

SOCIAL MEDIA CONTENT MODERATION: USER-MODERATOR COLLABORATION  
AND PERCEPTION BIASES

by

Kanlun Wang

A dissertation submitted to the faculty of  
The University of North Carolina at Charlotte  
in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy in  
Computing and Information Systems

Charlotte

2024

Approved by:

---

Dr. Lina Zhou

---

Dr. Dongsong Zhang

---

Dr. SungJune Park

---

Dr. Depeng Xu

---

Dr. Shi Chen



# ABSTRACT

KANLUN WANG. Social Media Content Moderation: User-Moderator Collaboration and Perception Biases (Under the direction of DR. LINA ZHOU)

Social media has emerged as a common platform for knowledge sharing and exchange in online communities. However, it has also become a hotbed for the diffusion of irregular content. Content moderation is crucial for maintaining a safe and healthy online environment by regulating the distribution of user-generated content (UGC).

Engaging users in content moderation fosters a sense of shared responsibility and empowers them to actively shape the environment of online communities. Leveraging the expertise of moderators leads to a deeper contextual understanding of content, thereby improving the overall consistency and legitimacy of content moderation in compliance with community or platform guidelines. Nevertheless, the collaborative effort of a more inclusive moderation process remains unexplored by previous studies. While there is increasing attention to fairness, transparency, and ethics in content moderation, prior research often assesses content moderation perceptions of users and moderators in isolation, resulting in a lack of comprehensive perceptual understanding of content moderation decision-making.

To address these limitations, this research proposes UMCollab, a user-moderator collaborative content moderation framework that incorporates the dynamics of user engagement and the domain knowledge of moderators into deep learning models to facilitate content moderation decision-making. Additionally, this research empirically investigates user perceptions of content moderation from the perspectives of review information comprehensiveness, user roles, and content familiarity.

UMCollab leverages graph learning to model user engagement, which is further enhanced by the creditability and stance of users' online discussions. It also employs attention mechanisms

to learn moderators' domain knowledge through their decisions on UGC in accordance with online community rules. Moreover, this research conducts an online experiment with participants with diverse backgrounds and roles regarding online engagement to complete a series of content moderation tasks and evaluate their perceptions of content moderation.

The findings of this dissertation hold significant potential for enhancing the effectiveness, fairness, transparency, and sense of community ownership in moderating UGC in social media. By providing theoretical, methodological, and technical contributions to content moderation, the research aims to improve the safety and success of online communities.

**Keywords:**

Content moderation, perception biases, user engagement, domain knowledge, deep learning, social media

## **ACKNOWLEDGMENTS**

This work is partially supported by a Truist Research Grant, a Graduate School Summer Fellowship, and a School of Data Science Summer Seed Grant. Any opinions, findings, or recommendations expressed here are those of the authors and are not necessarily those of the sponsors of this research.

## DEDICATION

First, I would like to express my sincere gratitude to Dr. Lina Zhou, the chair of my dissertation committee, for her continuous support, patience, motivation, and extensive mentorship throughout my dissertation research along with other research projects. Her guidance was invaluable during my Ph.D. journey, and I am deeply appreciative of the opportunity to work under her supervision, which led to the successful completion of multiple research projects and numerous noteworthy publications in interdisciplinary fields.

In addition to my advisor, I extend my heartfelt thanks to Dr. Dongsong Zhang for his co-mentorship during my Ph.D. studies. His expertise and dedication to research have broadened my perspective and expanded the scope of my work from various angles. I also deeply thank the other members of my dissertation committee, including Dr. SungJune Park, Dr. Depeng Xu, and Dr. Shi Chen, for their constructive suggestions and the time they dedicated to serving on my committee.

Moreover, I would like to extend a special thank you to my family, especially my wife, Yunzhou Zhu, and my lovely son, Jason Wang, for their unwavering emotional and mental support throughout my Ph.D. journey.

My appreciation also goes to my peers, including Zhe Fu, Wangjiaxuan Xin, Fei Peng, Jaewan Lim, Zhihui Liu, Abdulrahman Aldkheel, and Sisi Yuan for their support and research collaborations at UNC Charlotte.

Lastly, I would like to extend my special appreciation to those who have shared their friendship and provided continuous emotional support throughout my Ph.D. studies at Charlotte, including Pengfei Shi, Shanshan Yu, Yuchen Shi, and Yiru Luo.

## TABLE OF CONTENTS

LIST OF TABLES	xi
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS	xiv
<b>CHAPTER 1: INTRODUCTION</b>	<b>1</b>
1.1 Research Statement	1
1.2 Research Questions	6
1.3 Dissertation Roadmap	7
<b>CHAPTER 2: BACKGROUND AND LITERATURE REVIEW</b>	<b>9</b>
2.1 The Role of Content Moderation in Social Media	9
2.2 User Engagement in Social Media	10
2.3 Phase-based Categorization of Content Moderation	13
2.4 Human-based Content Moderation	16
2.4.1 Moderator-based Moderation	17
2.4.2 User-based Moderation	18
2.5 Automated Content Moderation	19
2.5.1 Matching/Hashing-based Approaches	20
2.5.2 Machine Learning-based Models	21
2.6 Issues with Content Moderation in Social Media	27
<b>CHAPTER 3: CHARACTERISTICS AND EFFICIENCY OF USER ENGAGEMENT IN CONTENT MODERATION</b>	<b>33</b>
3.1 Introduction	33
3.2 Related Work	35
3.3 A RoBERTa-based Framework	36

3.4 Experiments	38
3.4.1 Data Collection and Preparation	38
3.4.2 Baseline Models and Variant User Engagement Models	40
3.4.3 Performance Evaluation	41
3.5 Results	41
3.5.1 The Characteristics of User Engagement in Content Moderation	41
3.5.2 The Degree of User Engagement in Content Moderation	42
3.5.3 The Efficiency of User Engagement in Content Moderation	42
3.6 Discussion	45
<b>CHAPTER 4: PERCEIVED BIASES IN CONTENT MODERATION</b>	<b>47</b>
4.1 Introduction	47
4.2 Related Work	49
4.3 Theoretical Foundation and Hypotheses Development	51
4.3.1 Review Information Comprehensiveness	51
4.3.2 User Roles	55
4.3.3 Content Familiarity	58
4.4 Experiment Design	61
4.4.1 Procedure and Tasks	61
4.4.2 Variables and Measurements	65
4.4.3 Data Collection and Preparation	67
4.4.4 Participants	69
4.4.5 Data Analyses	69
4.5 Results	70



4.5.1 Content Review Efforts	70
4.5.2 Perceived Moderation Decision	73
4.6 Discussion	79
<b>CHAPTER 5: USER-MODERATOR COLLABORATIVE (UMCollab) CONTENT MODERATION</b>	<b>83</b>
5.1 Introduction	83
5.2 Related Work	85
5.3 Theoretical Foundations of User-Moderator Collaboration	87
5.4 The UMCollab Framework	88
5.4.1 User Engagement	89
5.4.2 Moderators' Domain Knowledge	91
5.4.3 Prediction of Moderation Decisions	93
5.5 Experiments	94
5.5.1 Data Collection and Preprocessing	94
5.5.2 Baseline Models	94
5.5.3 Performance Evaluation	95
5.6 Results	96
5.6.1 Model Performances	96
5.6.2 Ablation Study	97
5.7 Discussion	100
<b>CHAPTER 6: CONCLUSIONS</b>	<b>102</b>
6.1 Summary	102
6.2 Research Contributions	103
6.3 Research Implications	105

6.4 Practical Implications	106
6.5 Limitations and Future Work	107
REFERENCES	111
APPENDIX A: IRB APPROVAL	124
APPENDIX B: PRE-SCREENING SURVEYS	125
APPENDIX C: USER STUDY CONSENT FORM	136
APPENDIX D: PRE-EXPERIMENT SURVEYS	140
APPENDIX E: POST-REVIEW SURVEY	160
APPENDIX F: SUPPLEMENTARY TABLES	163

## LIST OF TABLES

Table 1: Feature Engineering-based Models for Content Moderation	22
Table 2: Representation Learning-based Models for Content Moderation	25
Table 3: Descriptive Statistics of Content Review Efforts	72
Table 4: The Results of Repeated ANOVA of Content Review Efforts	72
Table 5: Multiple-Comparison Results for Content Review Efforts across Review Information Comprehensiveness	72
Table 6: Multiple-Comparison Results for Content Review Efforts for Each User Role across Review Information Comprehensiveness	73
Table 7: Multiple-Comparison Results for Content Review Efforts for Each Content Familiarity across Review Information Comprehensiveness	73
Table 8: Descriptive Statistics of Perceived Moderation Decision	77
Table 9: The Results of Repeated ANOVA of Perceived Moderation Decision	77
Table 10: Multiple-Comparison Results for Perceived Moderation Decision across Review Information Comprehensiveness	78
Table 11: Multiple-Comparison Results for Perceived Moderation Decision for Each User Role across Different Review Information Comprehensiveness	78
Table 12: Multiple-Comparison Results for Perceived Moderation Decision for Each Content Familiarity across Different Review Information Comprehensiveness	78
Table 13: A Summary of Hypotheses Testing Results	79
Table 14: Performance Comparisons between the UMCollab and the Baseline Models	98
Table 15: Performance Comparisons between the UMCollab Model and Ablated Models	99
Table 16: Multiple Comparisons of Performance Deterioration of the Ablated Models	99
Table 17: Selected UGC for the User Study	163

Table 18: Selected Subreddits for the Four Domains	170
Table 19: The Results of the Homogeneity Test for the User Study	170
Table 20: Community Rule and Post Examples	170

## LIST OF FIGURES

Figure 1: Dissertation Roadmap	7
Figure 2: Phase-based Categorization of Content Moderation	14
Figure 3: User Engagement Patterns in Different Content Moderation Policies	30
Figure 4: The RoBERTa-based Framework for Content Moderation	37
Figure 5: The Distribution of User Comments among the Collected Posts	40
Figure 6: Performance Comparisons among Different Characteristics and Degrees of User Engagement	44
Figure 7: Efficiency of Content Moderation across Different Degrees of User Engagement	45
Figure 8: The Research Model	51
Figure 9: The Procedure of the User Study	62
Figure 10: Community Rules Review	64
Figure 11: Content Review	65
Figure 12: The Framework of UMCollab for Content Moderation	88
Figure 13: Model Performances of the UMCollab and Baseline Models	97
Figure 14: Performances Comparisons of the Ablated Models vs. the UMCollab Model	98

## LIST OF ABBREVIATIONS

ANN: Artificial Neural Network

API: Application Programming Interface

AUC: Area Under the receiver operating characteristic Curve

BERT: Bidirectional Encoder Representations from Transformers

BiLSTM: Bidirectional Long Short-term Memory

BiRNN: Bidirectional Recurrent Neural Networks

CBOW: Continuous Bag of Words

CNN: Convolutional Neural Networks

DADM: Discrimination-aware Data Mining

DeepNet: Deep Hybrid Neural Network

EFN: Emotion-based Fake News (Detection framework)

EFR: Error Finding Rate

FBM: Fundacion Barcelona Media

FNR: False Negative Rate

FPR: False Positive Rate

GCN: Graph Convolutional Network

GRU: Gated Recurrent Unit

HIT: The total number of work requests MTurk workers have attempted

LG: Logistic Regression

LIBSVM: A Library for Support Vector Machines

LIWC: Linguistic Inquiry and Word Count

LSTM: Long Short-term Memory

MIL: Multiple Instance Learning

MTurk: Amazon Mechanical Turk

NLP: Natural Language Processing

RNN: Recurrent Neural Network

RoBERTa: Robustly Optimized BERT Approach

SCP: Santa Clara Principles

SVM: Support Vector Machines

TF-IDF: Term Frequency - Inverse Document Frequency

TNT: Text Normalization based pre-training of Transformers

UGC: User-generated Content

## CHAPTER 1: INTRODUCTION

### 1.1 Research Statement

UGC encompasses various forms, such as text, images, videos, reviews, or testimonials, which users create and share on social media platforms. This content serves as a potent and valuable resource for both platforms and online users themselves. Ultimately, UGC fosters active user engagement, and in turn, stimulates higher purchase behaviors [1]. Social media platforms have emerged as widely adopted channels for the extensive dissemination of UGC, enabling users to share knowledge and experience on a large-scale online community. However, the inherent openness of these platforms also facilitates the spread of irregular content, such as unsubstantiated or false content. If left unmoderated, its dissemination on social media can endanger the well-being and trustworthy online community [2].

Most social media platforms have adopted intervention strategies for irregular content by incorporating governance mechanisms, which is “structure participation in a community to facilitate cooperation and prevent abuse” [3]. A typical intervention strategy for regulating UGC on social media platforms is content moderation, which is a process ensuring that UGC complies with the platforms’ policies and community standards [4]. According to a report by the Global Internet Forum to Counter Terrorism [5], the deployment of content moderation flags 98% of the videos removed from YouTube, and 93% of the removals are linked to accounts flagged by internal, proprietary spam-fighting tools on Twitter. Additionally, 99% of the Islamic State of Iraq and the Levant and Islamic State of Iraq and Syria and Al Qaeda-related terror content removed from Meta is detected preemptively before anyone from the community flags it, and sometimes even before it becomes visible on the platform. Through content moderation interventions, news organizations can influence the deliberative behavior of commenters [6], thereby impacting the



types of expressed comments (e.g., thoughtful or thoughtless) [7], as well as users' perceptions of the content they are commenting on [8]. The primary impact of content moderation on users is felt through a spectrum of punishments, which can range from content removal to the suspension of user accounts [9]. It is worth noting that this dissertation research focuses on content removal.

The approaches to content moderation can be broadly categorized into two types: *human moderation* and *automated moderation* [4]. Human moderation involves voluntary users [10], [11], [12] or commercially trained flaggers [13] hired by social media companies who manually flag or review users' content. Automated moderation harnesses advanced AI-based techniques, such as utilizing keyword blacklists to compare against UGC, to ensure efficient content moderation [4], [14], [15], [16]. This approach offers superior scalability compared to a manual review of UGC. Based on the specific needs and available resources of social media platforms, content moderation can be intervened at three different phases, each corresponding to different time frames of deployment. For instance, *pre/proactive-moderation* [17], [18], [19], [20], [21] reviews and approves UGC by moderators or systems before it is published on the platform, and it ensures that only appropriate content is visible to the public; *post-moderation* [22], [23], [24], [25] allows UGC to be published immediately, but moderators review and remove any inappropriate or rule-violating UGC after it has been posted. This approach allows for faster UGC delivery but may result in some inappropriate UGC being temporarily visible; and *reactive moderation* [26], [27], [27], [28] relies on user reports or flags to identify and review potentially problematic UGC. It takes action based on these reports, such as removing or addressing the reported UGC. Among different strategies for content moderation, this study focuses on *post moderation*, which has been widely adopted by most social media platforms [29].

Prior empirical research on content moderation has predominantly focused on users' perceptions, highlighting that these views are influenced by personal experiences and the transparency of the moderation process, with trust and fairness being key concerns (e.g., [30], [31]). Moreover, moderators face significant psychological stress [32] and encounter challenges related to the accuracy of AI systems [33], emphasizing the need for effective training and support mechanisms [34]. Policymakers play a crucial role in shaping content moderation through legislation, aiming to balance user protection with freedom of expression and continually updating standards to reflect evolving societal norms [35]. However, none of these studies has undertaken empirical investigations into how user perceptions of content moderation vary in content moderation. In particular, first, review information comprehensiveness promotes inclusivity and fairness by incorporating multiple perspectives of UGC and providing a comprehensive evaluation through a richer context. Second, different user roles offer unique insights from various user perspectives and foster collaboration and thorough content evaluation. Third, content familiarity enhances decision-making by leveraging pre-existing knowledge and encourages proactive community self-regulation. Nevertheless, these three aspects remain unexplored in prior content moderation research.

Effective content moderation is the key to boosting user engagement in online communities. When users feel secure and supported in a community, they're more likely to participate in positive interactions [36], [37]. User engagement in social media involves how much users interact with each other on the platform, including creating and sharing UGC, commenting, liking, and following others [38]. This engagement is valuable for content moderation because it allows moderators to gather feedback on objectionable content. User engagement serves as both a process and an outcome of interactions [39], providing significant opportunities to gain insights

from UGC, like their interactions and comments. However, the moderation decisions of users may be influenced by their personal biases and subjective perspectives, resulting in inconsistent application of content guidelines and potential favoritism towards specific viewpoints [40]. Users may not have the expertise or training to accurately identify certain types of problematic UGC. This could result in the removal of valid UGC or the allowance of harmful UGC to stay online [41]. Hence, devising an efficient framework that seamlessly incorporates diverse user engagement into content moderation poses a challenging yet highly valuable research endeavor. In this study, I denote *user engagement* as referring to active participation in the process of evaluating, flagging, and providing feedback on UGC within an online platform or community.

Moderators often have an extensive history of engagement within an online community. In general, moderators undergo training before being empowered to intervene in UGC [31]. Through their experience and training, moderators acquire domain knowledge and gain a deeper understanding of the topics and issues discussed in the UGC. This expertise enables them to interpret context, identify subtleties, and make more contextually appropriate decisions when assessing whether UGC adheres to guidelines. In addition, moderators frequently participate in shaping the community rules and standard operating procedures [42], [43]. These guidelines establish a framework for making decisions consistently and help ensure uniformity in their approach. Nevertheless, moderators may inadvertently reinforce existing biases and inadvertently create echo chambers where only certain opinions are tolerated, stifling diversity of thought. Moderators with expertise in a particular domain may be more likely to identify content that aligns with their own beliefs or background knowledge, leading to a bias in their moderation decisions and subsequently providing unfair treatment of UGC that challenges established beliefs or deviates from mainstream opinions. Thus, there is a notable gap in current research regarding the effective

alignment of moderators' domain expertise with UGC in the realm of content moderation. In this research, I denote *domain knowledge* as moderators' historical content moderation decisions made to UGC according to community rules and policies.

Given the technical advances in content moderation, deep learning-based approaches have shown compelling evidence of their efficacy. Those methods encompass a wide range of techniques to learn the representation of UGC, including classic vectorization techniques (e.g., paragraph2vec [44], word/character n-grams [45]) to word embeddings (e.g., GloVe [46], FastText [47], [48], GRU [49], [50], RNN [49], FastText [47], [48]), or to directly achieve the text classification task by leveraging pre-trained models (e.g., BERT [51] and RoBERTa [52]). Moreover, prior studies (e.g., [49], [50], [53], [54], [55], [56]) have attempted to incorporate user engagement for content moderation. They focused on the history of users' discussions to predict misinformation in videos [53] or news articles [49]. Some others focused on leveraging users' profiles or sources of news [54] or constructing users' social networks [55], [56] in the context of content moderation. However, to the best of our knowledge, none of the previous studies has considered a collaborative approach that integrates the dynamics of user engagement and moderators' domain knowledge in their model development. This can be due to the following notable challenges and limitations: 1) there is a lack of publicly available information about the moderators or individuals involved in particular content moderation interventions. This is primarily due to privacy and public relations-related concerns [57]; 2) user engagement and moderators' domain knowledge are dynamic given that online communities shift interests or adapt to new trends and technologies and that ongoing learning process enables moderators to make more informed and contextually appropriate moderation decisions; 3) user engagement in social media has different characteristics, such as user discussions (e.g., the interactive discussions

among online users within a specific post), temporality (i.e., the time of user engagement or discussion involvement), creditability (i.e., the quality of a comment or a post assessed by online users), and orientation (i.e., the orientation of a user comment directly responding to the original post or other relevant users' discussions). The characteristics of user engagement in moderating social media content are not yet fully comprehended; and 4) prior research works have not sufficiently addressed the effects of user engagement on moderating UGC. This is particularly notable given the limited understanding of how structure-based insights from user engagement and the domain knowledge of moderators can be leveraged to enhance the effectiveness of content moderation.

## **1.2 Research Questions**

To fill the aforementioned research gaps, this study aims to answer the following research questions:

RQ 1: What is the relationship between user engagement and the effectiveness of content moderation?

RQ 1.1: What major characteristics of user engagement impact the effectiveness of content moderation?

RQ 1.2: To what extent does the degree of user engagement impact the effectiveness of content moderation?

RQ 1.3: How can the efficiency of user engagement in content moderation be improved without compromising its effectiveness?

