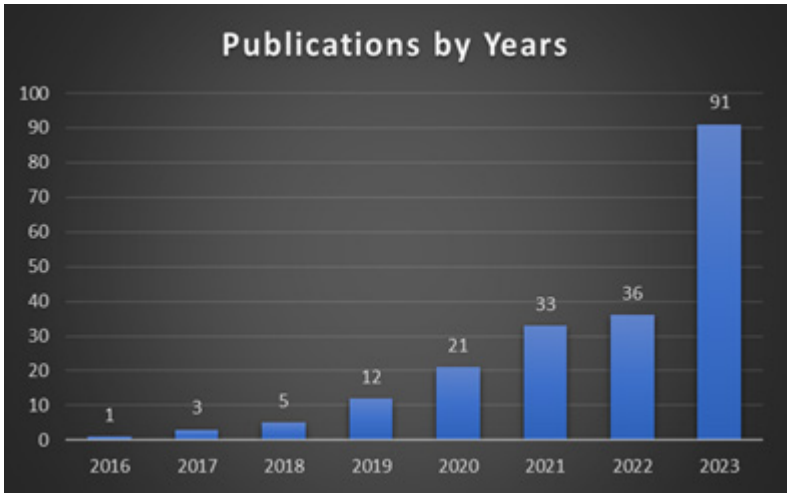
Data Extraction: The standardized metadata provided by WOS and Scopus databases for bibliometric analysis collects relevant information from each included study. Extracted data included Authors, Article Title, Journal Name, Author Keywords, Abstract, Author Address, Funding Text, Cited References, Cited Reference Count, Total Times Cited Count, Publisher, Publisher Address, International Standard Serial Number (ISSN), Year Published, Digital Object Identifier (DOI), Web of Science Categories, Research Areas, and other relevant details. Further data for RQ3 (such as h-index, country, and subject areas of journals) has been collected from the Scimago Database.

2.4. Data Analysis

We uploaded extracted data as RIS files to Zotero Reference Management Software to detect duplicate records of WOS and Scopus databases. After the identification of duplicates, we combined WOS and Scopus files manually with M.S. Excel and analyzed them with VOSviewer Bibliometric Analysis Software.

The bibliometric analysis uses publication metrics (e.g., average vs. total and single- vs. multi-authored publications), Citation metrics (e.g., average vs. total citations), and publication-citation metrics (e.g., h-index) for science mapping of the studied field. Science mapping includes the following methods to examine social networks based on contributors and knowledge clusters based on cited/citing publications and keywords (Lim and Kumar 2024): (1) Co-authorship analysis, (2) Co-citation analysis, (3) Bibliographic coupling analysis, (4) Co-occurrence of keywords analysis, (5) Citation analysis and (6) PageRank analysis. Furthermore, performative statistics relating to institutions (e.g., Universities and

Figure 2: Growth Pattern of Publications by Years



funding programs) can also be analyzed within bibliometrics (Benckendorff and Zehrer, 2013; Michael Hall, 2011).

In this context, we used the following analyses to overview and inspect the networks and connections among themes related to content moderation as well as classify the literature, highlighting conceptual structures that could produce insights through mapping:

a. Descriptive statistics based on metrics in RQ1, RQ2, RQ3, RQ4, R5, R6, R7, R8

b. Co-authorship analysis (authors and countries) in R3

Table 1: Top 10 Authors with the Highest Publications

Name	Number of Publication	Affiliation*	Department	Total Citations	Avg. Citations
Y. Gerrard	5	University of Sheffield	Sociological Studies	176	35.2
S. M. West	4	University of Southern California	Annenberg School for Communication and Journalism	246	61.5
N. Suzor	4	QUT	Faculty of Law	104	26
R. Gorwa	3	University of Oxford	Department of Politics and International Relations	243	81
T. Gillespie	3	Cornell; Microsoft	Department of Communication	142	47.3
A. Matamoros-fernandez	3	QUT	School of Communication	54	18
G. M. Masullo	3	University of Texas at Austin	School of Journalism and Media	39	13
Bright, J.	3	University of Oxford	Oxford Internet Institute	39	13
M. J. Riedl	3	University of Texas at Austin	School of Journalism and Media	35	11.6
J. E. Gray	3	QUT	Creative Industries Faculty	19	6.3

*Authors' affiliations and departments are taken from the publications; their current affiliations may differ from this list.

c. Bibliographic coupling analysis in R6
d.Co-occurrence of keywords analysis in R7, R8

3. Results

3.1. R.Q. 1: Growth Pattern of Publications by Year

According to the results presented in Figure 2, there has been a steady increase in publications between 2016 and 2022. The number of publications peaked in 2023 with noticeable momentum. The number of publications in 2023 is approximately equal to the total number in the last three years. This increase also aligns with the discussion presented in the introduction about increasing attention to safeguarding the internet.

