Exploring Content Moderation Research: Insights from a Bibliometric Analysis

Ozlem Ozan and Ali Rıza Sadıkzade
Yasar University, İzmir, Türkiye

ozlem.ozan@yasar.edu.tr
ali.riza.sadikzade@gmail.com

Abstract: Rapid technological advances have intensified user-content interactions, leading to real-world consequences and the implementation of complex regulation mechanisms such as AI filtering and industrial and user moderation. This study aims to introduce the contemporary topics surrounding the subject by comprehensively examining the content moderation research by conducting a bibliometric analysis of 202 publications between 2016 and 2023 from the Web of Science and Scopus databases. This study aims to identify the influential authors, universities, countries, journals, funding agencies, network maps of keywords, and co-authorship. The findings of this study demonstrate that the Queensland University of Technology is the most influential in the field. The United States of America, England, and Australia are the most productive countries. The National Science Foundation and the European Research Council are the most supporting funding institutions. New Media & Society, Social Media + Society, and Big Data & Society are the most influential journals. Ysabel Gerrard is the most productive author. Seven clusters occur in author collaboration networks. The network map of the keywords suggests that researchers mainly focus on social media; Facebook, Instagram, YouTube, and Twitter are the most investigated platforms. There is a shift from transparency to hate speech and misinformation among the research themes. The academic research has exhibited a consistent upward trajectory since 2016. Given the demonstrable momentum of interest in this field, it is reasonable to anticipate a further increase in research with a diverse array of academic disciplines.

Keywords: Content moderation, Bibliometric analysis, Bibliometric data, Social media

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

The internet's pervasive integration into daily life has blurred the lines between online and offline experiences, prompting extensive discussions on its multifaceted implications. As the internet's salience grows, discussions span a wide range, from the positive effects of social media, such as enhanced connectivity and access to diverse opinions, to concerns like increased polarization and manipulative misinformation (Wike et al. 2022).

Beyond organic issues, political actors leverage the internet to advance their agendas, employing tactics like disinformation campaigns witnessed in the 2016 US Presidential Election (Ferrara et al. 2020). The online landscape also poses challenges at personal and interpersonal levels, including data privacy, copyright, internet scams, and cyberbullying.

Given these issues, safeguarding online individuals and communities has become imperative. Legislative efforts, such as the US Communications Decency Act's Section 230, passed in 1996, aimed to limit online platform liabilities, although tech companies actively reviewed content to protect their user base (Angwin and Grassegger, 2017; Caplan 2018). Recent discussions point to an increased institutional authority, as seen in legislative examples like the European Union's Digital Services Act and Germany's NetzDG law.

The content moderation market has grown substantially, reflecting the heightened need for online safeguarding. Content moderation involves diverse methods, from algorithmic processes to user cooperation and active moderation teams (MSNBC 2021; Veglis 2014).

While typically executed by platforms, content moderation also involves online service providers enforcing ethical guidelines on clients and engaging in the moderation process (Byman 2022). Moderation methods vary based on platform size and financial structure, with larger platforms employing extensive moderation teams (Caplan 2018; Liu et al. 2022). Challenges persist, including concerns about the adequacy of human moderators in understanding content nuances and potential biases in moderation processes (Debre and Akram 2021).

Despite the rich insights from cited studies, comprehensive articles scarcely summarize the content moderation discourse. Therefore, the primary objective of this study is to conduct a bibliometric analysis of the literature, summarizing contemporary discussions through bibliographic analysis of academic articles. This research aims to contribute valuable insights for future researchers examining the evolving landscape of content moderation.

2. Methodology

This section covers research methodology adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (BMJ 2021), which mainly includes the key elements as rationale and

aim of the research, search 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. Bibliometric analysis, as defined by Broadus (1987), 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

2.2.1 Database selection

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

2.2.2 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.

2.2.3 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

2.3.1 Screening process

The authors of the study conducted the screening of titles and abstracts independently to identify potentially eligible studies.

2.3.2 Data extraction

The standardized metadata provided by WOS and Scopus databases for bibliometric analysis is used to collect 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.