RQ 2: What factors impact user perceptions in content moderation?

RQ 2.1: How does review information comprehensiveness impact perceptions of content moderation?

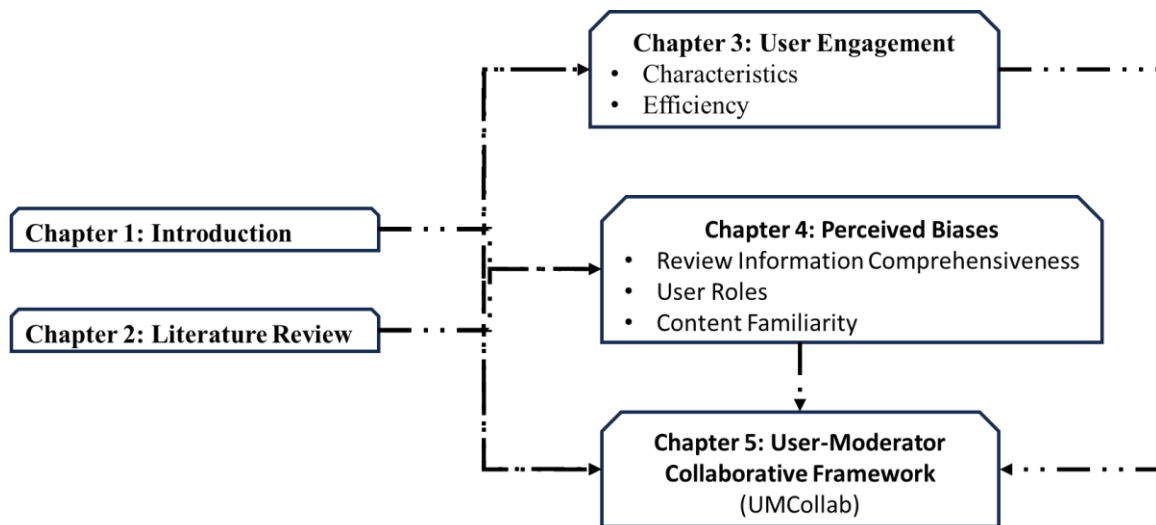
RQ 2.2: How do user roles impact perceptions of content moderation?

RQ 2.3: How does content familiarity impact perceptions of content moderation?

RQ 3: Can the effectiveness of content moderation be improved by collaboratively integrating user engagement and domain knowledge into a deep learning-based framework?

RQ 3.1: How does each component of the collaborative framework contribute to the effectiveness of content moderation?

### 1.3 Dissertation Roadmap



**Figure 1: Dissertation Roadmap**

The roadmap of the dissertation is articulated in Figure 1. In Chapter 1, I elaborate on the background and motivations for this dissertation research, as well as the identified research gaps and objectives. In Chapter 2, I discuss related work in relation to content moderation, including different content moderation categorizations, machine/deep learning-based models for content moderation, and issues associated with content moderation in social media. In Chapter 3, I explore the impact of user engagement characteristics and efficiency on content moderation, establishing

a fundamental understanding of the significance of user engagement in this context. In Chapter 4, I conduct an online user study to investigate user perceptions of the content moderation decision-making process, particularly focusing on review information comprehensiveness, user roles, and content familiarity. In Chapter 5, I introduce a novel framework for content moderation that leverages cutting-edge deep learning techniques to integrate user engagement and moderators' domain knowledge.

## CHAPTER 2: BACKGROUND AND LITERATURE REVIEW

This chapter provides an in-depth discussion of the role of content moderation and user engagement in social media. It also explores various categories of content moderation, with a particular focus on a range of machine/deep learning-based models for content moderation.

### 2.1 The Role of Content Moderation in Social Media

Content moderation plays a crucial role in monitoring UGC on social media platforms to ensure UGC compliance with community guidelines, legal regulations, and ethical standards, while also allowing for diverse opinions and viewpoints to be expressed [4]. The scale of content moderation on social media is immense, given the vast amount of UGC generated every second on the internet. Several factors contribute to the scale of content moderation [33], including, but not limited to, *volume* - millions of posts, comments, images, and videos are uploaded and shared every minute, making it challenging to review and moderate all UGC effectively [58]; *user diversity* - online platforms have a global reach, attracting users from all over the world. This diversity of users means that content moderation needs to be performed in multiple languages and consider cultural nuances and sensitivities specific to different regions [59], [60]; *format diversity* – as technology advances, new content formats and channels emerge, the format of UGC commonly extends beyond text-based content. Platforms must stay up-to-date with the latest trends and adapt their processes to handle emerging content types, such as images, videos, audio, and other forms of multimedia. Each format comes with its own set of challenges and requires specialized tools and techniques for moderation [61]; *moderation efficiency* - many platforms aim to provide real-time experience for their users, which requires UGC to be reviewed and moderated quickly and efficiently. This adds pressure to content moderation teams to identify and remove inappropriate or harmful UGC promptly [62]; *complexity* - the range of UGC categories that



require moderation is extensive, including, but not limited to misinformation, hate speech, harassment, nudity, violence, terrorism, self-harm, copyright infringement, and illegal activities. Each category requires a different approach and expertise in understanding the context and intent behind the UGC [60]. By looking into misinformation alone, it includes the unintentional/intentional spread of rumors, urban legends, fake news, etc. [63]. The openness and timeliness of social media platforms make them a hotbed for the creation and dissemination of misinformation. An alarming phenomenon of misinformation persisted with COVID-19 since the onset of the pandemic [64]. In addition, many studies have argued that false stories played an important role in political campaigns, especially during the 2016 presidential election and continued through the 2020 presidential election [65]; and *moderation strategies* - given the scale and complexity of content moderation, online platforms typically employ a combination of automated tools, machine learning algorithms, and human moderators to review and moderate UGC. The goal is to strike a balance between fostering a safe and inclusive online environment while respecting users' freedom of expression and avoiding over-censorship [4].

## **2.2 User Engagement in Social Media**

Engaging users in social media is essential for gathering data, understanding users' preferences, evaluating performance, optimizing content strategies, managing reputation, and gaining a competitive edge in the digital space [66]. User engagement generates valuable data that can be analyzed to gain insights into user behavior, preferences, and interests. By analyzing user engagement data, the derived knowledge enables social media platforms to tailor their UGC and communication strategies to better engage target users [67]. In addition, user engagement metrics, such as content credibility, likes, comments, shares, and click-through rates, allow us to assess the effectiveness of social media campaigns. By tracking engagement levels over time, we can

evaluate the impact of UGC and then determine which strategies are working and make data-driven decisions to optimize future efforts [68]. Evaluating engagement levels in relation to our marketing objectives can measure the effectiveness of social media campaigns and consequently allocate resources [69]. Most social media platforms have focused on leveraging content moderation to improve user engagement in order to provide a positive, supportive, and inclusive online environment that encourages users to participate, engage, and connect in meaningful ways [36], [37]. In particular, content moderation can facilitate a positive user experience and encourage diverse and respectful conversations by monitoring and removing inappropriate, offensive, or spammy content, which in turn contributes to the creditability and trustworthiness of social media platforms [15].

Engaging users in content moderation outlines users' rights and helps build a stronger and more vibrant online community [15]. Instead of relying solely on platform moderators, engaging users in content moderation allows for distributed moderation efforts by harnessing the collective vigilance of the community. The distributed workload of flagging and reporting inappropriate content helps identify and address violations more efficiently. This shared responsibility lightens the moderation workload and enables a more proactive approach to content moderation [28]. In addition, involving users in content moderation empowers the community and gives users a sense of ownership, so that users become active participants in shaping the platform and maintaining its integrity [70]. User engagement in content moderation brings diverse perspectives and contextual understanding to the process, which can provide insights and cultural nuances that might be missed by platform moderators alone. The collaborative approach improves the accuracy of content moderation decisions and helps prevent potential biases or misunderstandings [71]. By involving users in the moderation process, platforms demonstrate a commitment to openness and fairness.

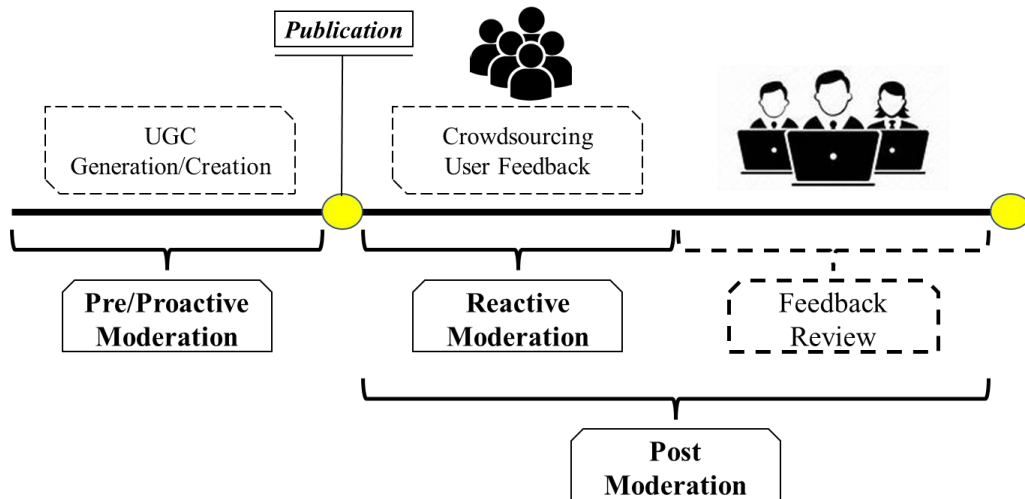
Users can witness firsthand how content moderation decisions are made, which helps foster trust and confidence in the platform's practices [72]. Furthermore, users can continuously provide feedback on moderation guidelines, reporting mechanisms, and overall user experience. This feedback loop enables platforms to refine their content moderation strategies, adapt to evolving challenges, and better align with the needs and expectations of the community [73].

However, there are still several unresolved issues related to user engagement in content moderation. When users are involved in content moderation, there is a risk of inconsistent enforcement of guidelines and standards. Different users may have varying interpretations of what is considered inappropriate or offensive UGC. This inconsistency can lead to confusion and frustration in the community and undermine the platform's credibility [40]. In addition, users may have personal preferences or biases that influence their judgment when moderating UGC. This can result in uneven treatment of similar types of UGC or unfair targeting of certain viewpoints. It requires careful oversight and clear guidelines to minimize the impact of biases [74]. Users may not possess the same level of expertise or understanding as dedicated platform moderators. They may struggle to identify nuanced or subtle violations, especially in complex or sensitive topics. This limitation can result in both false positives (removing UGC that does not actually violate community/platform guidelines) and false negatives (allowing inappropriate UGC to remain available in online communities) [41]. Furthermore, engaging users in content moderation opens up the possibility of abuse or manipulation [75]. Malicious users may exploit the reporting system to target and harass others, falsely flagging UGC as inappropriate. This can lead to censorship of legitimate content and create a hostile environment within the community. When users are involved in content moderation, platforms may face legal and liability challenges. If users make incorrect moderation decisions that result in harm or infringement of rights, the

platform may still be held responsible. Balancing user engagement while ensuring legal compliance and mitigating potential risks can be a complex task for platforms [76]. Engaging users in content moderation can have an emotional toll on those involved. Moderating disturbing, offensive, or graphic UGC can be mentally challenging and impact the well-being of users tasked with this responsibility [77], [78]. It is essential to prioritize user support, mental health resources, and clear guidelines to mitigate these risks. Thus, it is crucial for platforms to strike a balance between user engagement and professional moderation to address the disadvantages associated with user engagement. One potential solution to the problem is the design for contestability, whereby users can shape and influence the decision-making process in content moderation [79].

### **2.3 Phase-based Categorization of Content Moderation**

Content moderation encompasses a multifaceted and methodical decision-making procedure that incorporates the involvement of human moderators and the implementation of algorithms [80]. The interventions of content moderation can also be applied across various time frames, depending on the platform's requirements. By moderating UGC at various time intervals, platforms can increase the likelihood of detecting and addressing harmful or inappropriate UGC that may have been missed during previous moderation phases. This helps maintain a safer and more responsible online environment for users. Additionally, content moderation at different phases can also help account for varying user activity patterns and regional time differences, ensuring that moderation efforts are effective across different time zones and user demographics. In this section, each component will be explained in detail, providing a comprehensive understanding of its role and functionality.



**Figure 2: Phase-based Categorization of Content Moderation**

Based on the timing of content moderation interventions, it can be classified into three phases (see Figure 2):

- 1) *Pre/proactive moderation* [17], [18], [19], [20], [21] focuses on preventing the display or dissemination of problematic UGC and involves reviewing all UGC by human moderators and automated tools before it is posted or visible to other users, and it is typically used in more sensitive or high-risk areas, such as online forums for children or political discussion groups. Pre/proactive moderation ensures a high level of control over the UGC and enables platforms to curtail its potential to gain traction and reduce its impact on users and public discourse. By employing pre-screening measures, platforms can ensure that the majority of UGC visible to users adheres to community standards, fostering a more positive user experience. Moreover, pre/proactive moderation alleviates the burden on human moderators, particularly in platforms with large volumes of UGC. Automated systems can filter out a significant portion of problematic UGC, allowing human moderators to focus on more nuanced cases that require human judgment. However, pre/proactive moderation may sometimes incorrectly flag or block UGC that is actually harmless or permissible. This

can result in false positives, where UGC is mistakenly identified as problematic and restricted or removed. In addition, over-blocking can limit users' ability to express themselves freely and impede legitimate discussions or UGC sharing. Pre/proactive moderation systems may struggle to accurately interpret the context, intent, and cultural nuances of UGC. They often rely on patterns and keywords, which can lead to misinterpretations and incorrect enforcement of guidelines. This limitation may result in the removal of UGC that is intended as satire, parody, or harmless humor. In addition, pre/proactive moderation can be resource-intensive and time-consuming. Manually reviewing all UGC before it becomes visible can introduce significant delays in UGC publication, which may impact real-time or time-sensitive interactions. It can also require a large team of human moderators to handle the volume of UGC, resulting in higher operational costs.

- 2) *Post moderation* [22], [23], [24], [25] involves moderating UGC after it has been published. This method can be less time-intensive than pre/proactive moderation, yet can still help prevent harmful UGC from being visible on a platform. Post moderation allows users to publish their UGC immediately without delays or pre-approval, which effectively motivates users to engage in real-time conversations and share a broader spectrum of perspectives on the platform. In addition, post moderation creates an opportunity that enables users to provide immediate feedback on UGC through reporting mechanisms. Yet, with post moderation, users may come across objectionable or inappropriate UGC before it is reviewed and action is taken.
- 3) *Reactive moderation* [26], [27], [27], [28] involves reviewing UGC only after it has been flagged or reported by other users. The platform relies on user reports to identify potentially

harmful content first and then takes action(s) if the UGC violates community standards or terms of service. Reactive moderation allows users to express themselves and share their thoughts and opinions more freely without excessive pre-screening, which fosters a sense of freedom and encourages active participation. In addition, it allows for a more detailed and nuanced assessment of each reported piece of UGC and helps understand the intricacies of human communication, sarcasm, and subtle nuances that may be challenging for automated systems to grasp accurately. This personalized evaluation can lead to fairer decisions and avoid false positives that automated systems may generate. Nevertheless, reactive moderation makes problematic or rule-breaking UGC remain visible to users for a period of time before it is addressed. This delay in response can allow harmful UGC to spread, which in turn may have a negative impact on user experience, as users may encounter offensive UGC that goes against the platform's guidelines or policies. If the volume of UGC is high, it can be challenging for moderators to keep up with the influx of reports and UGC that require attention. This can result in slower response times and decreased effectiveness in addressing problematic content. In addition, this approach may not catch all instances of problematic UGC, as some users may not report it or may not recognize certain types of violations. Relying solely on user reports can create blind spots and limit the platform's ability to proactively address emerging issues.

## **2.4 Human-based Content Moderation**

Human moderators, as digital gatekeepers, are critical to guarding social media platforms with a decent digital presence. It is a traditional approach to regulating online behavior and UGC and offers several advantages over automated moderation systems. For example, human moderators can understand the intent behind a piece of UGC by considering factors that may not

be easily detected by automated systems and can adapt quickly to evolving trends, new forms of UGC, and emerging issues, ensuring that platforms can address emerging challenges effectively. In contrast, the implementation of human moderation can be time-consuming and resource-intensive, particularly for platforms with a large volume of UGC. Moderators may also face challenges in dealing with graphic or disturbing UGC, which can have a psychological impact on them. Moreover, human moderators may have biases or subjective viewpoints that can influence their decision-making process, necessitating ongoing training and oversight to maintain consistency and fairness. Human content moderation takes two sub-forms, including moderator- and user-based content moderation [4], [57].

#### **2.4.1 Moderator-based Moderation**

Moderator-based moderation refers to the practice of entrusting the responsibility of moderating and regulating UGC to a central authority or platform [13]. In this approach, a single entity, such as a host of human efforts from contractors (e.g., third-party moderation services) or power users (e.g., a group of high-reputation online users regulating an online community) manually reviews UGC [81]. With moderator-based moderation, there is a higher likelihood of consistent enforcement of community guidelines and policies across the platform, by allowing for specialized training and expertise to be concentrated within a single team so that moderators can receive comprehensive training on community guidelines, legal considerations, and emerging trends to enhance their ability to make informed decisions [30]. This helps establish clear standards for user behavior and fosters a more predictable and trustworthy user experience. In addition, having a moderator-based moderation allows for a streamlined and efficient content review process, which enables the platform to dedicate resources, such as a dedicated team of moderators, advanced tools, and standardized procedures, to handle content moderation effectively



[41]. However, moderator-based moderation also comes with certain disadvantages and challenges. For instance, Roberts 2016 [81] stated that “workers are dispersed globally, and the work is almost always done in secret for low wages by relatively low-status workers, who must review, day in and day out, digital content that may be pornographic, violent, disturbing, or disgusting.” Moderator-based moderation can be susceptible to censorship and biases. When moderation decisions are concentrated in the hands of a few individuals, there is a risk of subjective judgments and the suppression of certain viewpoints, even unintentionally [15]. Moreover, moderator-based moderation may struggle to account for the nuances of different cultural, regional, or linguistic contexts. Policies and guidelines established at a central level might not adequately address the diversity of perspectives and norms across different communities [3]. Moreover, moderation teams may face challenges in handling the sheer volume of UGC, resulting in delays in addressing violations or an inability to adequately moderate all UGC. Furthermore, moderator-based moderation can limit user agency and control over their own experiences. As a result, users may feel disempowered when their content is removed or their actions are restricted without clear explanations or avenues for appeal. Lack of transparency and opportunities for user input in the moderation process can erode trust in the platform [29]. Last but not least, relying on moderator-based moderation creates a single point of failure. If the moderation infrastructure experiences technical issues, downtime, or becomes compromised, it can disrupt the entire platform's content management and safety mechanisms, leaving it vulnerable to abuse or harmful content.

#### **2.4.2 User-based Moderation**

User-based moderation is based on the online users' triage to classify inappropriate UGC via user reporting [10], [11], [12]. This can involve giving power users tools to flag or report

harmful UGC, as well as moderating comments and other types of UGC. By involving the user community, the moderation workload is distributed, enabling platforms to scale their moderation efforts without solely relying on a designated moderation team. Different users may have different cultural backgrounds, experiences, and sensitivities, which can contribute to a more comprehensive approach to content moderation. It helps ensure that content is assessed from a wider range of viewpoints, reducing the potential for bias in the moderation process. Furthermore, involving users in the moderation process can also foster a sense of ownership and empowerment within the platform's community. Nevertheless, the design concept of reporting diverges significantly from the notion of user-constructed reporting and conflicts with the platforms' wish that online users did not have to encounter inappropriate UGC in the first place. With user-based moderation, different power users within the community may interpret platform guidelines differently, leading to inconsistent moderation decisions. This inconsistency can result in a lack of uniformity in content enforcement, potentially leading to confusion and frustration among users. In addition, power users participating in distributed moderation may not have the same level of expertise and training as dedicated moderators. This can lead to a higher likelihood of incorrect flagging or reporting of UGC, resulting in the potential removal of UGC that does not actually violate guidelines. To address these challenges, platforms often strive to incorporate community input, employ diverse moderation teams, offer transparent moderation policies, provide robust appeal mechanisms, and implement AI-based systems that can assist in content analysis and decision-making while mitigating biases.