3.2. R.Q. 2: Authors who had the highest publication productivity

Figure 3: Co-authorship Network Among Authors

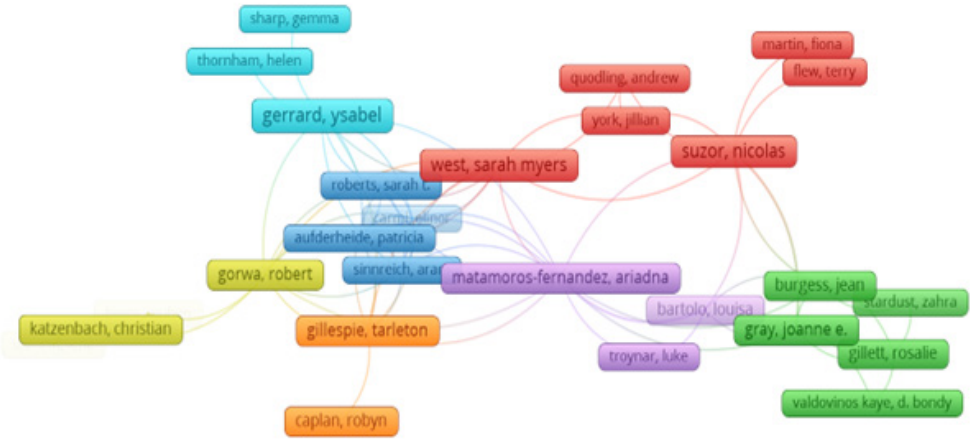
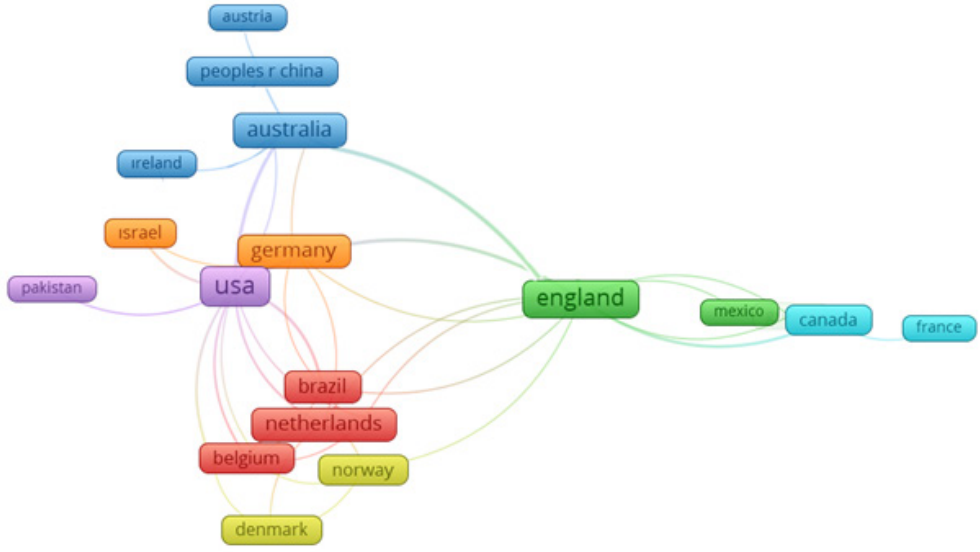


Figure 4: Co-Authorship Network Among Countries



The author with the highest publication productivity is Y. Gerrard, with five publications, followed by S.M. West and N. Suzor, with four publications. The University of Sheffield, the University of Southern California, and the Queensland University of Technology (QUT) are noticeable institutions with which the top three productive authors are affiliated. Additionally, Y. Gerrard, S.M. West, and R. Gorwa exhibit more significant influence and broader citation impact. In the authors' disciplines, it is noticeable that social sciences have the overwhelming majority, including sociology, communication and journalism, political sciences, and law.

Table 2: Top Journals

Top 10 Journals	Number of Publications	Total Citations	Avg. Citation	H-Index
New Media & Society	22	479	21.7	136
Social Media + Society	16	215	13.4	54
Big Data & Society	9	269	29.8	57
International Journal of Communication	9	58	6.4	52
Policy and Internet	9	122	13.5	38
Information Communication & Society	7	34	4.8	101
Internet Policy Review	7	146	20.8	24
Journal of Digital Media & Policy	7	85	12.1	9
Media Culture & Society	5	28	5.6	78
Computer Law & Security Review	4	32	8	49
Total	95	1468	15.45	

3.2. R.Q. 2: Authors who had the highest publication productivity

The co-authorship analysis reveals collaboration among 27 authors distributed across seven clusters, Figure 3. Clusters 1, 2, 3, 4, and 7 include seven authors listed in the top ten productive ones. S. M. West and N. Suzor are in Cluster 1. J. E. Gray is in Cluster 2; Y. Gerrard and A. Matamoros-Fernandez are in Cluster 3. R. Gorwa is in Cluster 4. Finally, T. Gillespie is in Cluster 7. According to the co-authorship analysis among countries, the USA, England, Australia, Germany, and Brazil are the most collaborative in content moderation, as shown in Figure 4.

3.4. R.Q. 4: Top Journals in Content Moderation Research

Ninety-nine journals have contributed to the subject in the given timeframe. According to the results in Table 2, New Media & Society has the highest citation number and H-Index. Furthermore, 10.91% (n=22) of the publications of the local dataset were published in this

journal. On the other hand, Big Data & Society has the highest average citation rate per paper and published 4.46 % (n=9) of the publications in this field of study.