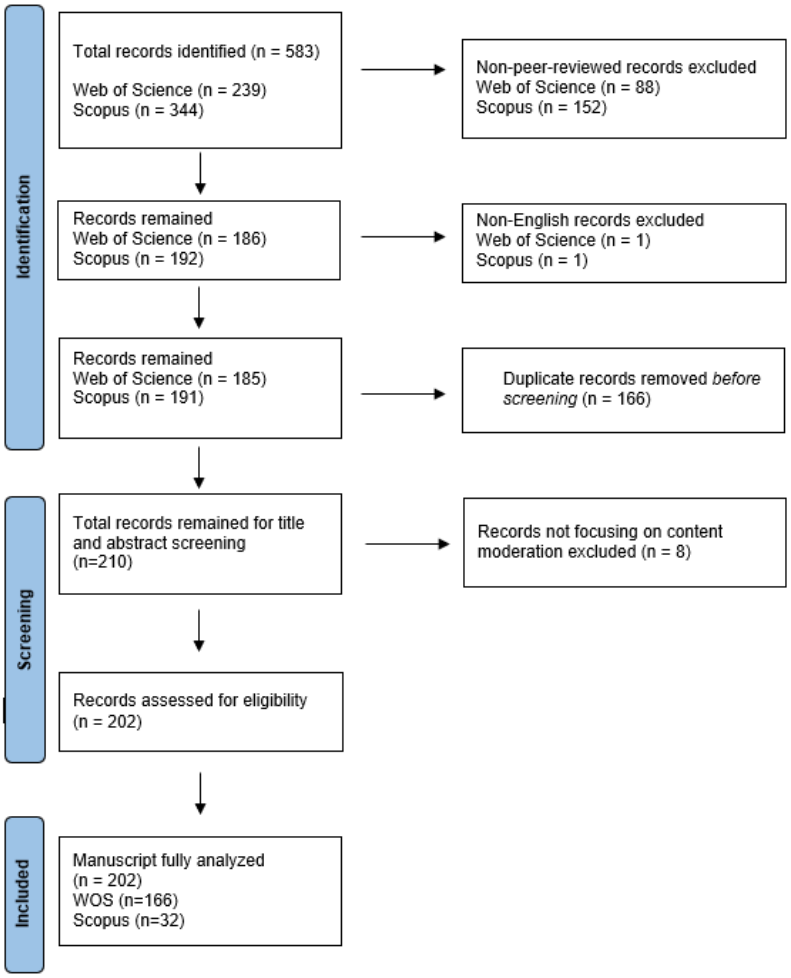

Figure 1: PRISMA framework of the study

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 MS 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 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:

- Descriptive statistics based on metrics in RQ1, RQ2, RQ3, RQ4, R5, R6, R7, R8
- Co-authorship analysis (authors and countries) in R3
- Bibliographic coupling analysis in R6
- Co-occurrence of keywords analysis in R7, R8

3. Results

3.1 RQ 1: Growth Pattern of Publications by Year

According to the results presented in Figure 2, there has been a steady increase in publications between 2016-2022. The number of publications peaked in 2023 with a 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.

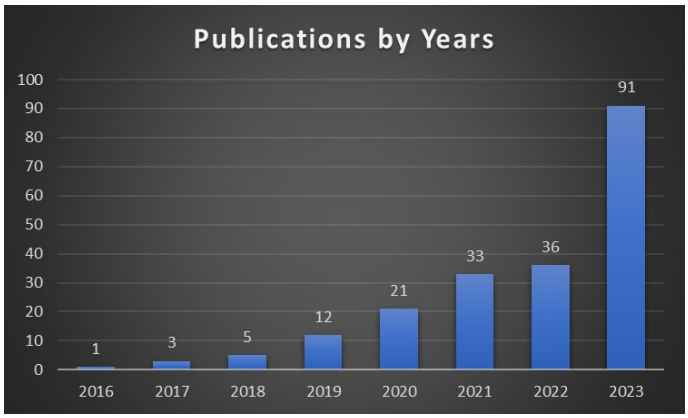

Figure 2: Growth pattern of publications by years

3.2 RQ 2: Authors who had the highest publication productivity

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 1: 10 Authors with the most publication

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

Name	Number of Publication*	Affiliation	Department	Total Citations	Avg Citations
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.