## **2.5 Automated Content Moderation**

Due to the vast and growing amount of UGC in social media, there is an insufficient number of human moderators to thoroughly examine every new piece of content [82], [83]. As a

result, automated content moderation become an emergent process in the content moderation process as it implements algorithms, machine learning, and artificial intelligence to automatically filter and moderate UGC on social media platforms [4], [14], [15], [84]. Automated content moderation allows platforms to have greater scalability by quickly processing and analyzing a vast amount of content, making it suitable for platforms with high levels of user activity. In addition, automated content moderation can apply real-time responses to UGC and reach high levels of consistency across a platform, which reduces the reliance on human moderators for routine and repetitive tasks, allowing them to focus on more complex cases. This can lead to cost savings and increased operational efficiency for platforms. To keep up with the volume of content created by users, social platforms—like Meta [85], YouTube [86], and Twitter [87]—are known to apply filtering mechanisms to make informative moderation decisions on their platforms. These techniques range from pattern matching to sophisticated machine learning techniques [33], [88]. Given that content moderation may involve multimedia data for classification, this review is focused on text-based UGC techniques.

### **2.5.1 Matching/Hashing-based Approaches**

Pattern matching typically leverages static word and source-ban lists (e.g., abusive language, pornographic sites, bot-generated content, hashes, IP addresses, and formatting restrictions) [53], [89] to compare with UGC, which achieves a high consistency and stability in content moderation, yet performs poorly due to highly contextual UGC in online space [4], [90]. As an example, the exact matching of bad words [91] proves ineffective over time as norms evolve, and users can figure out ways to circumvent the blacklist. On the other hand, over-blocking can be overwhelming in certain instances, as it may flag words that could be acceptable in a particular context. To maintain effectiveness, hashing is a typical technical solution that involves

transforming a known example of content into a unique 'hash' – a data string that serves to identify the original content uniquely [33]. Hashes are advantageous as they are simple to compute and generally have a smaller size compared to the original content. This makes it effortless to compare a given hash against a large table of existing hashes to determine if there's a match. Yet, it is necessary to update static filtering methods in order to adapt to the ever-changing nature of online behaviors [92] [93]. One example of this is the shared hash database for alleged terrorist propaganda that was created by Meta, YouTube, Microsoft, and Twitter [94]. By employing a shared hash database, the filtering mechanism can stay up to date and effectively capture the evolving trends and context-related aspects.

### **2.5.2 Machine Learning-based Models**

The multi-dimensional features of textual UGC or user behaviors are extremely helpful for triaging the violations, therefore providing an intelligent recommendation for moderation decisions [95]. On the basis of the exact properties or general features of UGC, previous research on machine learning-based approaches for content moderation has predominantly emphasized two design artifacts as if they were the definitive and final steps. I categorize existing machine learning-based approaches into two main categories: feature engineering- and representation learning-based approaches. Some exemplary approaches are summarized in Tables 1 and 2 respectively.

#### **2.5.2.1 Feature Engineering-based Models**

Feature engineering-based models for content moderation involve a detailed and iterative process of selecting, transforming, and creating features from UGC to enhance the effectiveness of machine learning algorithms in detecting and managing inappropriate or harmful content. This approach is critical because the quality and relevance of these features directly impact the performance and accuracy of the content moderation system.

**Table 1: Feature Engineering-based Models for Content Moderation**

Models		Context		Platform(s)/ Data Source(s)	Sample Size	Performance
Naïve and SVM [96]	Bayes	Offensive language		YouTube	2M+users from 18 videos	Sentence Level – Precision: 98.24% Recall: 94.34% User Level – Precision: 77.9% Recall: 77.8% AUC: 91.4%
LG and LIBSVM [97]		Toxicity -	hate	Wikipedia	100K	F1: 82%
		Multi-language speech detection		Twitter	795K	
LIBSVM [98]		Harassment detection		FBM (Kongregate, Slashdot, MySpace)	10K+	Precision: 39.4% Recall: 61.9% F1: 48.1%
Random Forest, AdaBoost, and LIBSVM [99]		Hate speech detection		Meta YouTube	142K+	F1: 79%
Decision Tree [89]	Tree	Abusive (flames)	messages	NewtWatch	460	Accuracy: 68%
Naïve [100]	Bayes	Offensive detection	language	NSM Usenet	1,525	Accuracy: 96.72%

M: millions; K: thousands; The best model performances are reported in the table.

For example, Sun and Ni [89] used 47 manually crafted linguistic rules to extract binary feature vectors and employed a decision tree to identify toxic content [89], and Razavi et al., [100] carried out the construction of an abusive language dictionary to extract lexicon-level features for detecting abusive content. Despite their strong generalizability when applied to data from various domains, handcrafted rules and lexicons may struggle to effectively handle implicit human expressions. Prior work has developed a fundamental approach by creating a straightforward classifier that utilizes TF-IDF to build a matrix representing word token frequencies and subsequently trains various classifier(s) to detect toxicity [101], harassment [98], or hate speech [99], [102]. In addition, an alternative approach in feature engineering-based methods [96] is to perform feature selection (e.g., bag of words:  $n$ -grams), where all words in a sentence are treated as features, disregarding their order and grammar. Then, an ablation analysis is conducted to determine the crucial features for content moderation. In relation to  $n$ -gram-based feature utilization, [97] used character trigrams represented by sequences of three characters to develop

input features to fit into a LIBSVM for hate speech detection. Among others, they focused on LIWC's lexicon-derived frequencies as features [53] or handcrafted emotional features [49]. However, these methods frequently grasp surface-level patterns rather than comprehending the underlying semantics and are often prone to errors in spelling, punctuation, and grammar [103].

### **2.5.2.2 Representation Learning-based Models**

Representation learning-based models for content moderation leverage deep learning techniques to automatically learn useful representations or features from UGC. For example, it employed neural networks to acquire surface-level representations by leveraging paragraph2vec [44] for joint modeling of comments and words, and then the CBOW-based distributed representations were utilized to proceed to a logistic regression classifier to identify abusive language [104]. In addition, Pavlopoulos et al. [105] employed three methods to depict a user comment. These methods involved utilizing DETOX [45] and CNN to portray each comment as a collection of word/character n-grams. Additionally, they employed RNN to process the comment tokens. By combining these techniques and employing a classifier, they successfully achieved the detection of abusive UGC. Building upon the foundations of classic deep learning models, researchers have been dedicated to enhancing existing methods. To achieve this, they have introduced novel models such as CNN-GRU [106], BiRNN [106], and BiRNN-attention [106]. These models aim to better learn the representation of UGC.

To initialize the word embeddings of UGC, GloVe [46], BERT [51], RoBERTa [52], RNN [49], FastText [47], [48], GRU [49], [50] have been widely used in content moderation, and followed by a variety of deep learning architectures, such as CNNs [47], LSTM [47], BiLSTM [56], [112], RNN [49], [113], FastText [47], [48] and transformers [107] to learn the contextual information from the text and predict the probability of a moderation decision or the types of

moderation that human moderators should make. In particular, Tan et al. [107] developed a pre-training model by initializing word embeddings using BERT [51] and reconstructing the embeddings via transformers from four operation, including substitution, transposition, deletion, and insertion, to detect hate speech in facilitating content moderation. Similarly, Lai et al. [42] leveraged BERT as a text encoder and used a rationale-style neural architecture to incorporate conditional delegation in content moderation. Furthermore, Badjatiya et al. [47] initialized the word embeddings with either random embeddings or GloVe embeddings [46], followed by CNN, LSTM, and FastText for hate speech detection in the context of content moderation.

Given the advances in deep learning research, there are a number of pre-trained models that are available to use in text classification tasks, such as BERT [51] and RoBERTa [52], the most commonly used pre-trained models, that show superior performance in content moderation. Specifically, prior studies leveraged either BERT [53], [106], [108], [109], [110] or RoBERTa [48], [53], [114] to capture the context-free meaning of UGC and then fine-tuned the parameters with a set of labeled data for content moderation. In addition to these two models, Chandrasekharan et al. [57] and Barbieri et al. [48] leveraged FastText to represent textual information of the comments but for different purposes. One is fine-tuned cross-community learning-based classifiers, one set of classifiers was obtained from 100 popular subreddits, and another set of classifiers was trained based on macro norm violations [57], and the other was fine-tuned for multiple tasks (e.g., emoji prediction, irony detection, hate speech detection, and sentiment analysis) [48]. In the pursuit of creating AI-based detection systems to recognize word perturbations (e.g., “shit”  $\rightarrow$  “sh\_t”, and “nigger”  $\rightarrow$  “ni66er”), previous studies also employed advanced techniques (e.g., Perspective API [115], Baidu [116], Huawei [117]) to produce a wide array of adversarial samples [109], [110].

**Table 2: Representation Learning-based Models for Content Moderation**

Models	Context	Platform(s)/ Data Source(s)	Sample Size	Performance
CBOW/paragraph2vec [104]	Hate speech detection	Yahoo Finance	951K+	AUC: 80.07%
CNN, LSTM, and FastText [47]	Hate speech detection	Twitter	16k	Precision: 93% Recall: 93% F1: 93%
FastText [57]	Comment moderation	Reddit	680	Accuracy: 86% Recall: 87.5%
RNN, CNN, and DETOX [105]	Comment moderation	Gazzetta, and Wikipedia	1.73M	AUC: 98.03%
TNT [107]	Comment moderation - hate speech	Yahoo News and Finance, Twitter, and Wikipedia	1.56M	AUC: 97.3% F1: 79.1%
BERT [108]	Comment moderation	Reddit	1,017	Accuracy: 97.34% Precision: 96.99% Recall: 95.68% F1: 96.33% EFR: 91.2%
BERT [109]	Text perturbations toxicity	Multiple Sources	1M+	Accuracy: 75.5%
BERT, RoBERTa, and Perspective API [110]	Text perturbations toxicity	NoisyHate [111]	131K+	
BERT, RoBERTa, XLNet, and Naive Bayes [53]	Misinformation detection	YouTube	180	Accuracy: 89.4%
BERT [42]	Comment moderation – toxicity	Wikipedia and Reddit	92K	-
CNN-GRU, BiRNN, BiRNN-Attention, and BERT [106]	Hate speech detection	Twitter and Gab	20k+	Accuracy: 69.8% Marco F1: 68.7% AUC: 85.1%
SVM, FastText, BiLSTM, and RoBERTa [48]	Multiple tasks	Multiple Sources	204k+	Marco F1: 69.4%
ANN and DeepNet [55]	Fake news detection	Fakeddit and BuzzFeed	800K+	Accuracy: 95.2% Precision: 90.9% Recall: 95.2% F1: 93.0%
MIL [50]	Rumor detection and stance detection	Twitter and PHEME	722	AUC: 91.9% Micro F1: 80.9% Marco F1: 79.0%
EFN [49]	Fake news detection	Sina Weibo	160K	Accuracy: 87.2% F1: 87.4%
BiLSTM [56]	Rumor detection	Sina Weibo and Twitter	4,654	Accuracy: 94.8%
Transformer [54]	Fake news detection	NELA-GT-2019 and Fakeddit	4K	Accuracy: 74.8% Precision: 82.4% Recall: 77.6% F1: 74.9%

M: millions; K: thousands; The best model performances are reported in the table.



Content moderation, especially on the internet and social media platforms, frequently entails a collaborative effort between humans and machines, rather than them working in isolation. Previous studies [49], [50], [53], [54], [55], [56] found that using both post content and its entire history of user engagement (i.e., all the comments) is effective in detecting the legitimacy of information. For instance, Serrano et al. [53] leveraged online users' comments to predict COVID-19-related misinformation videos on YouTube; Guo et al. [49] incorporated both word embeddings and emotion embeddings of user comments within a news article discussion; and Raza and Ding [54] proposed a transformer-based approach by concatenating both news content and social contexts (e.g., posts on news, source of news, user creditability) to facilitate fake news detection. However, incorporating the entire history of user engagement with explicit information structure is not cost-effective and it can easily increase the computational overhead. In a research endeavor, an LSTM-based model [55] was applied to acquire knowledge about the UGC of new articles, and the Clauset-Newman-Moore algorithm was utilized to configure user-to-user connections within a user community while also discerning the veracity of news content; similarly, an attention model [56] based on BiLSTM was used to combine word embeddings from user posts and various social features (such as user profile details like follower count and registration time). This fusion was done at the post level to predict the detection of rumors at an event level, which encompasses both the source post and its subsequent reposts; a hierarchical attention model [50] used both bottom-up and top-down propagation tree structures to iteratively combine user opinions from various user comments associated with a claim. Additionally, it incorporated the UGC information of a claim to jointly verify rumors and detect stances. Nevertheless, all the deep learning models mentioned above were designed to moderate individual units of UGC, such as a tweet or a user comment. In addition, social media interactions occur over time, and moderators' expertise grows through active

participation in community activities and implementing interventions. As a result, there is a gap in the development of models that effectively utilize the dynamics of user engagement and moderators' domain knowledge for decision-making in the moderation of UGC.

## **2.6 Issues with Content Moderation in Social Media**

In light of both the potential for regulatory measures and growing public criticism, social media companies are increasingly pledging to take greater action to contain harmful UGC on their platforms [118]. The impact of content moderation on users' fundamental rights and democratic values is considerable, as it involves online platforms autonomously defining UGC removal standards at a global level [23]. In this section, I outline the primary shortcomings of current content moderation interventions employed by social media platforms:

*Opacity*: although moderation systems have been developed across various platforms, they are typically proprietary and not accessible for public use or study (e.g., internal enforcement tools offered by Meta [85], New York Times [119], YouTube [86], and Reddit [120]). In larger communities, predictive systems are commonly employed to identify the most damaging edits, whereas, in smaller communities, they are utilized to identify any edits that could potentially be damaging [80]. Similarly, Juneja et al. [121] found that Reddit's moderation practices violated the SCP. These violations were evident in various aspects - the use of implicit community norms rather than clear content policies to guide removal decisions. Furthermore, Ma and Kou [122] conducted an inductive thematic analysis, stating that multiple layers of opacity were discovered in YouTube algorithmic punishments, resulting in a precarious situation for YouTubers engaged in video creation. Furthermore, YouTubers responded to moderation punishments by adopting a reflexive approach, gradually acquiring and utilizing practical knowledge of algorithms to cope with the situation. The lack of transparency in moderation can evoke feelings of unfairness and frustration

among social media users [123], prompting them to create folk theories for their future online activities [124] or develop biased beliefs to rationalize moderation decisions [15].

*Lack of accountability:* while responsible AI has gained significant importance in research and development, one persisting concern about content moderation relates to the lack of accountability [30], [31], [125], [126], [127], [128], [129]. For instance, deploying algorithms without any human oversight can be detrimental – real chaos caused by the launch of an unsupervised anti-porn algorithm on Tumblr [130]. In addition, many social media platforms outsource content moderation to third-party contractors [131], given the scale and volume of UGC on social media platforms. These moderators may not be adequately trained or equipped to handle nuanced decisions, resulting in inconsistent and sometimes erroneous UGC removals. Another relevant scenario is appeal processes for UGC removal decisions, which are limited or difficult to navigate. As Soha and McDowell [132] argued “Even in clear cases of fair use, it can often require months as well as legal help and expert knowledge of copyright law to achieve a successful fair use claim.” Users may find it challenging to challenge erroneous content removals or seek recourse for perceived unfair treatment.

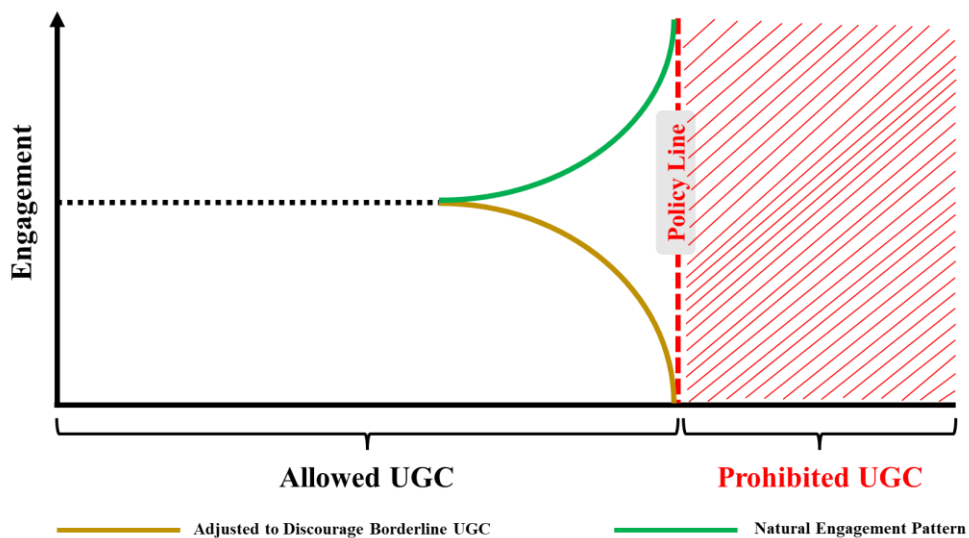
*Lack of explainability:* content moderation guidelines on social media platforms are often broad and open to interpretation. The lack of specific definitions for prohibited UGC can lead to inconsistent enforcement and confusion among users about what is acceptable. Even though the features (e.g., the degree of toxicity or nudity) incorporated in enforcement are in understandable terms to moderators, they may not be translatable to concepts that users would understand [133] or may not map cleanly to semantic concepts related to the relevant content policy [134]. Moreover, researchers have also highlighted that platforms moderate users and their content in an obscure manner, lacking sufficient explanations for their actions [30]. Sometimes, moderation decisions

are accompanied by concise, formal, and ambiguous explanations and are vaguely worded [30]. Nevertheless, Jhaver et al. [48] discovered that, on Reddit, an increase in the number of explanations provided in algorithmic moderation was correlated with a higher rate of users' content-generating behaviors aligning with the platform's policies [29].

*Bias and Subjectivity:* content moderation decisions can be influenced by the personal beliefs and biases of individual moderators or the platform's policies themselves. This subjectivity can lead to UGC removals that some users perceive as unfair or politically motivated, as Burk and Cohen [135] argued that the contextual factors needed to assess fair use standards cannot be programmed into automated systems – an argument supported by recent empirical studies of automated copyright enforcement that report substantial over-blocking of content on video sharing platforms [136], [137], [138]. Diakopoulos and Naaman's study [139] on news platform comment moderation revealed that media organizations recognize that moderators may introduce their own biases when evaluating standards. The issue of bias and subjectivity in content moderation decision-making has been well-evidenced. For instance, some critics [140] argued that social media platforms apply content moderation policies selectively, favoring certain high-profile users or allowing controversial content to remain online if it generates engagement and traffic. An empirical study [101] also showed that female moderators exhibit lower consistency in their content moderation decisions compared to male moderators. Moreover, it has been found that female moderators tend to be less sensitive to toxic content than their male counterparts. In addition to that, Marshall's discovery [141] highlighted the presence of algorithmic misogynoir within content moderation practices, as these systems were inherently influenced by the underlying principles of white colonialist culture. While there is an effort [101] in the trajectory of DADM to

promote fairness across individuals with diverse characteristics like race, gender, and religion, content moderation issues do not always align exclusively with these factors.

*The Bottom Line of Regulatory Strictness:* It can be observed from Figure 3 that online users tend to be more actively engaged with others when there is a natural intervention taken in place, but their behavior shifts in the opposite direction when stricter content moderation measures are implemented. This statement is also evidenced by an empirical study conducted by Seering et. al., [43] with 56 volunteer moderators from various online communities on three major platforms. Their investigation covered the entire journey of becoming and evolving as a moderator, managing misbehavior, and establishing community rules. The study particularly emphasizes the trade-off between the strictness of content moderation and users' expectations and calls to strike a balance between algorithmic and user-driven models of governance.