As for the countries of the journals, 35.35% (n=35) of the journals are published in the United Kingdom, followed by 27.27% (n=27) in the United States and 11.11% (n=11) in the Netherlands.

Table 3: Top Universities

Top 10 Universities	# of Publications	Total Citations	Avg. Citations
University of Oxford	12	294	24.5
Queensland University of Technology	12	264	22
University of Amsterdam	10	54	5.4
University of Michigan	8	120	15
Cornell University	5	111	22.2
University of Sheffield	5	179	35.8
University of Pennsylvania	4	21	5.75
University of Southern California	6	193	32.1
University of Texas Austin	4	51	12.7
University of Sydney	4	72	18.5
Total	70	1363	

3.5. R.Q. 5: Institutions and countries that have contributed the most to the literature on content moderation

According to the results shown in Table 3, the University of Oxford is the leading university in content moderation studies. QUT follows it. The University of Amsterdam is in third place. However, it has a low impact, with the most minor citations per publication.

According to the results shared in Table 4, seven of the ten institutions that funded research are governmental institutions. Despite being the most prominent research funder in terms of the number of projects it supports, the European Union ranks among the least influential organizations with a comparatively low average citation rate. Canada's Social Sciences and Humanities Research Council is the most influential organization, with 30.2 average citation rates. The Australian Research Council has the third-highest average citation (22.1). Microsoft stands out as the most successful non-governmental funding agency, boasting the highest citations per publication, 39.3. In comparison, Microsoft and Google appear more notable than Meta and Twitter, which funded only two.

Table 4: Top 10 Funding Organizations of the Research

Funding Organizations	# of Publications	Total Citations	Avg. Citations
European Union*	10	61	6.1
Social Sciences and Humanities Research Council of Canada	7	212	30.2
National Science Foundation	7	131	18.7
Australian Research Council	6	133	22.1
Israel Science Foundation	3	33	11
Microsoft	3	119	39.3
Research Council of Norway	3	13	3.3
William and Flora Hewlett Foundation	2	24	12
Google Inc**	3	14	4.6
Dutch Research Council	2	8	4
Total	38	687	18.1

*Funding programs under "European Union" include: European Research Council, European Union Tailor, European Union, European Commission Joint Research Centre, and European Union Nextgenerationeu Prtr
**Funding programs under "Google Inc" include Google Inc and Google - Project Be Positive Under The 2019 Google Org Impact Challenge on Safety Call

3.6. R.Q. 6: Publications which have received the highest number of citations

The majority of the most cited papers are written by the authors who have contributed the most. Similarly, they are in the journals that have contributed most to the field by the number of publications. Additionally, the keywords provided in Table 5 illuminate the scope of these articles. Notably, the most cited articles predominantly belong to the domain of social sciences. Keywords such as internet policy, platform transparency, online protest, transparency, artificial intelligence, platforms, and demonetization are prevalent in these highly cited works.

The bibliographic coupling analysis, with the minimum number of citations of a document parameter set at 15, reveals collaboration among 33 documents distributed across four clusters, Figure 5. Our analysis indicated that the field is not monolithic but encompasses distinct research areas. Cluster 1, the red one, focuses on Online Dynamics, covering communication, social dynamics, governance, and emotional impact. It bridges the gap between technical moderation mechanisms and the socio-cultural aspects of online interactions. Cluster 2, the green one, highlights the interplay between governance, technology, and societal implications of content moderation, focusing on regulatory frameworks, technological innovations, and ethical considerations. Cluster 3, the blue one, addresses Algorithmic Governance and Expression. It explores the relationship between

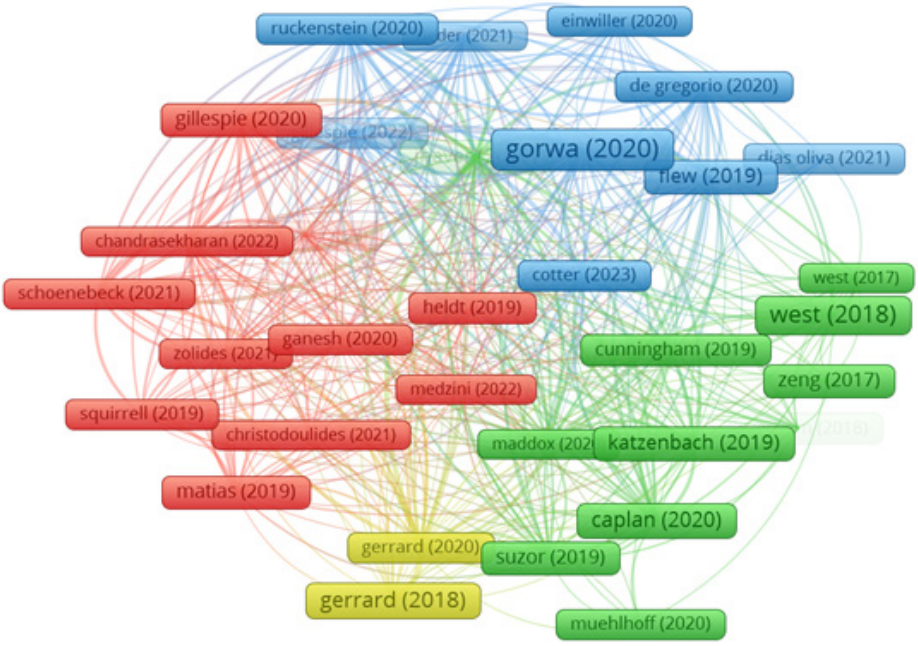
Table 5: Top 10 Publications Based on the Number of Citations