3.3 RQ 3: Networks among Authors

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.

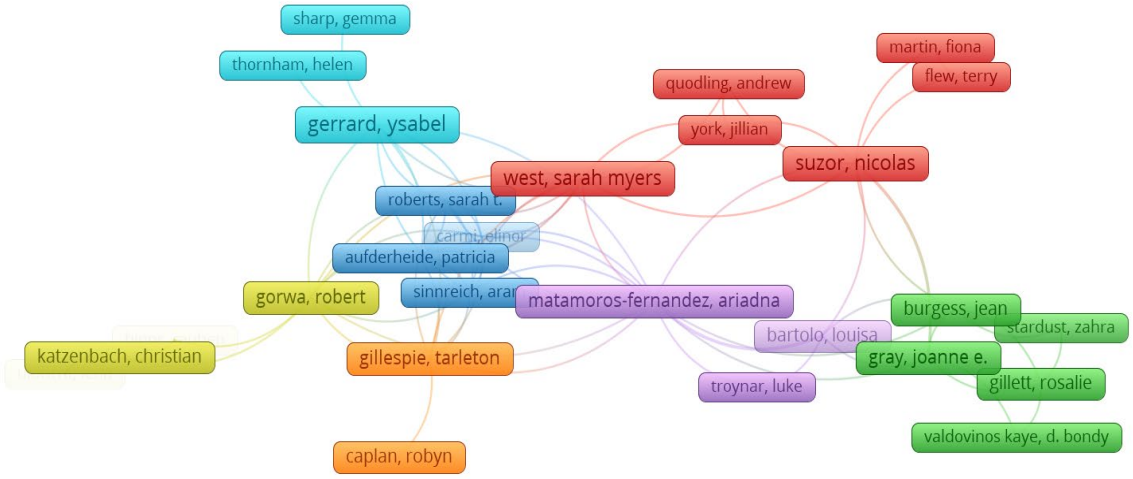

Figure 3: Co-authorship network among authors

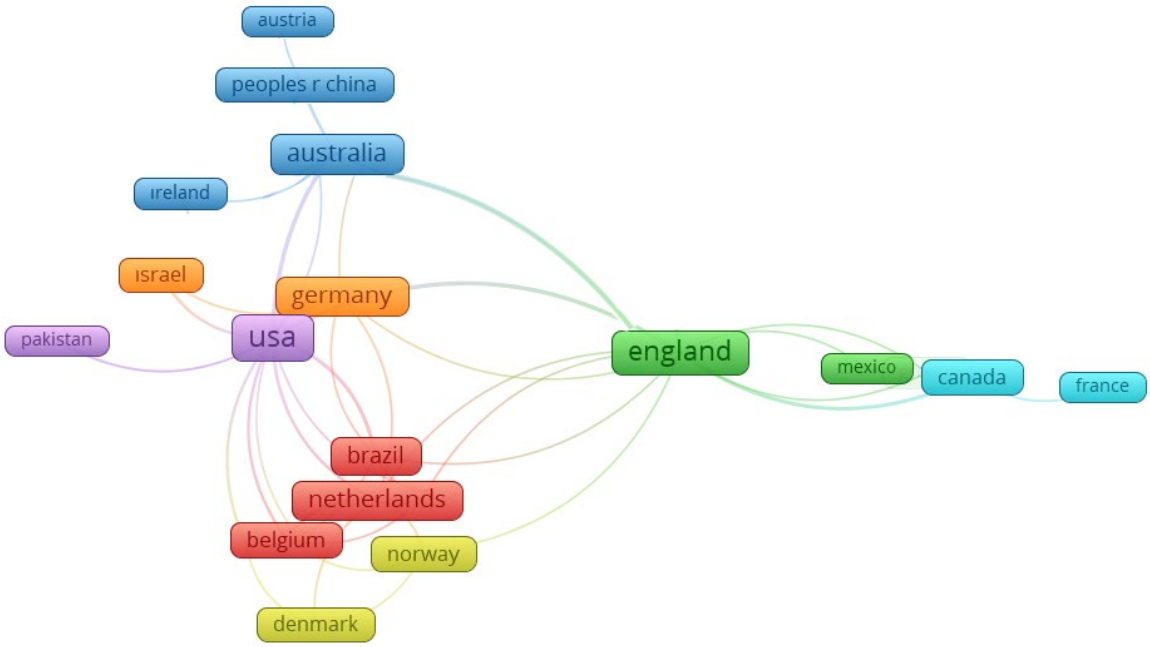

Figure 4: Co-authorship network among countries

3.4 RQ 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.

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	

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.