**Figure 3: User Engagement Patterns in Different Content Moderation Policies**

Addressing those issues is complex, requiring a careful balance between free speech, user safety, and responsible platform governance. The consequences of moderation punishments can significantly shape users' future behaviors [15] and have faced criticism for their substantial impact on restricting free expression [142]. Shneiderman et al. [143] assert that a significant challenge in

human-computer interaction research and practice is to design novel systems that enable users to comprehend the hidden algorithmic processes, thereby enhancing their ability to effectively manage their future actions. In response to this appeal, diverse researchers have recognized the significance of human-AI collaboration in bolstering trust in algorithmic decision-making [144]. Additionally, moderation systems that offer explanations of appeal were shown to enhance users' perceptions of fairness, trust, and transparency [145]. Jhaver et al.[29] showed that when users on Reddit were given moderation explanations, they became more inclined to understand the explicit UGC guidelines within particular subreddits. Furthermore, Kou and Gui [146] emphasized the importance of incorporating community context (such as shared values, knowledge, and community norms) in explanations. By doing so, users can gain a better understanding of how algorithms can be optimized to cater to the needs of end-users effectively. Last but not least, Cobbe [147] provided a theoretical summary of two effective strategies for countering algorithmic content moderation on social media: everyday resistance and organized resistance [83]. The former refers to the informal and individual acts of defiance or opposition towards algorithmic content moderation on social media platforms. It involves users finding subtle ways to bypass or subvert the moderation systems while still expressing their opinions or sharing content that may be deemed against platform policies or guidelines [148]. Unlike everyday resistance, organized resistance involves individual acts, organized resistance involves collaborative and coordinated actions aimed at addressing broader concerns related to content moderation practices [93]. In this dissertation, I focus on the combination of everyday resistance and organized resistance content moderation strategy to counter the persisting issues of content moderation.

However, user engagement on social media exhibits distinct attributes, including user discussions, and content creditability and stance, among others. Consequently, a comprehensive

understanding of the characteristics of user engagement in the context of social media content moderation remains elusive. This is particularly significant given the limited knowledge about the extent to which structure-based insights derived from user engagement can be employed to enhance the effectiveness and efficiency of content moderation. Additionally, social media interactions evolve over time, and moderators' expertise grows through active participation in community activities and implementing interventions. As a result, there is a gap in the development of models that effectively utilize the dynamics of user engagement and moderators' domain knowledge for decision-making in the moderation of UGC. Despite a few empirical studies that have delved into the decision-making processes of content moderation, focusing on understanding the perceived biases of content moderation from users, platforms and/or moderators, policymakers, and even bystanders, empirical investigations specifically probing into disparities in content moderation related to review information comprehensiveness, user roles, and content familiarity remain notably scarce.

## **CHAPTER 3: CHARACTERISTICS AND EFFICIENCY OF USER ENGAGEMENT IN CONTENT MODERATION**

### **3.1 Introduction**

Effective content moderation plays a pivotal role in fostering greater user engagement within online communities. When users feel secure and receive adequate support in a community, their propensity to actively participate in positive and constructive interactions is significantly heightened [36], [37]. Moreover, user engagement in social media pertains to the degree of interaction and communication exhibited by users towards fellow online users within a social media platform. This encompasses activities such as content creation, sharing, commenting, liking, sharing, and following other users [38]. User engagement holds intrinsic value for content moderation as it enables moderators to obtain feedback from the community regarding objectionable or inappropriate UGC. The significance of user engagement lies in its capacity to serve as both a process and outcome of user interactions [39], thereby presenting substantial opportunities for leveraging social networking techniques to derive insights from UGC. In this study, I define user engagement as online users' discussions/comments associated with a specific UGC.

Despite increasing research attention (e.g., [50], [53], [54], [55]) on leveraging user engagement for content moderation and establishing the effectiveness of incorporating user engagement in deep learning models, the focus has primarily been on the employment of users' profiles [49] or social networks [55], [56]. Among the few studies [49], [53], [54] that concentrated on user comments, both the content of a post and its entire history of user comments were utilized. In addition, incorporating the entire history of user engagement poses several notable technical challenges. First, storing and processing large volumes of user comments can be overwhelming,



especially for popular posts with numerous comments. This necessitates significant storage capacity and efficient data retrieval mechanisms. Second, deep learning models that incorporate vast amounts of text data require substantial computational resources for training and inference, leading to significant computational overhead. Third, user comments can contain a lot of irrelevant or redundant information that does not contribute to the moderation task. Filtering out noise without losing valuable context remains an ongoing challenge in information systems research. Fourth, the relevance of user comments may change over time. Older comments might not be as relevant as recent ones, and comments directly replying to the post might be more important than other comment threads, adding another layer of complexity to model development.

Consequently, research on content moderation that effectively engages with user comments remains scarce. More importantly, there is a lack of exploration regarding how the characteristics of user comments facilitate content moderation and the efficiency at which user comments can be processed while maintaining comparable model performance. Here, model performance refers to the binary content moderation decisions predicted by the developed model(s) compared to the ground truth data collected from the social media platform (i.e., Reddit). This study aims to address these research gaps by answering the following research questions.

RQ 1: What is the relationship between user engagement and the effectiveness of content moderation?

RQ 1.1: What major characteristics of user engagement impact the effectiveness of content moderation?

RQ 1.2: To what extent does the degree of user engagement impact the effectiveness of content moderation?

RQ 1.3: How can the efficiency of user engagement in content moderation be improved without compromising its effectiveness?

### 3.2 Related Work

Previous studies (e.g., [50], [53], [54], [55]) have found that user engagement is effective in detecting the legitimacy of information. Serrano et al. [53] used online users' comments to predict COVID-19-related videos with misinformation on YouTube, and Guo et al. [49] incorporated both word embeddings and emotion embeddings of user comments for fake news detection. In addition, Raza and Ding [54] proposed a transformer-based approach that combined news content and social contexts (e.g., posts on the news, news sources, and user creditability) to facilitate fake news detection. Kaliyar et al. [55] used an LSTM-based model to acquire knowledge about the UGC of news articles and applied the Clauset-Newman-Moore algorithm to configure user-to-user connections within a user community while discerning the veracity of news content. Similarly, an attention model [56] based on BiLSTM combined word embeddings from user posts and various social features (such as user profile details like follower count and registration time). This fusion at the post level aimed to detect rumors at an event level, which includes both the original post and its subsequent reposts. Additionally, a hierarchical attention model [50] utilized both bottom-up and top-down propagation tree structures to iteratively combine user opinions from various user comments associated with a claim, incorporating UGC information to jointly verify rumors and detect stances.

However, user engagement in social media displays distinct characteristics, including users' discussions, stances, and the creditability of UGC throughout online communications. Thus, a complete understanding of the characteristics of user engagement in the context of content moderation remains underexplored. This is particularly significant given the limited knowledge

about the extent to which structure-based insights derived from user engagement can be employed to enhance the effectiveness and/or efficiency of content moderation.

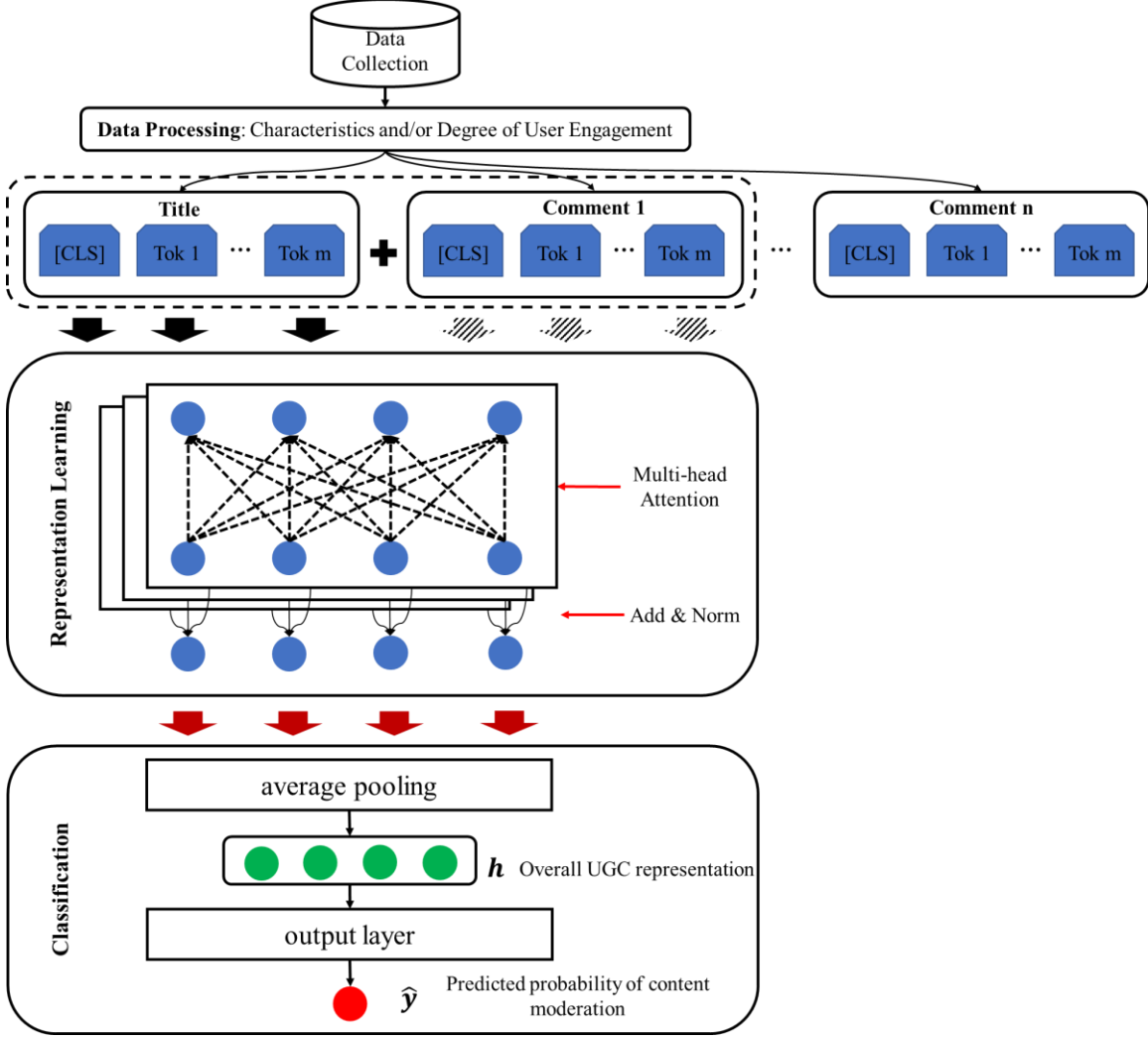
### 3.3 A RoBERTa-based Framework

RoBERTa [52] is pre-trained on the BookCorpus [149] and Wikipedia [150] datasets with a larger batch size of 8,000 and incorporates dynamic masking. This pre-training approach enables RoBERTa to understand nuanced contexts, making it highly proficient at identifying inappropriate or harmful content [151]. Moreover, RoBERTa employs a full-sentence training strategy to understand the relationships among sentences within the text. Given that the task involves using post titles and corresponding user comments for content moderation decisions, considering the contextual meanings of user comments with a post can be particularly advantageous. Therefore, I use a RoBERTa-based word embedding to enhance the word representation in this study. The overall structure of the RoBERTa-based framework can be found in Figure 4.

I formulate content moderation as a binary classification problem that classifies a UGC as either moderated or unmoderated. As shown in Equation 1, where  $u = \{t, c_1, c_2, \dots, c_n\}$  represent a UGC, which includes the title  $t$  and corresponding comments set  $c = \{c_1, c_2, \dots, c_n\}$ . Therefore,  $u$  serves as the source of input to the classification model.

$$y = f(\Theta, u) \quad (1)$$

where  $y$  denotes the classification result of a target UGC  $u$  (i.e., either moderated or unmoderated), and  $\Theta$  denotes the set of parameters of the classification function  $f(\cdot)$ .



**Figure 4: The RoBERTa-based Framework for Content Moderation**

The multi-head attention mechanism of RoBERTa can capture the relations among the words in the context. Specifically, the multi-head attention first maps each token to Query vector matrix  $Q$ , Key vector matrix  $K$ , and Value vector matrix  $V$ . After matching  $Q$  with  $K$  of all the tokens, the adjusted embedding vector for each token can be generated by the weighted sum of Value vector matrix  $V$  based on the  $Q$ - $K$  matching score (see Equation 2).

$$w = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Furthermore, I aggregate the word embedding vectors through an average pooling layer to generate an overall representation  $h$  for a UGC (i.e., post title and/or user comments) (see Equation 3).

$$h_u = \text{ave}([w_1, w_2, \dots, w_n]) \quad (3)$$

Finally, an output layer maps the features in the representation vector  $h$  to a probabilistic value and makes predictions on the softmax function of the UGC being moderated or not (see Equation 4).

$$\hat{y}_u = \sigma(W^T h_u + b) \quad (4)$$

where  $W \in R^{1 \times d}$ ,  $b \in R^1$  are learnable weight matrix and bias of the output layer respectively,  $\sigma(\cdot)$  is the sigmoid activation function which maps the output values to the range of 0 to 1,  $\hat{y}_u$  is the predicted probability of the UGC being moderated. Given that I formulate content moderation decision-making as a binary classification problem, I use the binary cross-entropy as the loss function for model training (see Equation 5).

$$\mathcal{L} = - \sum_{y_u \in \{y_u^+, y_u^-\}} y_u \log \hat{y}_u \quad (5)$$

where  $y_u^+$  and  $y_u^-$  denote the set of UGC labeled as moderated and unmoderated UGC respectively. By minimizing the loss, I train the model to generate classification results for content moderation.

### 3.4 Experiments

In this section, I introduce the data collection and preparation, model variations, and model performance evaluation respectively.

#### 3.4.1 Data Collection and Preparation

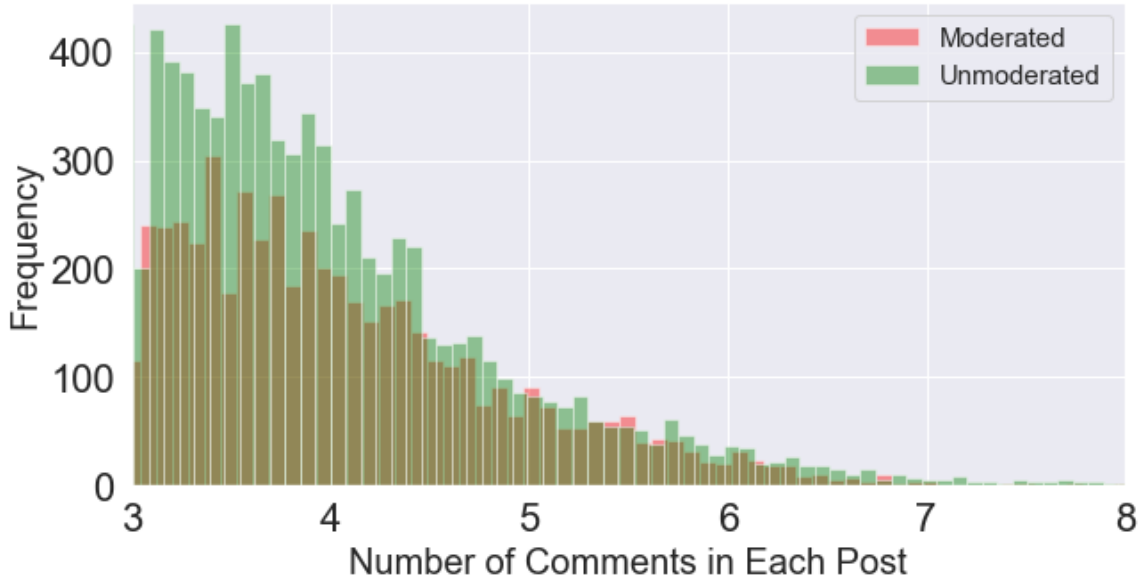
The data collection was limited to public online communities to comply with the platform's privacy policy. Thus, the procedure did not require approval from the Institutional Review Board

at the authors' institution. I chose Reddit as the platform for data collection. Each subreddit is a sub-online community that integrates a large fusion of UGC. I used the keyword “presidential election” to scrape Reddit posts via PRAW API<sup>1</sup> on a daily basis, over the course of six months before and after the 2020 presidential election day. In addition, I also collected the posts' metadata, such as post titles, post bodies, post-related comments, corresponding subreddit community names, voting scores, and postdates. Subsequently, I grouped the identified subreddits into 830 subreddits (i.e., online communities). To enhance the ecological validity of the study findings, I filtered inactive subreddits whose post frequency is below the average post frequency of all the identified subreddits, resulting in 245 subreddits.

Similar to the data collection method carried out by [152], [153], I performed another round of data collection by collecting the posts from each identified subreddit using the same timeframe. The re-collection process provides multi-dimensional information on whether the post content was being deleted by the original author or was being removed by a moderator of a subreddit or by an auto-moderator. Moreover, I snowballed the corresponding comments based on the posts that were collected from the identified subreddits and filtered out those comments with invalid responses (e.g., no-content comments or deleted comments). Additionally, I applied the following inclusion criteria to further filter the posts: 1) having more than 10 comments that directly replied to the original post and 2) all comments associated with each post should have karma scores. As a result, the final dataset contains 94 subreddits with 15,808 posts and 1,496,550 comments. The distribution of comments after the logarithm transformation is plotted in Figure 5.

---

<sup>1</sup> <https://praw.readthedocs.io/en/stable/index.html>



**Figure 5: The Distribution of User Comments among the Collected Posts**

### 3.4.2 Baseline Models and Variant User Engagement Models

To investigate the impact of characteristics and degree of user engagement on the performance of content moderation. I define two baseline models:

- Baseline 1: using post title only, and
- Baseline 2: using post title and randomized user comments.

Yet, the user engagement models are varied in two dimensions, including the characteristics and the degrees of user engagement. The former indicates various characteristics of user engagement, including

- temporality (i.e., comments sorted by reply time),
- creditability (i.e., received voting score credited by anonymous online users),
- orientation (i.e., directly responding to the original post), and
- credited-orientation (i.e., comments sorted by voting score while directly in response to the original post).

The latter stipulates the incorporation of the number of individual user comment(s) by

- incrementally adding 1~10 comments with the optimal characteristic of user engagement and post title to the model.

### **3.4.3 Performance Evaluation**

I randomly splitted the dataset into training and testing sets, using an 80/20 partition. The final dataset consists of 5,082 moderated posts and 7,564 unmoderated posts for training, and 1,261 moderated posts and 1,901 unmoderated posts for testing. I selected a set of widely used evaluation metrics, including accuracy, precision, recall, and F1, to measure predicted content moderation decisions in comparison to the decisions made by the moderator of the social media platform. Precision measures the proportion of moderated UGC actually being moderated by the platform; recall measures the proportion of actual moderated UGC that is predicted correctly; F1 is a harmonic mean of precision and recall; and accuracy is measured as the ratio of the sum of true positives and true negatives to all the predictions.

In addition, I used the model with the best characteristic of user engagement to evaluate the degree and efficiency of use engagement in content moderation. The degree of user engagement is measured by the number of user comments that are required to facilitate the satisfaction of model performance. The efficiency is measured by the expected duration of user engagement in minutes, which is the elapsed duration between the time of content posting to the time of receiving comment(s).

## **3.5 Results**

### **3.5.1 The Characteristics of User Engagement in Content Moderation**

The performance results of the model are illustrated in Figure 6. The findings indicate that the model integrating the orientation characteristic of user engagement achieves the best



performance, with an average accuracy of 82.65%. Following in descending order of performance are models incorporating temporality (81.89%), credited-orientation (80.71%), creditability (80.52%), baseline with randomized user comments (80.27%), and baseline with title only (75.9%). Moreover, the results in terms of precision, recall, and F1 show a similar pattern to those of accuracy.