Author(s)	Total Citations	Journal	Title	Keyword(s)
Gorwa et al. 2020	185	Big Data & Society	Algorithmic content moderation: Technical and political challenges in the automation of platform governance	algorithms, artificial intelligence, content moderation, copyright, platform governance, toxic speech
West 2018	139	New Media & Society	Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms	accountability, content moderation, free expression, social media, survey, transparency, user studies
Gerrard 2018	99	New Media & Society	Beyond the hashtag: Circumventing content moderation on social media	algorithms, anorexia, content moderation, eating disorders, hashtags, Instagram, Pinterest, pro-ana, social media, Tumblr
Caplan and Gillespie 2020	75	Social Media + Society	Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy	apocalypse, advertising, content moderation, demonetization, digital intermediaries, platforms, YouTube
Flew et al. 2019	63	Journal of Digital Media and Policy	Internet regulation as media policy: Rethinking the question of digital communication platform governance	media policy, digital platforms, platform capitalism, content moderation, classification, media regulation, intermediaries, platform governance
Jhaver et al. 2019	59	ACM Transactions on Computer-Human Interaction	Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator	content moderation, automated moderation, automod, platform governance, mixed-initiative, future of work
Katzenbac 2019	58	Internet Policy Review	Algorithmic governance	transparency, automation, politicization, regulation, social ordering, governance, predictive policing, content moderation, algorithmic governance

Table 5 (continued)

Zeng et al. 2017	47	Policy & Internet	How Social Media Construct "Truth" Around Crisis Events: Weibo's Rumor Management Strategies After the 2015 Tianjin Blasts	Internet censorship, online rumor, content moderation, emergency communication, online protest, collective action
Gillespie et al. 2020	46	Internet Policy Review	Expanding the debate about content moderation: scholarly research agendas for the coming policy debates	content moderation, platforms, internet policy, social media, regulation
Suzor et al. 2019	46	International Journal of Communication	What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation	content moderation, platforms, transparency, due process

Figure 5: Bibliographic Coupling Analysis of Publications



algorithmic governance, freedom of expression, and socio-cultural dimensions, focusing on ethical, legal, and technological aspects of content moderation, particularly algorithmic decision-making. Cluster 4, the yellow one, explores the impact of social media on mental health awareness, advocacy, and specific health issue discussions.

While each cluster has a unique thematic emphasis, common themes like governance and regulation appear prominently in Clusters 2 and 3, reflecting the ongoing discourse on legal and ethical aspects. Terms related to online behavior, such as communication and social dynamics, indicate a shared interest in understanding user interactions across all clusters. On the other hand, the clusters differ in their theme specificity. Cluster 1 covers a comprehensive range of topics, providing a holistic overview of content moderation research. In contrast, Clusters 2 and 3 delve into more specialized areas, with Cluster 2 focusing on governance and technology and Cluster 3 exploring algorithmic governance and freedom of expression.

The diversity within these clusters highlights the interdisciplinary nature of content moderation research. Researchers approach content moderation from various angles, from technical aspects to legal and ethical considerations, social dynamics, and mental health advocacy. This diversity underscores the need for collaborative efforts integrating insights from different clusters to develop comprehensive and effective content moderation strategies.

3.7. R.Q. 7: Keyword Analysis

A keyword represents the main topics explored in the document, aiding in indexing and categorization for readers. The dataset has 628 keywords, with the top 3 being social media, platform governance, and Facebook. 1.5 egocentric network analysis of the "Content Moderation" keyword, with a minimum occurrence parameter set at 5, revealed six clusters representing topic relationships, Figure 6. These clusters depict relationships among topics, with the thickness of connecting lines indicating the strength of keyword pairs, and the nodes' size signifies the keyword's frequency. Each cluster shows the interconnectedness and thematic cohesion among keywords within the broad context of content moderation.

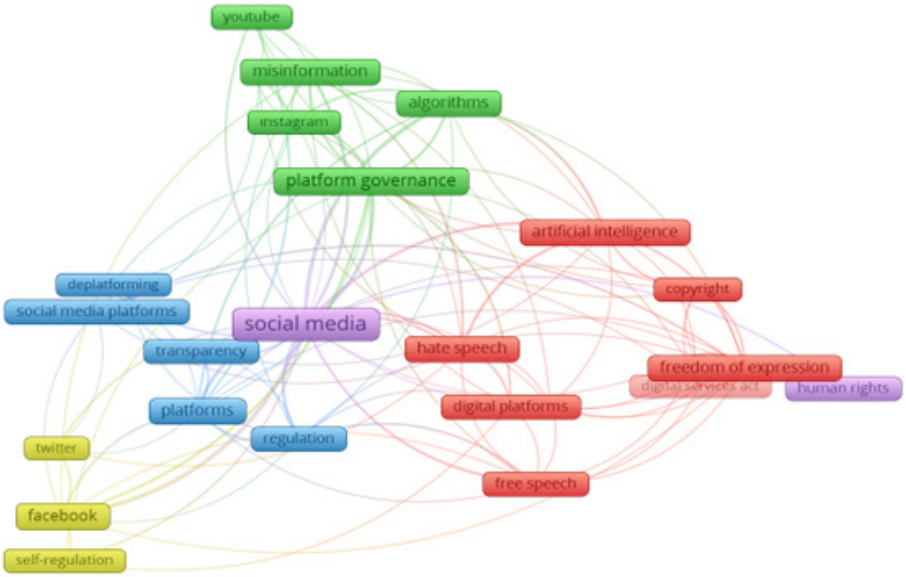
Cluster 1: Legal and Regulatory Aspects:

The keywords of this cluster are artificial intelligence, copyright, digital platforms, the Digital Services Act, free speech, freedom of expression, hate speech, online platforms, and platform regulation. This cluster focuses on the legal and regulatory aspects of content moderation. It covers topics like the legal implications of artificial intelligence, copyright issues, and the role of regulations in governing digital platforms. Terms like hate speech and freedom of expression suggest a focus on balancing regulatory measures with preserving free speech online.

Cluster 2: Technological and Platform-Specific Focus

The keywords of this cluster are algorithms, Instagram, machine learning, misinformation, platform governance, and YouTube. Cluster 2 centers on technological

Figure 6: Thematic Clusters and Relationships Among Keywords



aspects of content moderation and emphasizes algorithms, machine learning, and the problems posed by misinformation. Platform-specific terms such as Instagram and YouTube suggest a focus on understanding and addressing content moderation challenges unique to these platforms.

Cluster 3: Algorithmic Governance and Transparency

The keywords of this cluster are algorithmic governance, de-platforming, platforms, regulation, social media platforms, and transparency. This cluster emphasizes the intersection of algorithmic governance, transparency, and regulatory measures in content moderation. Terms like de-platforming suggest a focus on the decisions made by platforms regarding removing certain content or users. Social media platforms indicate a broader consideration of these issues within the social media landscape.

Cluster 4: Platform-Specific Analysis

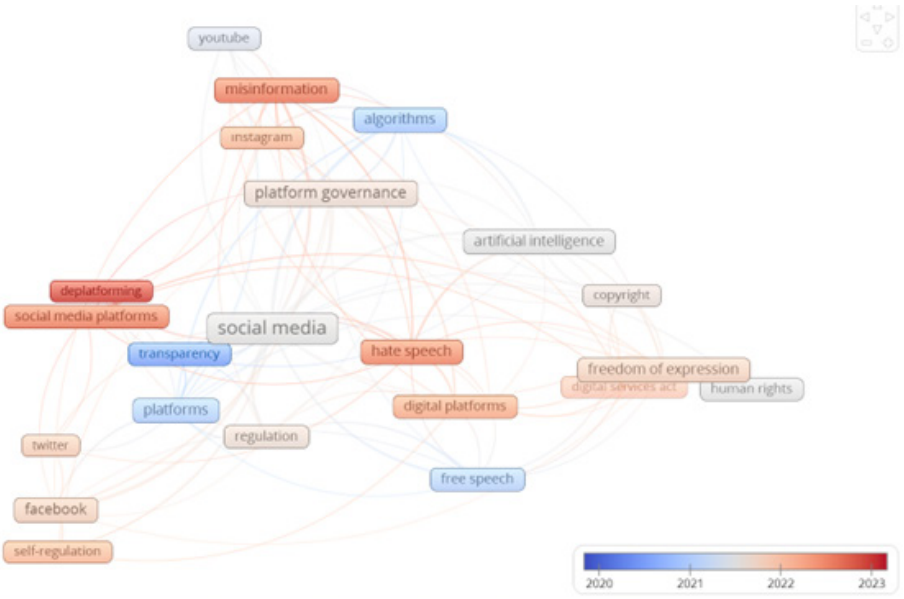
The keywords of this cluster are Facebook, self-regulation, and Twitter. Cluster 4 focuses on specific social media platforms—Facebook and Twitter. It suggests a detailed examination of content moderation issues within these platforms, including considerations of self-regulation and policies implemented by these companies to manage content.

Cluster 5: Human Rights and Social Media

The keywords of this cluster are human rights and social media. This cluster emphasizes the intersection of content moderation with human rights considerations. It suggests exploring the impact of content moderation practices on users' rights within the context of social media.

The network analysis reveals the multidimensional nature of the field, including legal, technological, platform-specific, governance, and human rights aspects.

Figure 7: Research Trends Between 2020 and 2023



3.8. RQ8: Analysis of research trends

1.5 egocentric network analysis was conducted on the keyword Content Moderation with a minimum occurrence parameter set at five. Overlay visualization was utilized to monitor the evolution of research trends over the past three years, Figure 7. In early 2020, the focus was on transparency. By late 2020, the emphasis had shifted to algorithms. In early 2021, the main areas of interest were free speech and platforms, while later in the year, the focus expanded to platform governance, social media, YouTube, human rights, artificial intelligence, copyright, and regulation. The trend continued to evolve in early 2022, with research centering on self-regulation, the Digital Services Act, Instagram, Twitter, Facebook, and freedom of expression. In late 2022, the focus shifted to hate speech and digital platforms. Early 2023 saw an emphasis on misinformation, and by late 2023, the main areas of interest were social media platforms and deplatforming.