3.5 RQ 6: Publications Which Have Received the Highest Number of Citations

Most 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 3 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.

Table 3: 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
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

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 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.

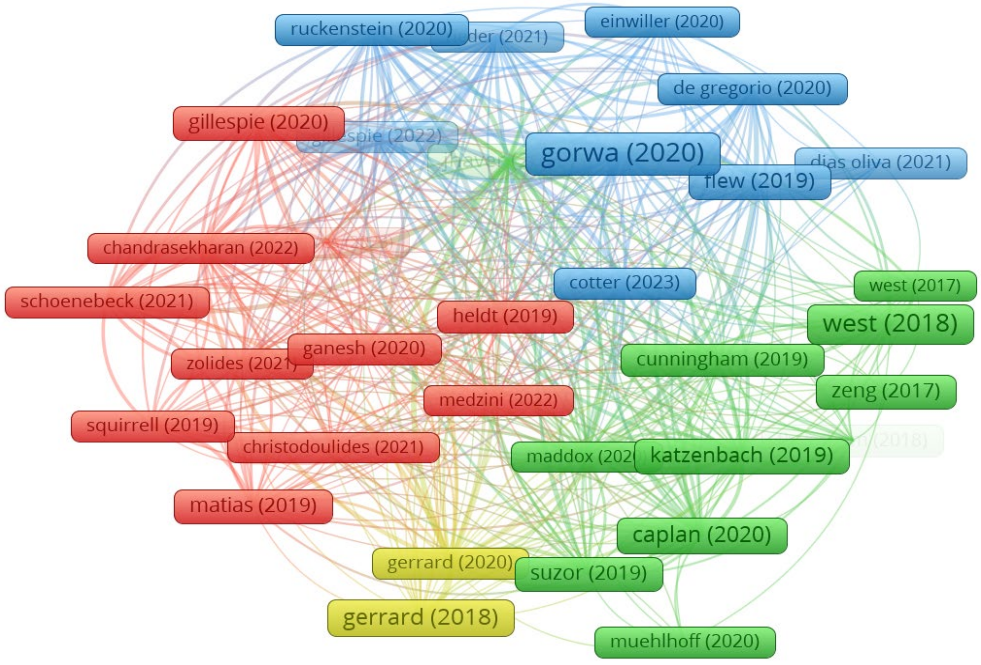

Figure 5: Bibliographic coupling analysis of publications

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.

3.6 RQ 5: Institutions and Countries That Have Contributed the Most to the Literature on Content Moderation

According to the results shown in Table 4, 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 5, seven of the ten institutions that funded research are governmental institutions. Despite being the most prominent research funder regarding 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 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	

Table 5: 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

Funding Organizations	# of Publications	Total Citations	Avg. Citations
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