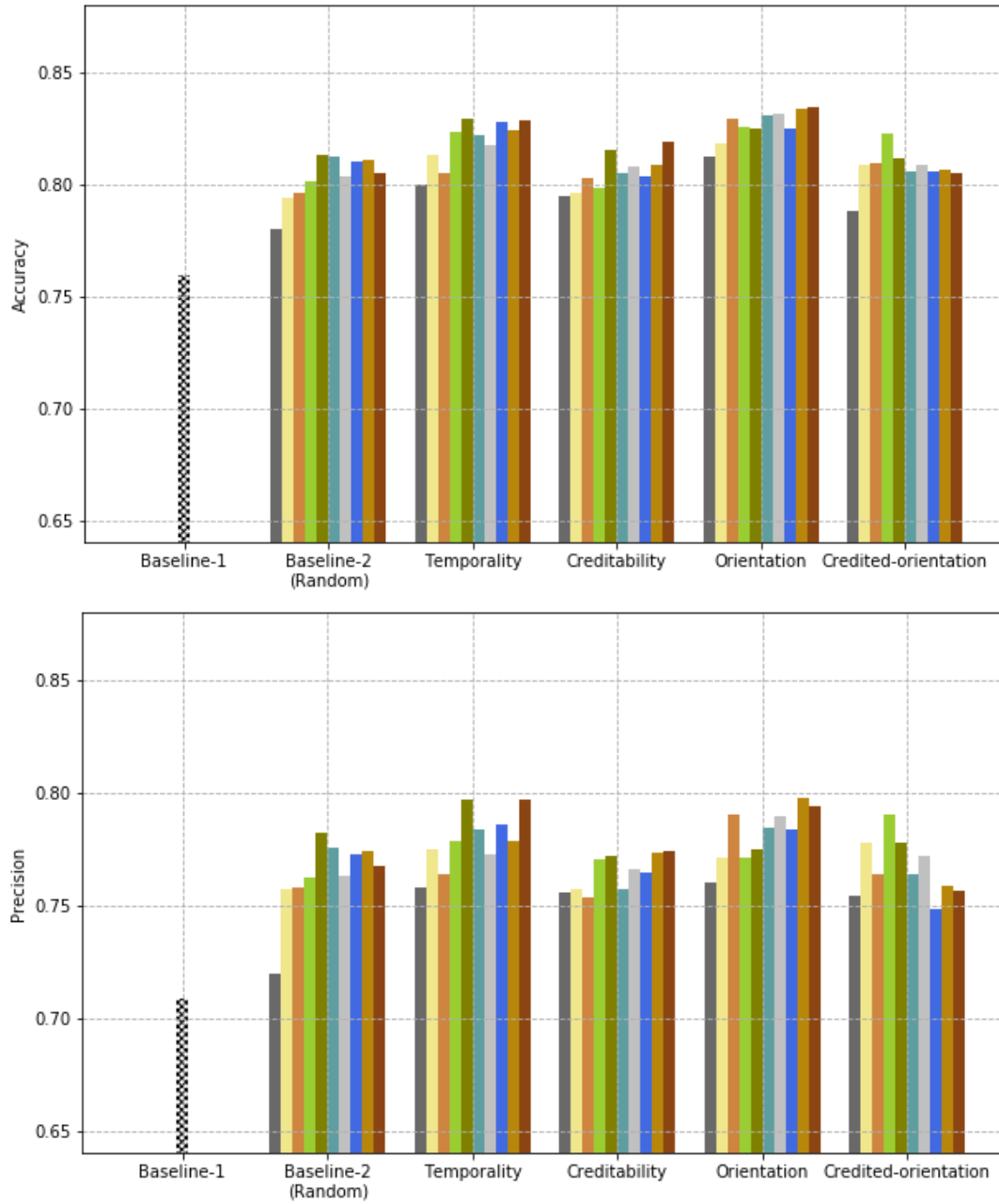
### **3.5.2 The Degree of User Engagement in Content Moderation**

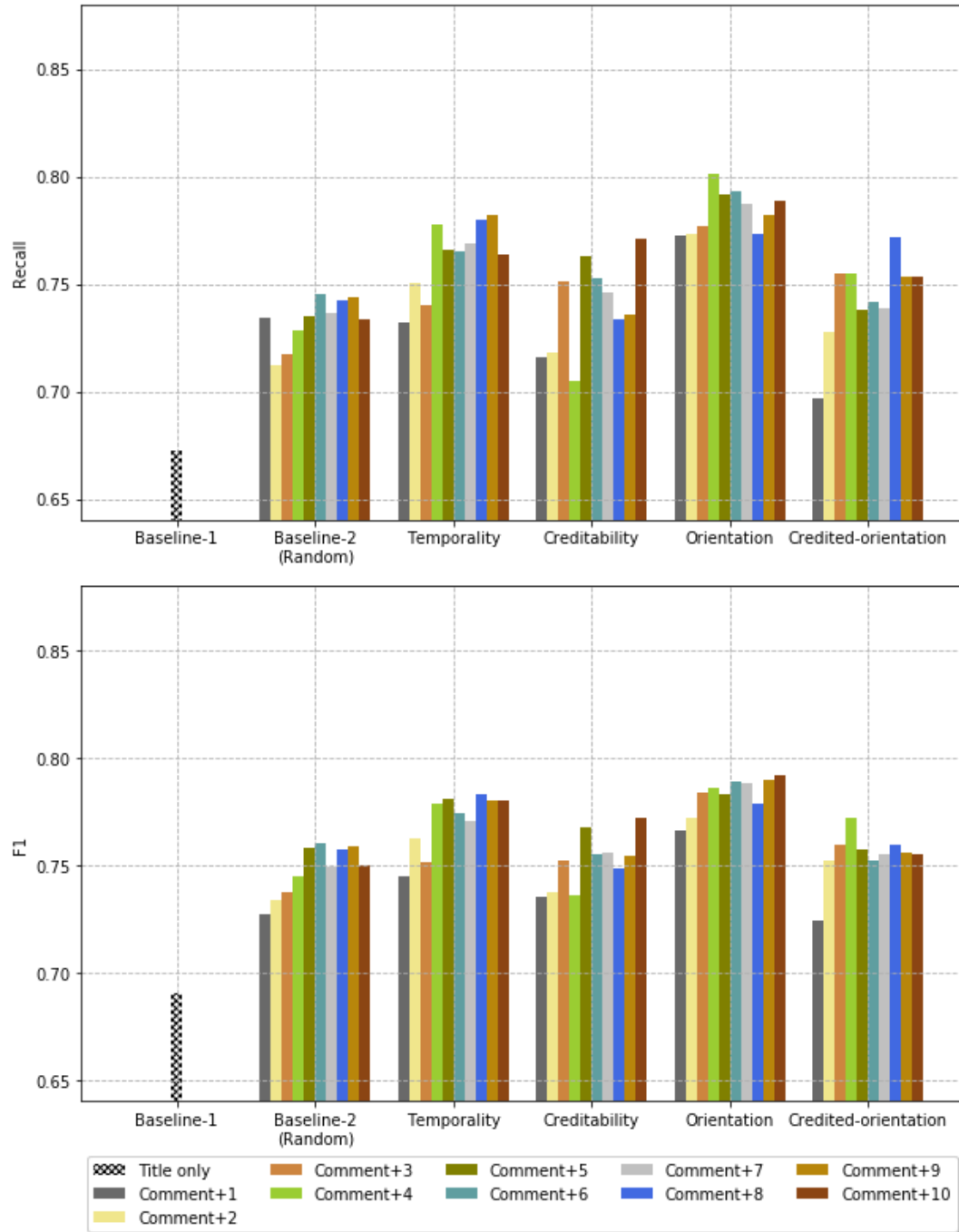
Based on the results of user engagement characteristics, I utilize the RoBERTa model with the orientation characteristic of user engagement to explore the impact of varying degrees of user engagement on content moderation performance. The results reveal that the model achieved 83.43% accuracy and 79.16% F1 score when incorporating the maximum number of user comments. It also achieves 79.77% precision with nine user comments and 80.10% recall with four user comments. Notably, model performance shows gradual improvement with the incorporation of up to three user comments, evidenced by increases in accuracy (from 81% to 83%), precision (from 76% to 79%), and F1 score (from 76% to 78%), and up to four user comments for recall (from 77% to 80%). However, further incorporation of user comments does not result in tangible performance improvement.

### **3.5.3 The Efficiency of User Engagement in Content Moderation**

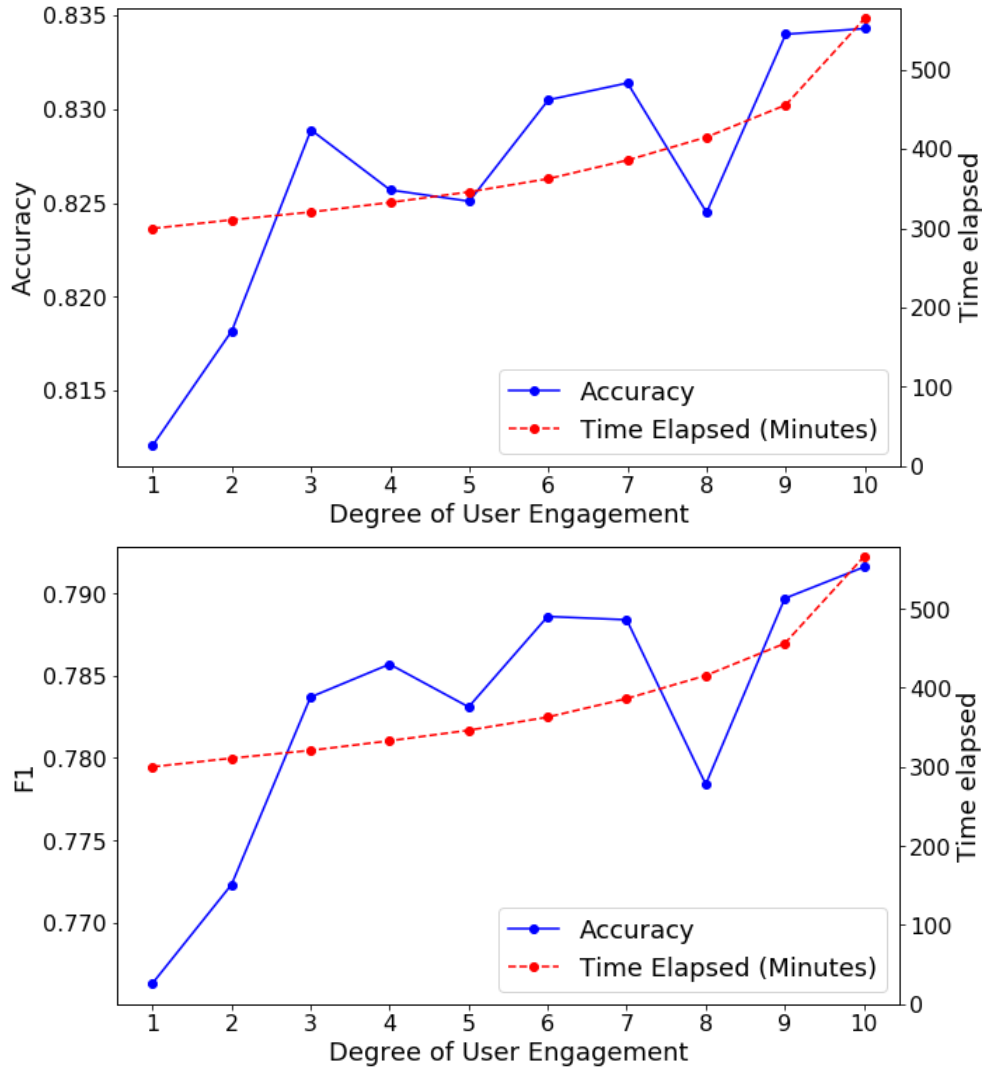
From an efficiency standpoint, Figure 7 illustrates a significant improvement in model performance in terms of accuracy between the first user comment (300 minutes) and the third user comment (321 minutes), and in terms of F1 score from the first user comment (300 minutes) to the fourth user comment (333 minutes). As the number of user comments increases, model performance varies from the fourth user comment (332 minutes) to the tenth (565 minutes). While

the model reaches its peak performance with the addition of ten user comments, the overall improvement is marginal.





**Figure 6: Performance Comparisons among Different Characteristics and Degrees of User Engagement**



**Figure 7: Efficiency of Content Moderation across Different Degrees of User Engagement**

### 3.6 Discussion

Incorporating user engagement into content moderation processes provides a more holistic and nuanced approach, enabling better decision-making in content moderation and fostering a healthier online community. The major findings of this research not only reveal the effectiveness of different characteristics and degrees of user engagement but also the efficiency of user engagement in content moderation.

To answer RQ 1.1, this research provides a shred of empirical evidence that incorporating the orientation characteristic of user engagement achieves the best performance in content

moderation, followed by the temporality characteristic. Such findings reveal that the rapidity and directness of user engagement can effectively reflect the relevancy of post content in social media, which in turn improves the performance of content moderation. To answer RQ 1.2, this research reveals that model performance improves by incorporating up to third user comments while fluctuating with further user comments expansion. The results indicate that using three user comments is sufficient to achieve satisfactory performance in content moderation. To answer RQ 1.3, the results of the efficiency of user engagement layout that the expected timely content moderation with user engagement is approximately five hours while achieving a robust performance.

This study makes multi-fold research contributions. This study provides the first empirical evidence for the effect of user engagement characteristics on content moderation in social media. Such evidence remains lacking in the literature. In addition, I examine user engagement in content moderation based on its degree rather than the entire history as previous studies do. This new perspective offers insights on how to improve the efficiency of content moderation by cultivating user engagement. Furthermore, the findings of this study not only serve as a guide for the development of content moderation techniques but also have practical implications for the design of online community policy and the optimization of moderation strategies.

## CHAPTER 4: PERCEIVED BIASES IN CONTENT MODERATION

### 4.1 Introduction

Content moderation is a highly contextual task given that moderation decisions about what is considered acceptable or undesirable are guided by an online community's norms or standards [154], [155], [156]. The norms of acceptability are not isolated but rather influenced by existing standards, and they can also be flexible and subject to change [101]. The question of what, if anything, should be filtered, or even removed, has consistently been a subject of intense societal debate (e.g., [30], [157], [158]). Excessively strict moderation can drive users to seek alternative platforms, causing platforms to redefine acceptable discourse through their terms of use, content policies, and enforcement measures [159]. However, adopting a flexible approach to content moderation regulation may expose the platform to the risk of allowing for harmful UGC that could potentially jeopardize the well-being of the online community [43]. Thus, it is challenging to strike a balance between users' expectations and the strictness of content moderation [160].

Prior research on content moderation has predominantly examined users' perceptions, emphasizing that these views are shaped by personal experiences and the transparency of the moderation process, with trust and fairness being key concerns [30], [31]. Additionally, moderators experience psychological stress [32] and encounter challenges related to AI accuracy [33], underscoring the necessity for effective training and support [34]. Policymakers play a critical role in influencing content moderation through legislation, striving to balance user protection with freedom of expression and continually updating standards to align with evolving societal norms [35]. However, the empirical investigation on the user perception of the content moderation decision-making process remains significantly scarce.

Drawn on the schema theory [161], the concept of content familiarity illuminates how individuals utilize existing knowledge to interpret and process information through pre-established mental frameworks or schemas. This proficiency enables reviewers to make well-informed, efficient, and consistent decisions, thereby enhancing content moderation practices that uphold community standards and user trust. Moreover, review content comprehensiveness in content moderation can be viewed as a reflection of democratic decision-making [162], specifically involving user discussions. Review content comprehensiveness fosters a diversified assessment of the UGC, which promotes a more equitable and inclusive decision-making process [163], contributing to informed and deliberative content moderation decisions [164]. Moreover, comprehending the disparities in how online users and moderators perceive content aids in designing effective content moderation policies and strategies. Nevertheless, none of the prior studies has conducted empirical investigations into the disparities in content moderation with respect to review information comprehensiveness, user roles, and content familiarity.

To fill the research gaps, I conducted a mixed-design online user study that recruited a diverse range of online participants. It is important to note that perceived bias does not necessarily imply actual bias; rather, it reflects how individuals perceive and interpret moderation actions. I aim to understand the perceived biases in content moderation by answering the following research questions.

RQ 2: What factors impact user perceptions in content moderation?

RQ 2.1: How does review information comprehensiveness impact perceptions of content moderation?

RQ 2.2: How do user roles impact perceptions of content moderation?

RQ 2.3: How does content familiarity impact perceptions of content moderation?

## 4.2 Related Work

Content moderation entails the oversight and management of UGC to ensure it adheres to community guidelines and legal standards. The primary goals are to prevent the spread of harmful UGC, safeguard users from abuse, and uphold the integrity of the platform [154], [155], [156]. The wide implementation of content moderation also results in several prominent cases, underscoring the intricacies of content moderation decision-making. For example, Facebook's approach to managing political content, especially during election periods, has led to debates regarding bias and the social media platform's influence on public opinion [30]. Similarly, YouTube's attempts to control hate speech through algorithmic moderation have faced criticism for both insufficient enforcement and excessive restrictions, revealing the challenges of context-sensitive moderation [157]. Twitter's strategies for handling misinformation, particularly in the realms of health and politics, highlight the difficulty of balancing accurate information dissemination with the preservation of free speech [158]. The primary factor intensifying debates surrounding content moderation is the involvement of various stakeholders—users, moderators, and policymakers—each holding divergent perspectives on what constitutes appropriate moderation practices [33].

Users' views on content moderation are often influenced by their personal experiences and the transparency of the moderation process. Trust and fairness are key themes in user perceptions, as users frequently question the fairness and consistency of moderation decisions. Perceived biases and a lack of transparency can undermine trust in the platform [30]. There is also a delicate balance between removing harmful UGC and protecting free speech. Users may perceive moderation as overly restrictive if they feel their expression is unfairly limited [32]. Furthermore, users seek clear



communication regarding moderation policies and decision-making processes. Transparency can improve trust and user satisfaction [31].

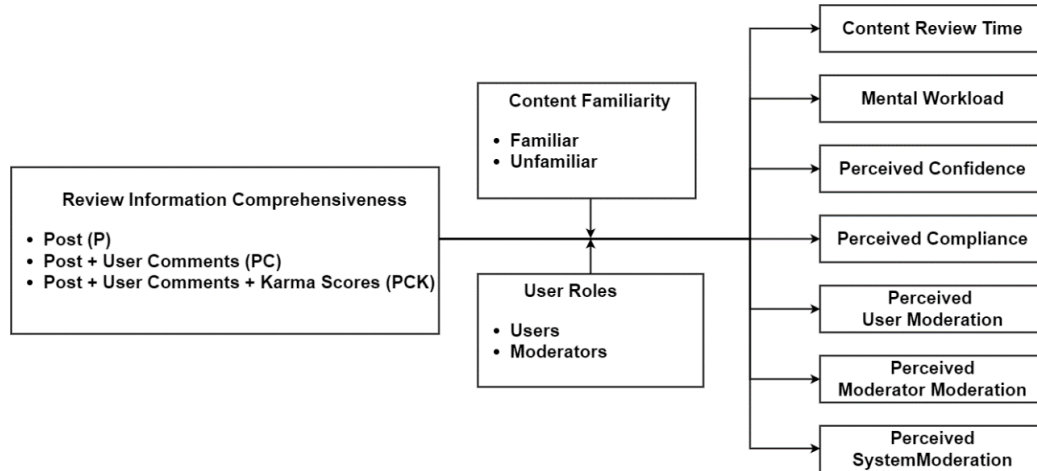
Moderators, whether human or algorithmic, are essential in enforcing content guidelines. Human moderators often face psychological stress due to continuous exposure to distressing content, which can affect their decision-making abilities and overall well-being [32]. The use of AI and algorithms for moderation raises concerns about accuracy and bias, as algorithms may struggle with making context-specific decisions, leading to incorrect content flagging [33]. Effective training and support systems are vital for moderators to perform their duties efficiently. Perceived inadequacies in support can affect their performance and morale [34].

Policymakers and regulators significantly influence content moderation practices through legislation and policy development. Developing legislation that balances user protection with freedom of expression is complex, requiring policymakers to consider diverse viewpoints and the global nature of social media [35]. The extent to which platforms should be held accountable for UGC remains debated, as regional differences in legal frameworks impact moderation practices [31]. Additionally, as societal norms change, so must the standards for acceptable UGC. Policymakers need to continuously update regulations to address new issues and technologies [30].

Despite previous research investigating content moderation from the perspectives of users, moderators, and policymakers, these studies were conducted in isolation. This approach lacks a comprehensive understanding of how users and moderators actually perform and perceive content moderation during the decision-making process. Furthermore, users and moderators may have varying online engagement experiences, leading to different levels of content familiarity with UGC. Hence, empirical evidence regarding the level of content familiarity in content moderation remains scarce. Moreover, content moderation decisions are influenced by the comprehensiveness of

review information presented to reviewers. However, it remains unclear to what extent the comprehensiveness of review information impacts these decisions. This gap in understanding calls for an integrated examination of the decision-making processes of both users and moderators, considering their levels of content familiarity and the detailed nature of the UGC they review.