The trends culminated in late 2022 when research interest prominently shifted towards deplatforming, misinformation, hate speech, and social media platforms in 2023. This trajectory highlights content moderation research's dynamic and adaptive nature, reflecting an evolving response to emerging challenges and contemporary issues within the digital landscape.

4. Discussion and Conclusion

Internet integration necessitates formal and informal measures to maintain a secure and positive online and offline environment. Thus, content moderation has become a vital topic today. Our analysis reveals a rise in publications on content moderation, particularly in 2023, corresponding with its increased significance due to the surge in hate speech and misinformation during the global COVID-19 pandemic and election periods. Keyword trends point to the complex relationship between the internet, social media platforms, and democratic ideals. The emphasis on transparency in early 2020 suggests a shared effort to increase openness and accountability online. However, the subsequent shift towards algorithms raises questions about the impact of automated content moderation systems.

As mentioned in the introduction, online platforms have intervened in the content they host in the virtual sphere since the proliferation of the internet as a medium. However, the term 'content moderation' may refer to not only an increasing interest in safeguarding of the virtual sphere but also the crystalized notion on which these interests (or worries) are addressed. The earliest mention of the term was in 2016, even though the point of discussion had previously existed; thus, 'content moderation' became a specific area in which the worries were platformed.

To understand this better, we can look into 2016, arguably the pathway to the 2016 U.S. Presidential elections, and the Brexit referendum has turned the spotlights to the virtual sphere. However, the same virtual sphere was also influential during the 2008 U.S. Presidential Elections. The online activities of "Obama Boys", the fans of Barack Obama (who was only a nominee then) were making their presence felt (Traister, 2008). Thus, the foundation of the term 'content moderation' in 2016 was a moment in which the momentum of interest in the field crystallized. That would parallel the change in people's perception of the virtual sphere and its safeguarding.

Considering the research method, it is possible for the term 'content moderation' to be used earlier in the journals that were not listed in the selected databases. Even with any earlier usage of the term, 'content moderation' became a point of interest in the mid-2010's. Moreover, although the study has used the word 'content moderation,' there seems to be no strong alternative to it; in our findings, there have been terms used such as "platform governance," "platform regulation," "internet governance," "algorithmic governance" and "self-regulation." Although these terms do not occur frequently, they diversify the perspectives within the literature.

Our analysis showed that there is an increasing amount of publication on content moderation following its introduction to academic writing, especially in 2023. This would expectly suggest that there is an increased interest within the area. But it also proves that content moderation is getting solidified as a term to address the safeguarding of the internet. It could also suggest that it is a fruitful area for interdisciplinary research.

For example, one of the most cited papers in the collection is from Ysabel Gerrard,

who focuses on user behavior to 'circumvent' content moderation. Her niche is content advising for behavior such as extreme diets that contribute to eating disorders like anorexia. Gerrard's study does show that researchers focusing on a particular online issue (for example, sports fans' behaviour online) could address the role of content moderation in their studies.

Influential authors and institutions are from Australia, England, and the USA. Judging by the influential authors and articles, it can be argued that social sciences like communication, law, and political sciences are the most cited and published academic fields. Other than social sciences, computer sciences are also visible within the field.

Although our dataset contains articles about alternative platforms or other branches of internet services, discussions regarding content moderation are overwhelmingly focused on major platforms such as Instagram, Facebook, Twitter (X), TikTok, and Reddit. However, it is surprising that the tech platforms themselves have yet to be absent from sponsoring these studies. Although it is expected that state grants would shadow private enterprises, the lack of platforms' involvement may indicate either their unwillingness to conduct their research or their willingness to keep such research private.Judging by the keywords artificial intelligence, freedom of speech, platform transparency, hate speech, misinformation, and human rights are the issues discussed within the literature, and deplatforming, hate speech, and misinformation are emerging discussions.

An issue requiring the attention of the literature seems to be Elon Musk's takeover of X; in this instance, a billionaire not only has bought one of the largest social media platforms existing on the web but seemingly has politicized the platform to his own by changing the content moderation policies. Not only are Musk's changes impacting real-life politics, but Musk himself had a certain success in canceling -and possibly discouraging- the legislation regulating social media content moderation practices.

This case especially underlines the importance of an important question: whether internet companies are reliable in terms of their effectiveness and dedication to safeguarding the web. This possibly emphasizes the importance of supranational organizations' regulation efforts (such as the European Union) and the possibility of international enforcement of such regulations on these platforms.

Another aspect that Musk's takeover points out is the ownership structures of these platforms, and related to Musk's takeover, questions on the nature of commercial ownership and the possibilities of duopoly may be asked.

The recent advancements in artificial intelligence (A.I.) will likely shape future research in the field. Although topics such as algorithms and automated learning have already been explored in the literature, future studies are expected to integrate A.I. more deeply into content moderation practices. Additionally, A.I. may also become a research subject, particularly in how it can be used to generate misleading content or provoke public outrage.