3.7 RQ 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 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.

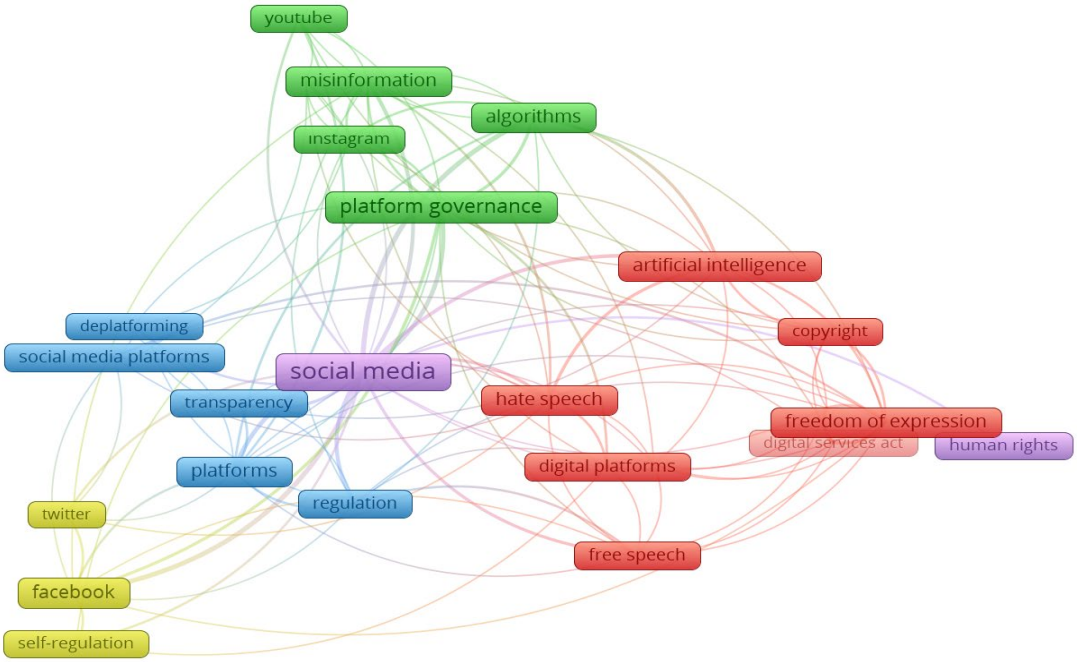

Figure 6: Thematic clusters and relationships among keywords

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

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.

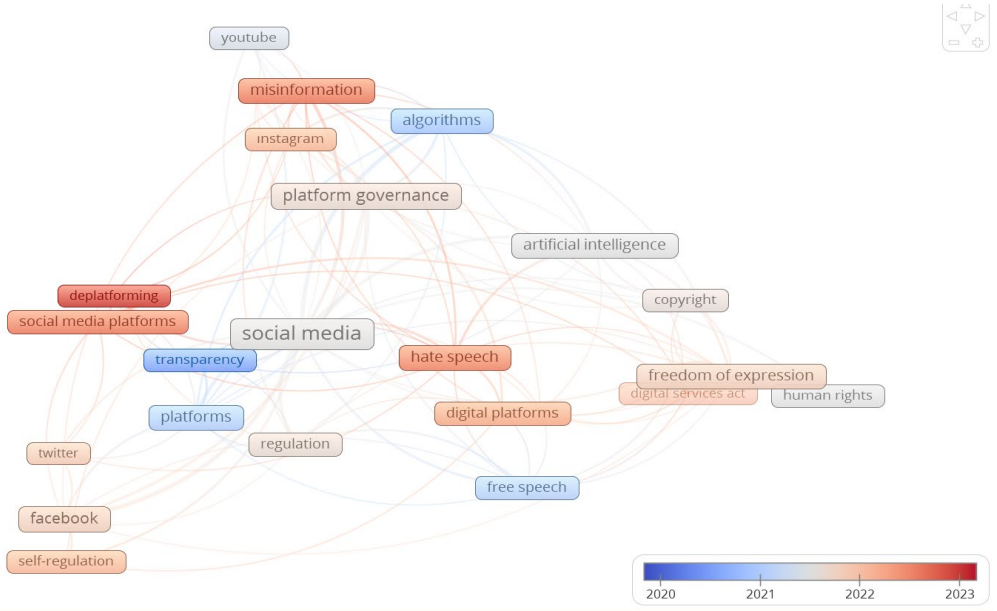

Figure 7: Research trends between 2020 and 2023

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, but the subsequent shift towards algorithms raises questions about the impact of automated content moderation systems.

Social media platforms, seen as tools for democratization, play a crucial role in global discourse. The keyword trends reflect a progression from a focus on free speech and platforms to a broader exploration of platform governance, social media dynamics, and algorithmic impacts on various aspects such as human rights, artificial intelligence, and regulation. However, the trends also highlight concerns about algorithmic recommendation systems' potential threats, such as fostering environments conducive to hate speech and misinformation spread.

Research predominantly focuses on major platforms like Facebook, YouTube, Twitter, and Instagram. This focus might inadvertently reinforce capital-driven models and neglect alternative platforms or narratives.