### 4.3 Theoretical Foundation and Hypotheses Development



**Figure 8: The Research Model**

This research aims to investigate the impact of review information comprehensiveness, user roles, and content familiarity on content moderation. The overall research model is depicted in Figure 8, with the hypotheses elaborated in the subsequent sections.

#### 4.3.1 Review Information Comprehensiveness

When users actively engage in online communities, they can comprehensively evaluate UGC from various perspectives, including the content itself, relevant discussions, community rules and policies, and credibility indicators such as anonymous voting for content quality. Review information comprehensiveness adds layers of complexity to the review process. With more information available, individuals would be able to conduct a more thorough analysis of UGC and make well-informed decisions by cross-referencing UGC from various sources. Additionally,

review information comprehensiveness often includes conflicting opinions or ambiguous data that reviewers need to resolve [165]. Addressing these ambiguities requires additional time and effort to reach a consensus or clear understanding, and effectively handling diversified content may necessitate higher levels of training and expertise. Information processing theory [166] explains that individuals may need more time to interpret and integrate this information correctly, particularly if they are not fully familiar with all the diverse aspects involved. Cognitive load theory [167] further supports this notion, indicating that thorough examination of multiple aspects can significantly increase cognitive load, making tasks more demanding and time-consuming.

Access to comprehensive information on UGC enables reviewers to develop a more in-depth understanding of content, thereby gaining deeper insights that support more precise and informed decisions. This improved understanding strengthens reviewers' confidence in their judgments and enhances their assurance regarding compliance with guidelines [81]. Additionally, review information comprehensiveness presents multiple viewpoints, which helps mitigate individual biases and promotes a balanced perspective that facilitates objective and equitable decision-making. Consequently, this balanced approach enhances confidence in the integrity of moderation outcomes. Moreover, review information comprehensiveness facilitates the cross-verification of facts and context, allowing reviewers to ensure consistency, reliability, and a nuanced understanding that might otherwise be overlooked. This thorough evaluation supports decisions made with greater confidence. Furthermore, review information comprehensiveness reduces the impact of individual biases by offering a range of perspectives. This objectivity ensures that reviewers evaluate UGC based on its merits and adherence to guidelines rather than personal preferences, leading to more confident assessments of UGC.

In addition, detailed contextual information helps moderators better grasp user intent, resulting in more consistent and fair decisions. This thorough understanding allows moderators to distinguish between harmful and benign UGC, thereby aligning their decisions more closely with community guidelines. Furthermore, comprehensive information promotes the uniform application of community guidelines, reducing the likelihood of inconsistent decisions that could undermine trust in the moderation system. Roberts [81] points out that inconsistency often arises from a lack of sufficient context, leading to perceived biases and unfair treatment. By ensuring all relevant information is available, platforms can enhance consistency and improve the perception of fair compliance with community rules.

In addition, review information comprehensiveness exposes reviewers to examples of UGC that have previously sparked controversy or debate within the community [168]. Reviewers can anticipate that similar content may elicit strong reactions and reports, based on past instances where diverse perspectives clashed over the UGC's appropriateness or compliance with guidelines. In addition, reviewers exposed to comprehensive UGC develop a keen sense of what constitutes deviations from community norms or guidelines. They can recognize when UGC strays from accepted standards or values, making them more likely to anticipate that such UGC could prompt user reports.

Moreover, incorporating comprehensive information into content moderation processes introduces variability and complexity compared to the consistency of solitary post reviews. The multidimensional nature of this information can significantly increase the workload for moderators and potentially diminish the cohesion and effectiveness of the moderation efforts. This variability may impact trust in moderators' capacity to maintain a harmonious environment within platforms or community settings. In addition, delays in addressing content-related issues may further raise

concerns regarding the efficiency and responsiveness of moderation teams. Conversely, review information comprehensiveness empowers reviewers to more accurately discern patterns of compliance with guidelines, enhancing their ability to identify content that aligns with established norms. This capability fosters greater confidence that moderators will recognize compliant content and reduce the likelihood of intervention.

According to the technology acceptance model [169], individuals are more inclined to adopt and utilize technology, including automated systems, when they perceive it as both useful and user-friendly. Comprehensive review information exposes reviewers to various instances where automated systems effectively manage diverse UGC types and scenarios, thereby demonstrating their utility in content moderation. This exposure contributes to an increased confidence among reviewers in the efficacy of automated moderation systems. Moreover, expectancy theory [170] posits that individuals' expectations regarding future events are shaped by their beliefs about the likelihood of those events occurring and the associated outcomes. Reviewers exposed to comprehensive review information recognize consistent patterns in which automated systems apply rules and criteria consistently across diverse content contexts. This observed consistency reinforces their expectation that automated systems will continue to be utilized for content moderation tasks.

Therefore, the first set of hypotheses is proposed as follows:

***H1: For online content review, adding comments to the review post itself will lead to***

- (a) an increase in content review time of content moderation,
- (b) an increase in mental workload of content moderation,
- (c) an increase in perceived confidence of content moderation,
- (d) an increase in perceived compliance of content moderation,

- (e) an increase in perceived user moderation,
- (f) an increase in perceived moderator moderation, and
- (g) an increase in perceived systems moderation.

***H2: For online content review, adding comments and karma scores to the review post itself will lead to***

- (a) an increase in content review time of content moderation,
- (b) an increase in mental workload of content moderation,
- (c) an increase in perceived confidence of content moderation,
- (d) an increase in perceived compliance of content moderation,
- (e) an increase in perceived user moderation,
- (f) an increase in perceived moderator moderation, and
- (g) an increase in perceived systems moderation.

#### **4.3.2 User Roles**

Moderators often undergo specific training to handle content moderation effectively [31]. They are equipped with the knowledge and skills to recognize various types of content violations and handle them efficiently. This experience reduces the mental strain associated with decision-making. In addition, moderators have clear guidelines and are involved in the development of community rules and standard operating procedures to follow [42], [43]. These guidelines provide a framework for consistent decision-making, reducing the cognitive burden of determining appropriate actions. Moreover, moderators are more likely to focus on patterns and recurring violations rather than individual UGC pieces. This analytical approach can be less mentally demanding than making subjective decisions on a case-by-case basis. Unlike moderators, regular users may not have the necessary skills or knowledge to handle challenging situations effectively.

They may feel overwhelmed by the responsibility of identifying and reporting UGC violations without clear guidance, especially when dealing with borderline cases or complex issues.

When users review posts with additional information, they may feel more confident in their decisions due to the perceived thoroughness of their review process. Self-efficacy theory [171] supports this by indicating that individuals feel more capable and assured when they believe they have sufficient information to make informed decisions. The additional context can enhance users' perceived self-efficacy in moderation tasks. In addition, unlike moderators, users are likely to perceive higher compliance with community guidelines when they can see how content is rated and commented on by other users. Social proof theory [172] suggests that people look to the behavior of others to guide their own actions. When users see positive comments and high karma scores, they may infer that the content complies with community standards, reinforcing their perception of compliance.

Users often share similar experiences and perspectives with other users [163]. When they encounter UGC that they believe violates platform guidelines, they may feel that fellow users are more likely to understand their concerns and report them appropriately. In some online communities, there may be a strong emphasis on self-policing and community moderation [10]. Users may believe that their peers are more attuned to the community's values and will report UGC that goes against those values. In addition, some users might fear that moderators could be biased in their content moderation decisions or might abuse their power to silence dissenting opinions. Trusting fellow users to report UGC violations may be seen as a way to circumvent potential biases. Users may perceive their fellow users as more neutral and unbiased compared to platform-appointed moderators [30]. They might believe that moderators could be influenced by various factors, whereas other users are more likely to act in the community's best interest. Moderators are

typically trained to enforce platform guidelines consistently and accurately [31]. They may have a better understanding of the community's standards and the nuances of content moderation, leading them to trust their own judgment over other users. Moderators are aware that some users might report content based on personal biases or disagreements, or even with malicious intent. They exercise caution when considering reports from other users to avoid acting on false or misleading information. Platform tools and systems are designed to aid in content moderation efficiently. They offer specific functionalities, such as content flagging, user reporting, and automated content filtering, which regular users may not have access to or might not be aware of [4], [14], [15], [84]. In addition, platform tools can provide moderators with valuable data and insights, such as user behavior patterns and historical context, that help in making informed decisions. Regular users may not have access to this data, which could be critical in understanding the broader context of the reported UGC. The presence of additional information may lead users to believe that the platform's moderation systems are actively involved in evaluating UGC. Systems theory [173] posits that complex interactions within a system can influence perceptions of how that system operates. When users see indicators of engagement and evaluation, they may infer that automated moderation systems are effectively managing and monitoring the content.

Therefore, the second set of hypotheses is proposed as follows:

***H3: Compared with moderators, users reviewing posts with comments exhibit***

- (a) an increase in content review time of content moderation,
- (b) an increase in mental workload of content moderation,
- (c) an increase in perceived confidence of content moderation,
- (d) an increase in perceived compliance of content moderation,
- (e) an increase in perceived user moderation,



- (f) an increase in perceived moderator moderation, and
- (g) an increase in perceived systems moderation.

***H4: Compared with moderators, users reviewing posts with comments and karma scores exhibit***

- (a) an increase in content review time of content moderation,
- (b) an increase in mental workload of content moderation,
- (c) an increase in perceived confidence of content moderation,
- (d) an increase in perceived compliance of content moderation,
- (e) an increase in perceived user moderation,
- (f) an increase in perceived moderator moderation, and
- (g) an increase in perceived systems moderation.

### **4.3.3 Content Familiarity**

Given the diverse range of UGC prevalent on social media platforms, encountering both familiar and unfamiliar UGC is not uncommon. Schema theory [161] posits that familiarity with specific content types enables individuals to swiftly recognize patterns and accurately discern anomalies. Consequently, when evaluating familiar posts, individuals are better equipped to interpret their significance, thereby reducing the likelihood of misinterpretation. Furthermore, individuals familiar with UGC often possess a deeper understanding of its typical characteristics, facilitating quicker judgments and reducing the need for consultation or collaboration with others. They can independently assess and decide on UGC according to their accumulated knowledge and experience. Automatic processing theory [174] further supports this perspective by suggesting that when individuals encounter familiar information, their cognitive systems can process it more efficiently and automatically. This streamlined process reduces the cognitive load associated with

information processing, freeing up mental resources that can be directed toward higher-level decision-making tasks during the UGC review process.

When individuals with weaker content familiarity, they often experience uncertainty about their ability to make accurate moderation decisions. The inclusion of additional review information can significantly enhance their perceived confidence by providing context and cues about the UGC's reception and appropriateness. According to self-efficacy theory [175], individuals gain confidence in their abilities when they have access to comprehensive information that supports their decision-making. This is particularly relevant for individuals unfamiliar with the UGC, reinforcing their belief that they can accurately judge the content. Moreover, individuals with weaker content familiarity may struggle to independently assess whether content complies with community guidelines. Additional information, such as user comments and karma scores, can act as indicators of compliance, assisting these individuals in aligning their judgments with community standards. For those who are unfamiliar with the UGC, the visible approval or disapproval reflected in supplementary information provides a heuristic for determining compliance, thereby enhancing their perception of the content's adherence to rules.

When individuals who are unfamiliar with UGC, they may find it challenging to independently evaluate the appropriateness of posts. In such cases, the presence of additional information, such as user comments and/or karma scores acts as a proxy for community judgment, significantly influencing their perception of user moderation. According to social proof theory [172], individuals tend to rely on the behavior and opinions of others to guide their own actions, especially in situations of uncertainty. For individuals who are not well-versed in the content, visible community engagement through comments and karma scores can create a strong impression that the content has been thoroughly vetted by other users, thus increasing the perceived

level of user moderation. In addition, content familiarity can lead individuals to depend more on external indicators to gauge the validity of content moderation. The availability heuristic [176], a cognitive shortcut where individuals rely on immediate examples that come to mind, suggests that users are likely to assume that content with extensive comments and high karma scores has been scrutinized more rigorously by moderators. This heuristic is particularly influential for users unfamiliar with the UGC, as they are more inclined to trust visible signs of engagement and assume that moderators have actively reviewed such UGC. Consequently, the perceived level of moderator involvement is increased. Last but not least, the presence of user comments and karma scores can enhance individuals' trust in the platform's automated moderation systems. Systems theory [177] posits that the complex interactions within a system can significantly shape individuals' perceptions of how that system operates. When individuals lack familiarity with the UGC, they are more likely to believe that the platform's systems are effectively monitoring and evaluating the UGC based on visible indicators of user engagement. The assumption is that such metrics are integrated into the platform's moderation algorithms, thereby increasing their perception in the effectiveness of systems moderation.

Therefore, the third set of hypotheses is proposed as follows:

***H5: The lower the content familiarity, the greater the positive effect of adding comments to post itself on***

- (a) content review time in content moderation,
- (b) mental workload in content moderation,
- (c) perceived confidence in content moderation,
- (d) perceived compliance in content moderation,
- (e) perceived user moderation in content moderation,

- (f) perceived moderator moderation in content moderation, and
- (g) perceived systems moderation in content moderation.

***H6: The lower the content familiarity, the greater the positive effect of adding both comments and karma scores to post itself on***

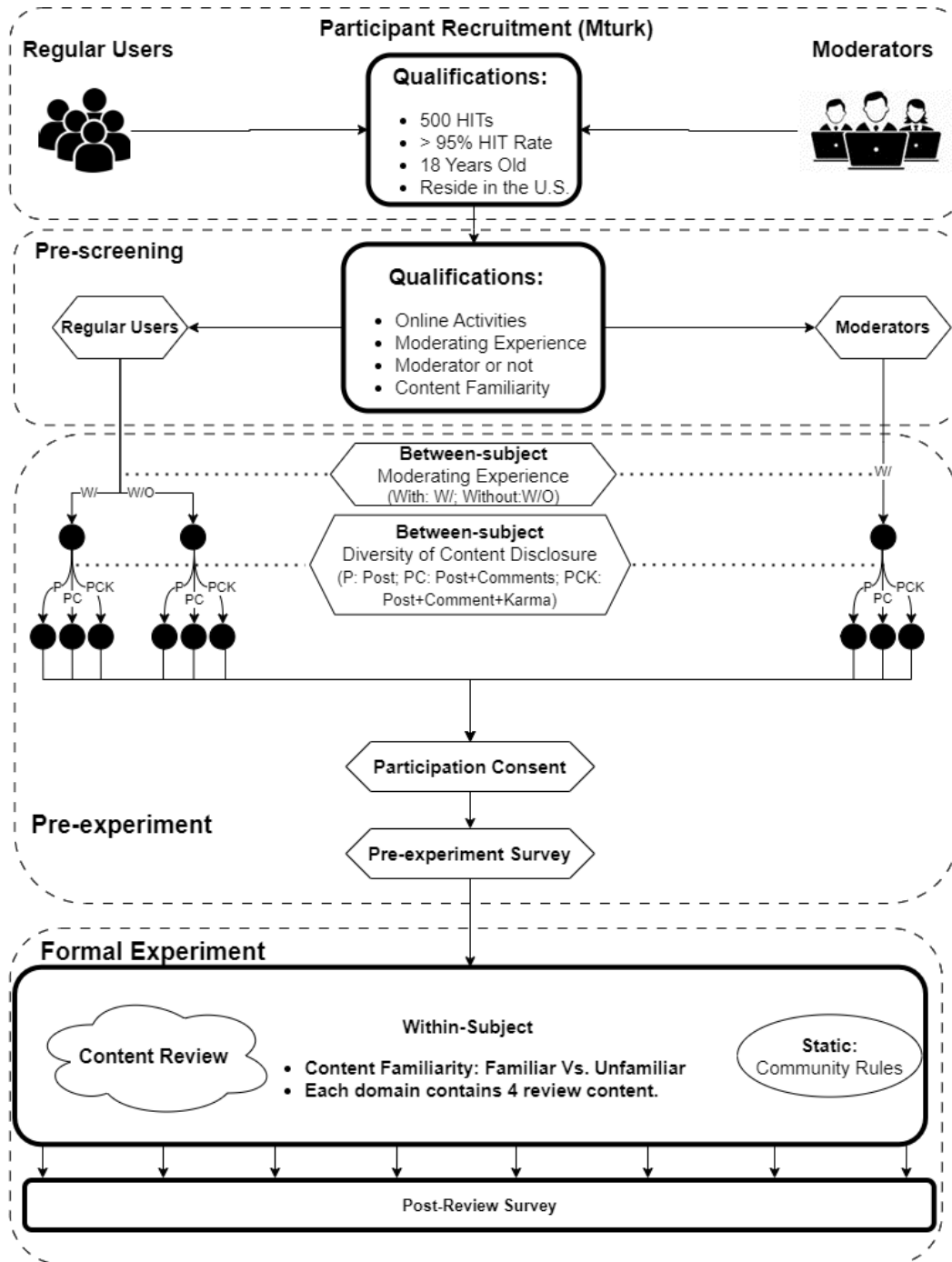
- (a) content review time in content moderation,
- (b) mental workload in content moderation,
- (c) perceived confidence in content moderation,
- (d) perceived compliance in content moderation,
- (e) perceived user moderation in content moderation,
- (f) perceived moderator moderation in content moderation, and
- (g) perceived systems moderation in content moderation.

## **4.4 Experiment Design**

This section provides details about the mixed-factor design for testing the research hypotheses. It introduces the experiment procedure, material preparation, instruments, and data analysis, respectively. The study (IRB-24-0623, see Appendix A) has been approved by the Institutional Review Board at the University of North Carolina at Charlotte.

### **4.4.1 Procedure and Tasks**

The experiment consists of four distinct phases (see Figure 9): participant recruitment, pre-screening, pre-experiment, and the formal experiment.



**Figure 9: The Procedure of the User Study**

Participant Recruitment Phase: Mturk was selected as the recruitment platform for this study due to the greater diversity of its workers compared to participants recruited through other methods [178]. Despite existing concerns about the quality and validity of data obtained from MTurk [179], this study targeted MTurk workers residing in the United States who were 18 years

of age or older and had completed more than 500 approved HITs with an approval rate exceeding 95%.


**Pre-Screening Phase:** A survey (see Appendix B) was utilized for the pre-screening phase. The survey questions are tailored based on participants' user roles: regular users with content moderation experience, regular users without content moderation experience, and moderators. All eligible participants must have actively engaged in online activities such as sharing posts, commenting on or replying to posts or comments, or moderating user-generated content. Participants were also required to indicate their familiarity with one of the selected domains: health, sports, fashion, or gaming. In addition, participants without moderation experience must not have served as moderators nor had their UGC moderated by social media platforms or online communities. Conversely, those with moderation experience must not have served as moderators but must have had their UGC moderated (e.g., UGC removal or deletion) by social media platforms or online communities. Moderators must have served as moderators in an online community or on a social media platform. Only participants meeting these qualification criteria were allowed to proceed to the consent and pre-experiment survey.

**Pre-Experiment Phase:** Upon obtaining participants' consent for participation (see Appendix C), a pre-experiment survey was administered (see Appendix D). This survey collected detailed demographic information, such as gender, education, ethnic origin, and IT expertise. It also assessed participants' perceived mental stress [180]. Additionally, both groups of users with and without moderation experience completed questionnaires regarding their experience with moderated UGC, while the moderator group completed questionnaires about their content moderation experience.

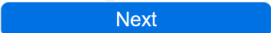
**Formal Experiment Phase:** In the formal experiment, all three participant groups first reviewed a set of community rules (see Figure 10). Subsequently, they were asked to review eight pieces of UGC (see Appendix F – Table 17) sequentially. After reviewing each piece of content, participants completed a post-review survey (see Appendix E).

The order of UGC and types (including post, post with comments, and post with comments and karma scores, as shown in Figure 11) of UGC distributions was randomized. Participants' content review behaviors, such as the start and end times of UGC review and button clicks, were logged into the designed system. Survey responses were collected using Qualtrics.

**Community Rules**

You must click on  to carefully read ALL the rules to proceed to the next step.