Considering that the real-life impact of online content includes more than electoral processes and may influence riots and protests, it is likely that topics like hate speech, fake news, misinformation, and disinformation will stay relevant. A likely case study would be the 2024 U.S. Presidential Elections, for not only major platforms have followed changes in content moderation, but one of the candidates, Donald Trump, has founded his own social media platform, Truth Social, in which he can communicate with the electorate unfiltered and unmoderated.

Acknowledgements

This article is an extended version of the conference proceeding titled "Exploring Content Moderation Research: Insights from a Bibliometric Analysis," presented at the European Conference on Social Media, held on May 21, 2024, in Birmingham, UK. The original conference paper can be accessed via DOI: 10.34190/ecsm.11.1.2114.

References

Amnesty International. (2022, September, 29). *Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – new report.* https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Angwin, J. & Grassegger, H. (2017, June, 28). Facebook's secret censorship rules protect white men from hate speech but not black children. *ProPublica.* https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

Audureau, W. (2022, August, 10). Alex Jones trial: A record fine intended to make misinformers pay. *LeMonde',*. https://www.lemonde.fr/en/les-decodeurs/article/2022/08/10/alex-jones-trial-a-record-fine-intended-to-make-misinformers-pay_5993081_8.html

Barrett, P.M. (2020). *Who moderates the social media giants? A call to end outsourcing.* NYU STERN. https://bhr.stern.nyu.edu/wp-content/uploads/2024/02/NYUContentModerationReport_FINALVERSION.pdf

Benckendorff, P., & Zehrer, A. (2013). A network analysis of tourism research. *Annals of tourism research, 43,* 121-149. https://doi.org/10.1016/j.annals.2013.04.005.

Broadus, R.N. (1987). Toward a definition of "bibliometrics". *Scientometrics,* 12(5–6),373–379. https://doi.org/10.1007/BF02016680.

Byman, D. (no date). *Content moderation tools to stop extremism.* Lawfare (2022nd ed.). https://www.lawfaremedia.org/article/content-moderation-tools-stop-extremism

Caplan, R. (2018) *Content or context moderation? Artisanal, community-reliant, and industrial approaches.* Data & Society.https://datasociety.net/library/content-or-context-moderation/

CBS News. (2022, November, 14). Musk fires outsourced content moderators who track abuse on Twitter. *CBS News.* Available at: https://www.cbsnews.com/news/elon-musk-twitter-layoffs-outsourced-content-moderators/

CBS News. (2024, March, 25). Judge tosses out X lawsuit against hate-speech researchers, saying Elon Musk tried to punish critics. *CBS News.* Available at: cbsnews.com/news/elon-musk-x-lawsuit-dismissed-hate-speech/.

Center for Countering Digital Hate. (2023). Twitter fails to act on 99% of Twitter Blue accounts tweeting hate. https://counterhate.com/research/twitter-fails-to-act-on-twitter-blue-accounts-tweeting-hate/#about.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for information Science and Technology, 62*(7), 1382-1402. https://doi.org/10.1002/asi.21525.

Debre, I. and Akram (2021, October, 26) 'Facebook's language gaps weaken screening of hate, terrorism'. *The Associated Press.* https://apnews.com/article/the-facebook-papers-language-moderation-problems392cb2d065f81980713f37384d07e61f?utm_source=copy&utm_medium=share

De Guzman, C. (2024, March, 27). What to Know About Meta's 'Political Content' Limit—and How to Turn It Off on Instagram. Time. https://time.com/6960587/meta-instagram-political-content-limit-off-setting-default/

Dey, M. (2023, December, 29). Elon Musk's X fails to block California's content moderation law. *Reuters.* https://www.reuters.com/sustainability/society-equity/elon-musks-x-fails-block-californias-content-moderation-law-2023-12-29/

D.W. News. (2023, October, 11). German anti-racism body leaves X over 'rise in hate speech'. *https://www.dw.com/en/german-anti-racism-body-leaves-x-over-rise-in-hate-speech/a-67065363.*

Ferrara, E. et al. (2020) 'Characterizing social media manipulation in the 2020 U.S. presidential election', *First Monday* [Preprint]. https://doi.org/10.5210/fm.v25i11.11431

Field of Vision. (2017, April, 14. *Field of Vision – The Moderators* [Video]. Youtube. https://www.youtube.com/watch?v=k9m0axUDpro&t=1s

Freedland, J. (2024, August, 9) You know who else should be on trial for the U.K.'s far-right riots? Elon Musk. *The Guardian.* https://theguardian.com/commentisfree/article/2024/aug/09/uk-far-right-riots-elon-musk-x

Fung, B. (2021, January, 9) 'Twitter bans President Trump permanently'. CNN. https://edition.cnn.com/2021/01/08/tech/trump-twitter-ban/index.html

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society, 20*(12), 4492-4511. https://doi.org/10.1177/1461444818776611

Hall, C. M. (2011). Publish and perish? Bibliometric analysis, journal ranking and the assessment of research quality in tourism. *Tourism Management, 32*(1), 16–27. https://doi.org/10.1016/j.tourman.2010.07.001.

Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. Proceedings of the *National Academy of Sciences, 118*(32), e2101967118. https://doi.org/10.1073/pnas.2101967118.

Klepper, D. & O'Brien, M. (2022, December, 4). As Musk is learning, content moderation is a messy job. A.P. News. https://apnews.com/article/kanye-west-elon-musk-twitter-inc-entertainment-technology-0bf6e0ab969a60cd38abd9358ee5fd47

Koseoglu, M.A., Rahimi, R., Okumus, F., & Liu, J. (2016). Bibliometric studies in tourism. *Annals of Tourism Research, 61,* 180–198. https://doi.org/10.1016/j.annals.2016.10.006.

Levy, S. (2022, September, 23) 'Bogus Fears of Censorship Could Spell the End of Content Moderation'. *Wired*. https://www.wired.com/story/plaintext-bogus-fears-of-censorship-could-spell-the-end-of-content-moderation/

Lim, W.M. & Kumar, S. (2024). Guidelines for interpreting the results of bibliometric analysis: A sensemaking approach. *Global Business and Organizational Excellence, 43*(2), 17–26. https://doi.org/10.1002/joe.22229.

Liu, Y., Yildirim, P., & John Zhang, Z. (2022). Implications of revenue models and technology for content moderation strategies. *Marketing Science, 41*(4), 403–419. https://doi.org/10.1287/mksc.2022.1361.

Maheshwari, S. (2023, March, 23) 'What to Know About Today's Congressional Hearing on TikTok'. *New York Times.* https://www.nytimes.com/2023/03/23/technology/tiktok-congress-hearing.html

MSNBC (2021, February, 21) 'Why Content Moderation Costs Social Media Companies Billions'. [Video]. Youtube. https://www.youtube.com/watch?v=OBZoVpmbwPk

Özbudun, S.Y. (2022, October, 16) 'Basında sansür yasası ve Türkiye'yi bekleyen ihtimaller'. *Politikyol.* https://www.politikyol.com/basinda-sansur-yasasi-ve-turkiyeyi-bekleyen-ihtimaller/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaf, J.M., Akl, E.A., Brennan S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj, 372*. https://doi.org/10.1136/bmj.n71

Phillips, T. (2024, August, 31). X goes offline in Brazil after Elon Musk's refusal to comply with local laws. *The Guardian.* https://theguardian.com/technology/article/2024/aug/31/x-offline-brazil-elon-musk

Perrin, A. and Atske, S. (2021, March, 26) 'About three-in-ten U.S. adults say they are "almost constantly" online', Pew Research Center. https://www.pewresearch.org/short-reads/2021/03/26/about-three-in-ten-u-s-adults-say-they-are-almost-constantly-online/
Pritchard, A. (1969). Statistical bibliography or bibliometrics? Journal of Documentation, 24, 348–349.

Sato, M. (2022, April, 14). Buying Twitter 'is not a way to make money,' says Musk in TED interview. *The Verge.* https://www.theverge.com/2022/4/14/23025343/elon-musk-twitter-takeover-ted-talk-quote-stock-buyout.

Scarcella, M. (2024, September, 4). Elon Musk's X wins appeal to block part of California content moderation law. *Reuters.* https://www.reuters.com/legal/elon-musks-x-wins-appeal-block-part-california-content-moderation-law-2024-09-04/

Suzor, N.P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? toward meaningful transparency in commercial content moderation. *International Journal of Communication,* 1526 – 1543. https://ijoc.org/index.php/ijoc/article/view/9736

Tangalakis-Lippert, K. (2022, November, 7). While Elon Musk said 'comedy is now legal' following his acquisition of Twitter, jokes about the new owner and criticism over his takeover are getting users blocked and suspended. *Business Insider.* https://businessinsider.com/musk-said- comedy-now-legal-twitter-jokes-getting-users-suspended-2022-11

Traister, R. (2008, April, 14). Hey, Obama boys: Back off already! *Salon.* https://salon.com/2008/04/14/obama_supporters/

Treisman, R. (2024, March, 26). Meta is limiting how much political content users see. Here's how to opt out of that. *NPR*. https://npr.org/2024/03/26/1240737627/meta-limit-political-content-instagram-facebook-opt-out

Tüfekçi, Z. (2018, March, 10) 'YouTube, the great radicalizer', *New York Times.* https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Veglis, A. (2014). Moderation techniques for social media content. In G. Meiselwitz (Ed.), *Social Computing and Social Media* (pp. 137-148). Cham: Springer International Publishing (Lecture Notes in Computer Science). https://doi.org/10.1007/978-3-319-07632-4_13

West, S. M. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society,* 4366 – 4383. https://doi.org/10.1177/1461444818773059

Wike, R. Silver L. Fetterolf, J. Huang, C. Austin, S. Clancy, L and Gubbala, S. (2022, December, 6).'Social Media Seen as Mostly Good for Democracy Across Many Nations, But U.S. is a Major Outlier'. *Pew Research Center.* https://www.pewresearch.org/global/2022/12/06/social-media-seen-as-mostly-good-for-democracy-across-many-nations-but-u-s-is-a-major-outlier/

Zupic, I. & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods, 18*(3), 429–472. https://doi.org/10.1177/1094428114562629