Prominent authors and institutions from Australia, England, and the USA dominate content moderation research, suggesting expertise and collaboration concentration in these regions. In contrast, countries unprepared for AI may face challenges in dealing with the ethical and legal considerations related to evolving content moderation practices.

In conclusion, while the internet and social media platforms offer potential for a more democratic world, content moderation research trends indicate potential threats to democracy. Algorithm-driven polarization of online spaces can lead to environments that limit exposure to diverse perspectives. Addressing this requires a nuanced approach that balances algorithmic content curation benefits with democratic values preservation. Initiatives prioritizing transparency, user agency, and algorithmic accountability are vital, as is ongoing interdisciplinary research and collaboration among scholars, policymakers, and platform developers. Prioritizing transparency, user agency, and algorithmic accountability can help navigate the complex landscape of content moderation and mitigate the unintended consequences of algorithmic systems, harnessing the internet and social media platforms' potential to contribute positively to democratic discourse and global idea exchange.

References

Angwin, J. and Grassegger, H. (2017) “Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children”, *ProPublica,* [online], 28 June. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

Benckendorff, P. and Zehrer, A. (2013) “A Network Analysis of Tourism Research”. *Annals of Tourism Research, Vol* 43, October 2013, pp 121-149.

Broadus, R. N. (1987). “Toward a definition of ‘bibliometrics’”, *Scientometrics*, Vol 12, pp 373–379. https://doi.org/10.1007/BF02016680

Byman, D. (2022) “Content Moderation Tools to Stop Extremism”, *Lawfare,* [online], 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, [online], https://datasociety.net/library/content-or-context-moderation/

Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E. and Herrera, F. (2011) “Science mapping software tools: Review, analysis, and cooperative study among tools”. *Journal of the American Society for Information Science and Technology,* Vol 62, no. 7, pp 1382–1402. https://doi.org/10.1002/asi.21525

Debre, I. and Akram F. (2021) “Facebook's language gaps weaken screening of hate, terrorism”. *The Associated Press*, [online], 26 October, https://apnews.com/article/the-facebook-papers-language-moderation-problems-392cb2d065f81980713f37384d07e61f

Ferrara, E., Chang, H., Chen, E., Muric, G. and Patel, J. (2020) “Characterizing social media manipulation in the 2020 US presidential election”, *First Monday*, [online], https://doi.org/10.5210/fm.v25i11.11431

Lim, W. M. and Kumar, S. (2024) "Guidelines for interpreting the results of bibliometric analysis: A sensemaking approach", *Global Business and Organizational Excellence*, Vol 43, No. 2, pp. 17–26. https://doi.org/10.1002/joe.22229

Liu, Y., Yildirim, P. and Zhang, Z. (2022) "Implications of Revenue Models and Technology for Content Moderation Strategies", *Marketing Science, Vol* 41, No. 4, pp. 403–419. https://doi.org/10.1287/mksc.2022.1361

Michael Hall, C. (2011) "Publish and perish? Bibliometric analysis, journal ranking and the assessment of research quality in tourism", *Tourism Management*, Vol 32, No. 1, pp. 16–27. https://doi.org/10.1016/j.tourman.2010.07.001

MSNBC (2021) "Why Content Moderation Costs Social Media Companies Billions". *MSNBC*, [online], 27 February, https://www.youtube.com/watch?v=OBZoVpmbwPk (Accessed: 9 February 2024)

BMJ, 2021;372:n71 https://doi.org/10.1136/bmj.n71

Veglis, A. (2014) *Moderation Techniques for Social Media Content in* Meiselwitz, G. eds. *Social Computing and Social Media*. Cham: Springer International Publishing, pp. 137–148.

Wike, R., Silver, L., Fetterolf, J., Huang, C., Austin, S., Clancy, L. and Gubbala, S. (2022) "Social Media Seen as Mostly Good for Democracy Across Many Nations, But US is a Major Outlier", *Pew Research Center*, [online], 22 January, 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. and Čater, T. (2015) "Bibliometric Methods in Management and Organization", Organizational Research Methods, Vol 18, No. 3, pp. 429–472. https://doi.org/10.1177/1094428114562629