<p><b>1 Posts need to be directly related to Europa Universalis</b></p> <p>Posts must be related to Europa Universalis (i.e., a grand strategy video game). Just the title of the post being relevant does not qualify.</p>	<p><b>2 No memes, reaction pictures, or similar</b></p>
<p><b>3 No piracy or key resellers</b></p> <p>No links to pirated materials, pirated game mods, or key resellers. General discussion of piracy or leaked content is allowed.</p>	<p><b>4 No unapproved giveaways, surveys, or petitions</b></p>
<p><b>5 Follow our self-promotion limit</b></p> <p>Users may only make one self-promotional submission per seven days.</p>	<p><b>6 Follow the spirit of the rules</b></p>









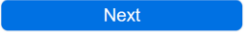
**Figure 10: Community Rules Review**


**Best government type in 2022 and why do you think so?**

I'm curious to see what others think the best government type is in 2022, maybe different ones for different reasons? Lets see your opinions!

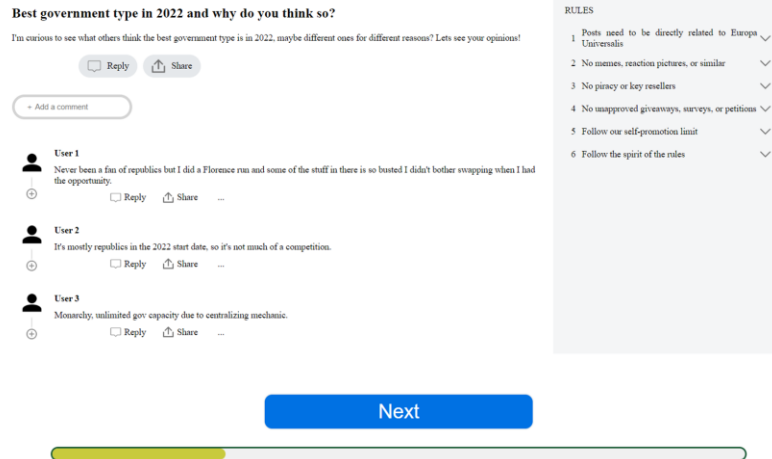
**RULES**

- 1 Posts need to be directly related to Europa Universalis 
- 2 No memes, reaction pictures, or similar 
- 3 No piracy or key resellers 
- 4 No unapproved giveaways, surveys, or petitions 
- 5 Follow our self-promotion limit 
- 6 Follow the spirit of the rules 

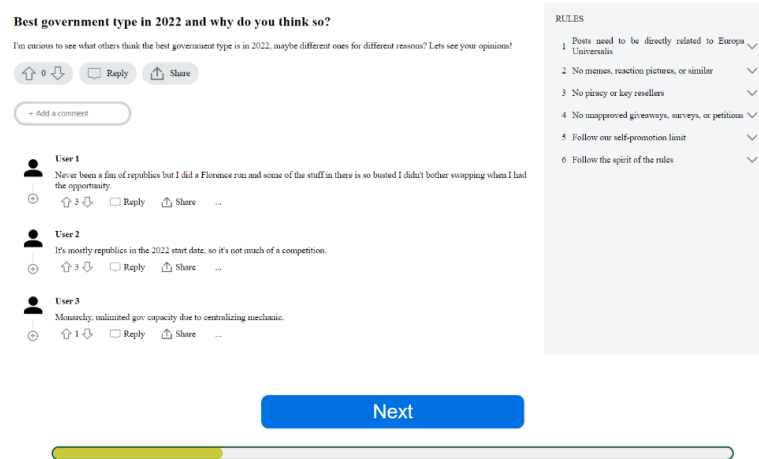




**(a) Post**



### (b) Post with Comments



### (c) Post with Comments and Karma Scores

**Figure 11: Content Review**

## 4.4.2 Variables and Measurements

Three independent variables are defined for this study, including review information comprehensiveness, user roles, and content familiarity. Additionally, seven dependent variables are examined through the lens of content review efforts and perceived moderation decision.

### Independent Variables:

- *Review Information Comprehensiveness*: This variable is operationalized through three conditions (see Figure 11): the interface displays post content only (P), post content with user comments (PC), and post content with both user comments and karma scores (PCK).



- *User Roles*: This variable is operationalized based on participants' self-identification in the pre-screening survey, categorizing them into two types: users and moderators. It is important to note that the users with and without content moderation experience are merged in this study.
- *Content Familiarity*: This variable is measured by their self-identification of domain familiarity in the pre-screening survey and is operationalized at two levels: familiar and unfamiliar.

### **Dependent Variables:**

Building on previous survey studies [42], [43], the dependent variables in this study are contextualized into two categories: content review efforts and perceived moderation decision. All dependent variables are measured using a seven-point Likert scale, except for mental workload.

Content review efforts are assessed from two aspects:

- *Content Review Time*: This is measured by recording the duration of participants' content review, from the end of the community rules review to the end of the content review.
- *Mental Workload*: This is assessed using three statements selected from the NASA Task Load Index [181]. Participants rate their responses, ranging from extremely agree to extremely disagree, to the following statements: "I felt that the task was mentally demanding", "I felt irritated, stressed, and annoyed versus content, relaxed, and complacent during the task", and "I felt successful in accomplishing what I was asked to do" (reverse coded). Responses to these statements are aggregated on a scale from 3 to 21.

Perceived moderation decision are measured using five variables:

- *Perceived Confidence*: This is measured by participants' responses to the statement, "Please rate the level of confidence in your moderation decision," ranging from extremely confident to extremely unconfident.
- *Perceived Compliance*: This is assessed by participants' responses to the statement, "I think this content complied with the community guidelines/rules," ranging from extremely agree to extremely disagree.
- *Perceived User Moderation*: This is evaluated based on participants' responses to the statement, "I anticipate that online users will feel the necessity for moderation of this content," ranging from extremely agree to extremely disagree.
- *Perceived Moderator Moderation*: This is measured by participants' responses to the statement, "I anticipate that community moderators will moderate this content," ranging from extremely agree to extremely disagree.
- *Perceived System Moderation*: This is assessed by participants' responses to the statement, "I am confident that the community moderation system will moderate this content," ranging from extremely agree to extremely disagree.

#### **4.4.3 Data Collection and Preparation**

I selected Reddit as the source for data collection. First, Reddit is a large-scale social media-based online community; second, the social aggregation on Reddit has less concentrated user networks, which allows online users to express their opinions naturally [182]; and third, the outcomes of content moderation are available, which can serve as the ground truths for this study [152], [153]. I selected four classic domains of online communities with different levels of content moderation, including fashion, health, sports, and gaming. Accordingly, the top 10 popular

subreddits were selected from each of the four domains. I leveraged a Pushshift API<sup>2</sup> to scrape posts from 40 subreddits daily across the four different domains from Aug 24 to October 28, 2022, resulting in 104,674 posts. Among these collected posts, the moderated ones account for 35% of the fashion, 30% in health, 26% in sports, and 13% in the gaming domain.

Since manual moderation is time-consuming, the efficiency of content moderation depends on several factors such as the type and volume of content, the complexity of the moderation rules, and the availability of human moderators. To enhance the ecological validity of the study findings, I used a PRAW API<sup>3</sup> to perform a second round of data collection of the collected posts two months later to validate whether the post content was moderated or not. Thereafter, I used a snowballing approach to collect the corresponding comments on all the posts. The metadata includes post content, post time, comment content, comment time, karma score, etc.

I meticulously reviewed and selected two moderated posts and two unmoderated posts from each domain, resulting in a total of 16 social media posts across four different domains. The post-selection criteria were as follows: (1) posts must be in English, (2) posts must receive a number of user comments relevant to the topic of discussion, (3) posts must receive karma scores for the post itself or users' comments, and 4\_ post must be controversial and lead a hot discussion within the community. Additionally, I extracted community rules from the respective communities (see Appendix F-Table 18) to aid participants in understanding these rules before making content moderation decisions.

---

<sup>2</sup> <https://github.com/pushshift/api>

<sup>3</sup> <https://praw.readthedocs.io/en/stable/>

#### **4.4.4 Participants**

A total of 2,130 participants were recruited for this study through Mturk. Of these, 1,021 participants completed the entire study, while the submissions of 601 participants were rejected due to incompleteness and quality concerns. Consequently, the final sample comprises 420 participants, including 139 users without content moderation experience, 141 users with content moderation experience, and 140 moderators. The gender distribution of the participants includes 253 males and 167 females. The age distribution is as follows: 115 participants are aged between 18-25 years, 115 between 26-30 years, 117 between 31-35 years, 27 between 36-40 years, 18 between 41-45 years, 3 between 46-50 years, 9 between 51-55 years, 12 between 56-60 years, and 4 between 61-65 years. Regarding educational attainment, 15 participants hold a high school degree or equivalent, 2 have an associate degree, 325 possess a bachelor's degree, and 78 have a master's degree or higher. In terms of employment status, 401 participants are full-time employees, 12 are part-time employees, 6 are self-employed, and 1 is retired. Last but not least, participants with and without content moderation experience received \$2 as a reward, and participants who were moderators received \$4 as a reward.

#### **4.4.5 Data Analyses**

To test the research hypotheses, I first conducted three sets of MANOVA analyses using Wilks' Lambda tests for all dependents, content review efforts, and perceived content moderation separately. In addition, I also conducted a repeated measures ANOVA to analyze the effects of review information comprehensiveness, user roles, and content familiarity. Additionally, I employed the Bonferroni multiple comparisons to examine review information comprehensiveness within each setting, operationalized by user roles and content familiarity. Both user roles and content familiarity serve as moderating factors in this research.

## 4.5 Results

In this section, I present the research findings from the online experiment. It is worth noting that a homogeneity test (See Appendix F – Table 19) is not required given the similar sizes of study participants among different user groups [183]. The results of the MANOVA indicate that there is a significant difference in the combination of all dependent variables based on review information comprehensiveness ( $p < .01$ ) and user roles ( $p < .01$ ). Similar differences are observed in the combinations of content review efforts and perceived moderation decision, with a marginal effect ( $p < .1$ ) of user roles on the combinations of perceived content moderation. Additionally, the interaction between user roles and review information comprehensiveness shows significant effects ( $p < .01$ ) on the combination of all dependent variables, content review efforts, and perceived moderation decision, but no significant effects on other interactions.

### 4.5.1 Content Review Efforts

The descriptive statistics for content review time and mental workload are presented in Table 3. The results of the repeated measures ANOVA and multiple comparisons for review information comprehensiveness are reported in Tables 4 and 5, respectively. Additionally, the results of multiple comparisons of review information comprehensiveness for review time and mental workload across different user roles and content familiarity are reported in Tables 6 and 7.

The results of content review time reveals a significant main effect for review information comprehensiveness ( $p < .001$ ) and a marginal effect for user roles ( $p < .1$ ), but no significant effect for content familiarity ( $p > .1$ ). Additionally, the interaction effect between review information comprehensiveness and user roles is significant ( $p < .001$ ), indicating that the effect of review information comprehensiveness is moderated by user roles. However, the interaction effect between review information comprehensiveness and content familiarity is not significant ( $p > .1$ ).

Moreover, the content review time for participants reviewing PCK is significantly longer than for PC ( $p < .001$ ) and P ( $p < .001$ ). The results also show that the difference in content review time between P and PC is significant for users ( $p < .001$ ) but not for moderators ( $p > .1$ ), while the difference in content review time between P and PCK is significant for both users ( $p < .001$ ) and moderators ( $p < .001$ ). Therefore, H3(a) is supported and H4(a) is not supported. Furthermore, the difference in content review time is significant ( $p < .001$ ) between P and PC and between P and PCK regardless of content familiarity, thus both H5(a) and H6(a) are not supported. Based on the multiple comparison results, H1(a) is partially supported and H2(a) is supported.

The ANOVA on mental workload indicates a significant main effect for review information comprehensiveness ( $p < .05$ ), user roles ( $p < .05$ ), and content familiarity ( $p < .05$ ). Additionally, the interaction effect between review information comprehensiveness and user roles is marginal ( $p < .1$ ), while the interaction effect between review information comprehensiveness and content familiarity is not significant ( $p > .1$ ). Furthermore, the mental workload for participants reviewing PCK is significantly greater than for those reviewing P ( $p < .01$ ), but not significantly different from those reviewing PC ( $p > .1$ ). The results also demonstrate that the difference in mental workload between P and PC is marginally significant for users ( $p < .1$ ) but not for moderators ( $p > .1$ ), whereas the difference in mental workload between P and PCK is significant for moderators ( $p < .05$ ) but not for users ( $p > .1$ ). Consequently, H3(b) is supported, while H4(b) is not supported. Regardless of content familiarity, the difference in mental workload is significant between P and PCK ( $p < .05$ ), but not between P and PC ( $p > .1$ ). Therefore, both H5(b) and H6(b) are not supported. Based on the multiple comparison results, H1(b) is not supported, and H2(b) is partially supported.

**Table 3: Descriptive Statistics of Content Review Efforts**

User Roles	Content Familiarity	Review Information Comprehensiveness	Content Review Time	Mental Workload
Moderators	Familiar	P	32.83(70.57)	12.21(2.50)
		PC	46.7(63.37)	12.49(2.22)
		PCK	115.78(152.60)	12.99(2.37)
	Unfamiliar	P	23.77(23.21)	12.23(2.24)
		PC	44.84(48.81)	12.18(2.29)
		PCK	117.46(148.02)	12.72(2.49)
Users	Familiar	P	23.22(61.22)	12.05(3.00)
		PC	62.96(148.08)	12.45(2.41)
		PCK	74.25(134.98)	12.21(2.63)
	Unfamiliar	P	20.85(32.98)	11.98(2.91)
		PC	76.18(205.60)	12.33(2.47)
		PCK	70.4(121.23)	12.26(2.48)

**Table 4: The Results of Repeated ANOVA of Content Review Efforts**

Independent Variables and Interactions	Content Review Time	Mental Workload
User Roles	(3.299)<.070†	(4.525).034*
Content Familiarity	(.010).922	(4.125).042*
Review Information Comprehensiveness	(66.892)<.001***	(4.343).013*
User Roles* Review Information Comprehensiveness	(15.983)<.001***	(2.674).069†
Content Familiarity* Review Information Comprehensiveness	(.759).468	(.886).413

F values are reported in parentheses; \*\*\*:  $p < .001$ ; \*:  $p < .05$ ; †:  $p < .1$ .

**Table 5: Multiple-Comparison Results for Content Review Efforts across Review Information Comprehensiveness**

	Review Information Comprehensiveness (I) (J)	Content Review Time (I-J) <sup>a</sup>	Mental Workload (I-J) <sup>a</sup>
P	PC	(-32.504)<.001***	(-.247).279
	PCK	(-69.304)<.001***	(-.429).009**
PC	PCK	(-36.799)<.001***	(-.182).647

<sup>a</sup>: mean difference; \*\*\*:  $p < .001$ ; \*\*:  $p < .01$

**Table 6: Multiple-Comparison Results for Content Review Efforts for Each User Role across Review Information Comprehensiveness**

User Roles	Review Information Comprehensiveness (I) (J)		Content Review Time (I-J) <sup>a</sup>	Mental Workload (I-J) <sup>a</sup>
Moderators	P	PC	-17.471	-.116
		PCK	-88.317***	-.636*
	PC	PCK	-70.846***	-.519†
Users	P	PC	-47.538***	-.377†
		PCK	-50.290***	-.222
	PC	PCK	-2.753	.155

<sup>a</sup>:mean difference; \*\*\*:  $p < .001$ ; \*:  $p < .05$ ; †:  $p < .1$

**Table 7: Multiple-Comparison Results for Content Review Efforts for Each Content Familiarity across Review Information Comprehensiveness**

Content Familiarity	Review Information Comprehensiveness (I) (J)		Content Review Duration (I-J) <sup>a</sup>	Mental Workload (I-J) <sup>a</sup>
Familiar	P	PC	-26.807***	-.340
		PCK	-66.989***	-.474*
	PC	PCK	-40.182***	-.133
Unfamiliar	P	PC	-38.202***	-.153
		PCK	-71.618***	-.384*
	PC	PCK	-33.416***	-.231

<sup>a</sup>:mean difference; \*\*\*:  $p < .001$ ; \*:  $p < .05$

#### 4.5.2 Perceived Moderation Decision

The descriptive statistics for perceived moderation decision measures, including perceived confidence, perceived compliance, perceived user moderation, perceived moderator moderation, and perceived system moderation are presented in Table 8. The results of the repeated measures ANOVA and multiple comparisons for review information comprehensiveness are reported in Tables 9 and 10, respectively. Additionally, the results of multiple comparisons of review information comprehensiveness for perceived moderation decision measures across different user roles and content familiarity are reported in Tables 11 and 12.



The results of perceived confidence reveal a marginal main effect for review information comprehensiveness ( $p < .1$ ), but no significant effects for user roles ( $p > .1$ ) and content familiarity ( $p > .1$ ). Additionally, the interaction effect between review information comprehensiveness and user roles is marginally significant ( $p < .1$ ); however, the interaction effect between review information comprehensiveness and content familiarity is not significant ( $p > .1$ ). Furthermore, the perceived confidence for participants reviewing PCK is marginally greater than for those reviewing P ( $p < .1$ ), but not significantly different from those reviewing PC ( $p > .1$ ). The results also indicate that the difference in perceived confidence between P and PC is significant for moderators ( $p < .05$ ) but not for users ( $p > .1$ ), and the difference in perceived confidence between P and PCK is marginally significant for moderators ( $p < .1$ ) but not for users ( $p > .1$ ). Therefore, H3(c) and H4(c) are not supported. Furthermore, the difference in perceived confidence between P and PC is not significant ( $p > .1$ ) regardless of content familiarity, but the difference in perceived confidence between P and PCK is significant ( $p < .05$ ) for reviewing familiar content. Thus, H5(c) and H6(c) are not supported. Based on the multiple comparison results, both H1(c) and H2(c) are partially supported.

The results of perceived compliance reveal a marginal main effect for review information comprehensiveness ( $p < .1$ ), but no significant effects for user roles ( $p > .1$ ) and content familiarity ( $p > .1$ ), nor their interaction effects ( $p > .1$ ). Furthermore, the perceived compliance for participants reviewing PCK is marginally greater than for those reviewing P ( $p < .1$ ), but not significantly different from those reviewing PC ( $p > .1$ ). The results also indicate that the differences in perceived compliance between P and PC and between P and PCK are not significant for both moderators and users ( $p > .1$ ). Therefore, H3(d) and H4(d) are not supported. Additionally, the difference in content review time between P and PC is not significant ( $p > .1$ ) regardless of

content familiarity, but the difference in content review time between P and PCK is marginally significant ( $p < .1$ ) for reviewing unfamiliar content. Thus, H5(d) is not supported but H6(d) is supported. Based on the multiple comparison results, H1(d) is not supported, and H2(d) is partially supported.

The results of perceived user moderation reveal a significant main effect for review information comprehensiveness ( $p < .05$ ) and content familiarity ( $p < .05$ ), but not for user roles ( $p > .1$ ). Additionally, the interaction effect between review information comprehensiveness and user roles is significant ( $p < .05$ ); however, the interaction effect between review information comprehensiveness and content familiarity is not significant ( $p > .1$ ). Furthermore, the perceived user moderation for participants reviewing PCK is significantly greater than for those reviewing P ( $p < .01$ ), but not significantly different from those reviewing PC ( $p > .1$ ). The results also indicate that the difference in perceived user moderation between P and PC is significant for users ( $p < .001$ ) but not for moderators ( $p > .1$ ), and the difference in perceived user moderation between P and PCK is significant for users ( $p < .001$ ) but not for moderators ( $p > .1$ ). Therefore, both H3(e) and H4(e) are supported. Furthermore, the difference in perceived user moderation between P and PC is not significant ( $p > .1$ ) regardless of content familiarity, but the difference in perceived user moderation between P and PCK is significant ( $p < .001$ ) for reviewing unfamiliar content. Thus, H5(e) is not supported but H6(e) is supported. Based on the multiple comparison results, both H1(e) and H2(e) are partially supported.

The results of perceived moderator moderation reveal a marginal main effect for user roles ( $p < .1$ ), but not for review information comprehensiveness ( $p > .1$ ) and content familiarity ( $p > .1$ ). Additionally, the interaction effect between review information comprehensiveness and user roles is significant ( $p < .05$ ); however, the interaction effect between review information

comprehensiveness and content familiarity is not significant ( $p > .1$ ). Furthermore, there is no significant difference among P, PC, and PCK in perceived moderator moderation ( $p > .1$ ). The results also indicate that the difference in perceived moderator moderation between P and PC is significant for users ( $p < .01$ ) but not for moderators ( $p > .1$ ), and the difference in perceived moderator moderation between P and PCK is significant for users ( $p < .05$ ) but not for moderators ( $p > .1$ ). Therefore, H3(f) and H4(f) are supported. Furthermore, the difference in perceived moderator moderation between P and PC is not significant ( $p > .1$ ) regardless of content familiarity, but the difference in perceived moderator moderation between P and PCK is significant ( $p < .05$ ) for reviewing unfamiliar content. Thus, H5(f) is not supported but H6(f) is supported. Based on the multiple comparison results, both H1(f) and H2(f) are partially supported.

The results of perceived system moderation reveal a significant main effect for review information comprehensiveness ( $p < .05$ ), but not for user roles ( $p > .1$ ) and content familiarity ( $p > .1$ ). Additionally, there is no significant interaction effect ( $p > .1$ ). Furthermore, the difference in perceived system moderation is significant between PCK and P ( $p < .05$ ), but not between PC and P ( $p > .1$ ). The results also indicate that the difference in perceived system moderation between P and PC is significant for users ( $p < .01$ ) but not for moderators ( $p > .1$ ), and the difference in perceived system moderation between P and PCK is significant for users ( $p < .01$ ) but not for moderators ( $p > .1$ ). Therefore, H3(g) and H4(g) are supported. Furthermore, the difference in perceived system moderation between P and PC is not significant ( $p > .1$ ) regardless of content familiarity, but the difference in perceived system moderation between P and PCK is significant ( $p < .01$ ) for reviewing unfamiliar content. Thus, H5(g) is not supported but H6(g) is supported. Based on the multiple comparison results, both H1(g) and H2(g) are partially supported.

Table 8: Descriptive Statistics of Perceived Moderation Decision

User Roles	Content Familiarity	Review Information Comprehensiveness	Perceived Confidence	Perceived Compliance	Perceived User Moderation	Perceived Moderator Moderation	Perceived System Moderation
Moderators	Familiar	P	1.9(0.86)	5.21(1.04)	5.18(1.14)	5.3(1.07)	5.22(1.15)
		PC	2.12(0.95)	5.27(1.29)	5.12(1.13)	5.08(1.29)	5.26(1.26)
		PCK	2.17(0.84)	5.33(1.19)	5.13(1.31)	5.26(1.27)	5.27(1.13)
	Unfamiliar	P	1.97(0.83)	5.16(1.05)	5.02(1.18)	5.14(1.07)	5.19(1.16)
		PC	2.18(0.98)	5.27(1.28)	4.9(1.21)	5.17(1.41)	5.16(1.29)
		PCK	2.11(0.91)	5.36(1.04)	5.2(1.18)	5.3(1.33)	5.37(1.17)
Users	Familiar	P	2.13(0.95)	5.17(1.11)	4.83(1.47)	4.91(1.51)	5.03(1.39)
		PC	2.09(0.96)	5.25(1.17)	5.12(1.3)	5.24(1.29)	5.25(1.2)
		PCK	2.18(0.98)	5.29(1.21)	5.15(1.37)	5.11(1.47)	5.22(1.35)
	Unfamiliar	P	2.15(1.00)	5.12(1.25)	4.71(1.5)	4.92(1.52)	4.97(1.42)
		PC	2.13(0.99)	5.25(1.19)	5.05(1.22)	5.18(1.39)	5.25(1.24)
		PCK	2.17(1.00)	5.27(1.21)	5.12(1.37)	5.2(1.35)	5.28(1.25)

Table 9: The Results of Repeated ANOVA of Perceived Moderation Decision

Independent Variables and Interactions		Perceived Confidence	Perceived Compliance	Perceived User Moderation	Perceived Moderator Moderation	Perceived System Moderation
User Roles		(2.333).127	(.661).416	(2.629).105	(3.760).053†	(1.837).175
Content Familiarity		(.665).415	(.160).690	(5.966).015*	(.001).970	(.006).939
Review Information Comprehensiveness		(2.783).062†	(2.766).063†	(4.615).010*	(2.171).114	(3.531).029*
User Roles* Review Information Comprehensiveness		(2.698).068†	(.075).927	(4.285).014*	(3.485).031*	(1.554).212
Content Familiarity* Review Information Comprehensiveness		(1.525).218	(.236).790	(2.299).101	(1.293).275	(1.559).211

F values are reported in parentheses; \*:  $p < .05$ ; †:  $p < .1$ .

**Table 10: Multiple-Comparison Results for Perceived Moderation Decision across Review Information Comprehensiveness**

Review Information Comprehensiveness (I) (J)		Perceived Confidence (I-J) <sup>a</sup>	Perceived Compliance (I-J) <sup>a</sup>	Perceived User Moderation (I-J) <sup>a</sup>	Perceived Moderator Moderation (I-J) <sup>a</sup>	Perceived System Moderation (I-J) <sup>a</sup>
P	PC	(-.094).239	(-.095).407	(-.113).340	(-.098).544	(-.124).224
	PCK	(-.120).076†	(-.147).062†	(-.216).007**	(-.150).121	(-.180).029*
PC	PCK	(-.026)1	(-.052)1	(-.103).453	(-.052)1	(-.056)1

<sup>a</sup>:mean difference; \*\*:  $p < .01$ ; \*:  $p < .05$ ; †:  $p < .1$

**Table 11: Multiple-Comparison Results for Perceived Moderation Decision for Each User Role across Different Review Information Comprehensiveness**

User Roles	Review Information Comprehensiveness (I) (J)		Perceived Confidence (I-J) <sup>a</sup>	Perceived Compliance (I-J) <sup>a</sup>	Perceived User Moderation (I-J) <sup>a</sup>	Perceived Moderator Moderation (I-J) <sup>a</sup>	Perceived System Moderation (I-J) <sup>a</sup>
Moderators	P	PC	-.216*	-.083	.088	.096	-.002
		PCK	-.202†	-.160	-.066	-.061	-.112
	PC	PCK	.014	-.077	-.154	-.157	-.110
Users	P	PC	.028	-.107	-.313***	-.293**	-.246**
		PCK	-.037	-.134	-.365***	-.240*	-.247**
	PC	PCK	-.065	-.027	-.052	-.053	-.001

<sup>a</sup>:mean difference; \*\*\*:  $p < .001$ ; \*\*:  $p < .01$ ; \*:  $p < .05$ ; †:  $p < .1$

**Table 12: Multiple-Comparison Results for Perceived Moderation Decision for Each Content Familiarity across Different Review Information Comprehensiveness**

Content Familiarity	Review Information Comprehensiveness (I) (J)		Perceived Confidence (I-J) <sup>a</sup>	Perceived Compliance (I-J) <sup>a</sup>	Perceived User Moderation (I-J) <sup>a</sup>	Perceived Moderator Moderation (I-J) <sup>a</sup>	Perceived System Moderation (I-J) <sup>a</sup>
Familiar	P	PC	-.091	-.074	-.114	-.050	-.128
		PCK	-.160*	-.122	-.135	-.079	-.118
	PC	PCK	-.069	-.048	-.022	-.028	.010
Unfamiliar	P	PC	-.097	-.116	-.112	-.146	-.120
		PCK	-.079	-.172†	-.296***	-.222*	-.241**
	PC	PCK	.017	-.056	-.184†	-.076	-.121

<sup>a</sup>:mean difference; \*\*\*:  $p < .001$ ; \*\*:  $p < .01$ ; \*:  $p < .05$ ; †:  $p < .1$

## 4.6 Discussion

This study provides a conceptualized understanding of the perception biases in content moderation, specifically elucidating the effects of review information comprehensiveness, user roles, and content familiarity. The summary of hypotheses testing results is reported in Table 13.

**Table 13: A Summary of Hypotheses Testing Results**

Hypotheses	Content Review Time	Mental Workload	Perceived Confidence	Perceived Compliance	Perceived User Moderation	Perceived Moderator Moderation	Perceived System Moderation
H1: The effect of adding user comments	P	N	P	N	P	P	P
H2: The effect of adding both user comments and karma scores	Y	P	P	P	P	P	P
H3: The moderating effect of user roles on content review with user comments	Y	Y	N	N	Y	Y	Y
H4: The moderating effect of user roles on content review with both user comments and karma scores	N	N	N	N	Y	Y	Y
H5: The moderating effect of content familiarity on content review with user comments	N	N	N	N	N	N	N
H6: The moderating effect of content familiarity on content review with both user comments and karma scores	N	N	N	Y	Y	Y	Y

Y = Supported; P = Partially Supported; N = Not Supported

To address RQ 2.1, this study reveals that review information comprehensiveness not only increases the content review time and mental workload, along with a higher perception of confidence and compliance in content moderation but also enhances the perceived content moderation intervention expectations towards users, moderators, and systems. Accordingly, platforms need to consider the increased time and cognitive effort required for reviewing content with rich online content. This may necessitate additional staffing, more robust training programs, and perhaps the development of specialized roles focused on handling complex cases. To mitigate the increased cognitive load, platforms might invest in advanced technological aids, such as AI-powered tools that can pre-filter or flag content needing detailed human review. These tools can handle preliminary assessments, allowing moderators to focus on more complex, nuanced decisions. Understanding that review information comprehensiveness can also inform the design of automated systems, which can be fine-tuned to prioritize UGC with extensive information to

make decisions, which can streamline the moderation process and ensure that complex cases are addressed efficiently. Technology can help in organizing and presenting UGC in a more digestible format, reducing the cognitive burden on reviewers. Additionally, platforms can use insights from UGC to refine their content policies. Clear guidelines on what constitutes a violation, supported by examples from reviewed UGC, can help make accurate moderation decisions. This can lead to better alignment between community expectations and platform enforcement practices.

To address RQ 2.2, this study demonstrates that content review time and mental workload increase for users but not for moderators. Users typically lack the specialized training and experience that moderators possess, but moderators are often trained to quickly identify key issues and make decisions efficiently, using well-developed heuristics and guidelines. The presence of additional review information, such as user comments and karma scores, increases the cognitive demands on individuals who are not accustomed to processing such data [184]. Users, unlike moderators, do not have pre-established schemas for quickly integrating this information, resulting in higher cognitive load and extended review time. It calls attention to developing user interfaces that simplify the presentation of UGC, such as using visual aids such as highlights, summaries, and categorization, which helps reduce cognitive load for users, making the review process more efficient and less mentally taxing. AI-driven tools that provide automated summaries or highlight important information can also assist users in quickly understanding the core points of the UGC, thus decreasing the time and mental effort required for review. Additionally, users progressively enhance their expectations for content moderation interventions towards their peers, moderators, and systems as more review information is incorporated during the review of UGC. With the increased availability of review information, users perceive the content moderation process to be more transparent and fair. This perception of fairness and consistency raises their expectations that

all involved parties—peers, moderators, and systems—will apply the same standards and thoroughness in their moderation efforts. Consistency in applying rules and guidelines is crucial for maintaining trust in the moderation process [30]. Furthermore, providing users with more review information empowers them to participate actively in the content moderation process. When users feel their voice is heard and their input is valued, their expectations for effective and accurate moderation interventions by peers, moderators, and systems increase [34]. Implementing feedback systems where users can receive insights on their moderation decisions and understand how they align with those of moderators and automated systems is beneficial. This feedback not only helps users refine their expectations and improve their content review skills but also creates a sense of ownership and understanding of the moderation process, leading to more realistic expectations.

To address RQ 2.3, this study demonstrates that, when reviewing unfamiliar UGC, incorporating more review information during content moderation enhances the perception of compliance and shapes content moderation intervention expectations towards users, moderators, and systems. When individuals review unfamiliar UGC, they often pay closer attention to detail, which underscores the importance of specific guidelines and their application. This heightened scrutiny arises because reviewers approach unfamiliar content without pre-existing biases or preconceived notions, making them more reliant on community feedback—such as user comments and karma scores—to form opinions about content compliance. These user comments and karma scores serve as community signals that indicate the general acceptability and quality of the content. In contrast, familiarity with UGC can lead to subjective judgments influenced by personal experiences or preferences. Evaluating unfamiliar UGC with the support of community feedback encourages a more objective assessment based on collective input rather than personal biases. The



findings of this study advocate for reviewers to approach unfamiliar UGC without preconceived biases, emphasizing the value of community feedback in forming impartial assessments. Furthermore, the study suggests that social media platforms should provide clear guidelines to assist reviewers in analyzing contextual information. This includes looking for detailed explanations and consensus within the UGC to better understand the community's perspective on content compliance. These guidelines will help reviewers to accurately interpret the community's expectations and apply moderation practices consistently.

## **CHAPTER 5: USER-MODERATOR COLLABORATIVE (UMCollab) CONTENT MODERATION**

### **5.1 Introduction**

Content moderation is a highly nuanced and contextually situated task given its inherently context-dependent nature. The determinations of what qualifies as acceptable or objectionable UGC are intricately tied to the norms and standards within a given online community [154], [155], [156]. These criteria of acceptability are not isolated; instead, they are influenced by existing standards and can exhibit flexibility and susceptibility to change based on the dynamics of platform trend shifts and users' acceptability [101]. The issue of what, if any, UGC should be filtered or entirely removed has consistently been a topic of fervent societal discussion [101].

User engagement creates inherent value in the realm of content moderation as it empowers moderators to gather feedback from the community regarding objectionable or inappropriate UGC [39]. The importance of user engagement lies in its dual role as both a process and a result of user interactions, offering significant potential for harnessing social networking methods to extract insights from UGC (such as users' interactions and comments, content creditability, and stance [185]). Nonetheless, users' moderation choices can be swayed by their personal prejudices and subjective outlooks, leading to irregular adherence to content guidelines and the likelihood of favoring certain viewpoints [40]. In addition, users might lack the expertise or training needed to reliably spot certain forms of troublesome content. This situation could lead to either the removal of legitimate UGC or the retention of harmful UGC [60].

Moderators frequently possess a significant track record of involvement within an online community. Generally, moderators receive training before overseeing UGC [31]. Through this training, moderators build knowledge in their area and develop a deeper understanding of the

subjects discussed in the UGC. This expertise allows moderators to understand the context, notice nuances, and make decisions that align with the guidelines [42], [43]. In addition, moderators also often take part in creating community rules and standard procedures. These rules provide a structure for consistent decision-making and help ensure moderators follow a uniform approach to content moderation interventions[126]. However, there is a risk that moderators might reinforce existing biases and create spaces where only certain opinions are allowed, limiting diverse viewpoints.

Given the complementary role of users and moderators in content moderation, it would be promising to integrate viewpoints from both users and moderators. However, such a collaborative approach remains scarce. This research gap is attributed to a few significant research challenges: 1) there's little public information available about the moderators who perform content moderation, which primarily resulted from the concerns about privacy and public images of online users [57]; 2) user engagement is a complex construct, encompassing not only textual contents but also user interactions and various factors that influence the topic of discussion, such as content quality and the stance toward different online users; and 3) user engagement and moderator expertise change over time as online communities may evolve with shifting interests and new trend adaptation, such dynamics make the content moderation more challenging. Therefore, the development of an effective content moderation framework that seamlessly integrates user engagement with moderator expertise presents a demanding yet potentially rewarding research undertaking.

To fill the research gap, this study aims to answer the following research questions by introducing UMCollab - a user-moderator collaborative framework for content moderation. The UMCollab framework employs a dual approach. Firstly, it dynamically integrates user engagement through graph learning to map users' interactive discussions concerning UGC, further enhancing

this integration with metrics of UGC creditability and stance. Secondly, the framework incorporates the domain knowledge of moderators by embedding historical content moderation decisions, made per community rules, into deep learning models to facilitate content moderation.

RQ 3: Can the effectiveness of content moderation be improved by collaboratively integrating user engagement and domain knowledge into a deep learning-based framework?

RQ 3.1: How does each component of the collaborative framework contribute to the effectiveness of content moderation?

## 5.2 Related Work

Content moderation, particularly on the internet and social media platforms, often involves a collaborative effort between humans and machines rather than either of them alone. Previous studies (e.g., [50], [53], [54], [55]) have found that using both the content of a post and its entire history of user engagement (i.e., all comments) is effective in detecting the legitimacy of information. For example, Serrano et al. [53] used online users' comments to predict COVID-19-related misinformation videos on YouTube; Guo et al. [49] incorporated both word embeddings and emotion embeddings of user comments within a news article discussion; and Raza and Ding [54] proposed a transformer-based approach that combined news content (e.g., posts on the news, news sources) and social contexts (e.g., user creditability) to facilitate fake news detection. However, incorporating the entire history of user engagement with explicit information structure is not cost-effective and can significantly increase the computational overhead. Kaliyar et al. [55] applied an LSTM-based model to acquire knowledge about the content of news articles and used the Clauset-Newman-Moore algorithm to configure user-to-user connections within a user community while discerning the veracity of news content. In addition, an attention model based

on BiLSTM combined word embeddings from user posts and various social features (such as user profile details like follower count and registration time). This fusion at the post level aimed to detect rumors at the event level, which includes both the original post and its reposts [56]. Additionally, a hierarchical attention model utilized both bottom-up and top-down propagation tree structures to combine user opinions from various comments associated with a claim to jointly verify rumors and detect stances [50].

There are a few domain knowledge-based approaches to content moderation (e.g., [186], [187], [188]); however, these methods typically use data from a single target community or focus on a specific type of content (e.g., hate speech, fake news, and rumors). As a result, these approaches face limitations when being generalized to cross-community content moderation. In contrast, cross-community moderation [57] exhibits greater robustness in tackling data scarcity and imbalance by utilizing a substantial corpus of prior moderator decisions through an ensemble of classifiers. This approach achieved a preferable performance in terms of accuracy and recall in identifying users' comments that require removal.

While integrating the comments of online users and the domain knowledge of moderators has demonstrated effectiveness in detecting a variety of types of illegitimate UGC, these models are predominantly task-oriented and do not reflect the broader practices of content moderation. This broader practice includes a range of interventions, such as ensuring content compliance with community rules, evaluating content relevance within an online community, assessing content quality, and considering online users' opinions on UGC. Additionally, the social contexts of online users are often not accessible across different social media platforms, particularly when users opt to keep their profiles private. Moreover, the two most pertinent content moderation studies [57], [105] focus exclusively on comment removal, overlooking the comprehensive content moderation

ecosystem. Social media interactions unfold over time, and moderators' expertise is honed through active engagement in community activities and the implementation of various interventions. Consequently, there is a notable gap in the model development that effectively incorporates the dynamics of user engagement and the domain knowledge of moderators for making comprehensive moderation decisions.

### **5.3 Theoretical Foundations of User-Moderator Collaboration**

The advocates of civil society and academic human rights argue that fully automated decision-making systems, devoid of human-in-the-loop involvement, pose significant risks [189]. Drawn on the diffusion of innovations theory [190], online user discussions often reveal new trends, memes, or challenges. The dynamics of user conversations would mirror the subject being discussed within UGC. In addition, the online disinhibition effect theory [191] also states that individuals may exhibit altered behavior and participate in more intense online discussions because of their perceived anonymity and the absence of social cues. This is particularly true in anonymous social media platforms such as Reddit. Users' positions on a topic can either remain steadfast or transform, influenced by the actions of those in their online circles, thereby fostering the propagation of specific viewpoints in digital discussions [192]. According to the expertise and expert performance theory [193], individuals (e.g., moderators) with extensive domain knowledge make better decisions within their specific domain due to their deep understanding and experience. In addition, the dual-process theory [194] also posits that decision-making involves two systems, one being intuitive and fast and the other being analytical and deliberate, and domain knowledge can influence the interplay between these two systems. The former pertains to predefined automated content moderation systems and mechanisms created by experienced moderators, while

the latter mirrors how these moderators make content moderation decisions by drawing upon their contextual understanding of the domain [195].

Regarding content moderation practices, the cognitive assemblages under discussion comprise a blend of human and algorithmic interactions, operating at various reflective, temporal, and perceptual levels [196]. By leveraging user discussions and the domain knowledge of moderators, the content moderation system can become more flexible and adaptive to the dynamic nature of online content. It can better address new challenges and adjust moderation criteria based on user feedback and community standards. This continuous exposure to real-world examples allows the system to improve over time and make more informed decisions.

#### 5.4 The UMCollab Framework

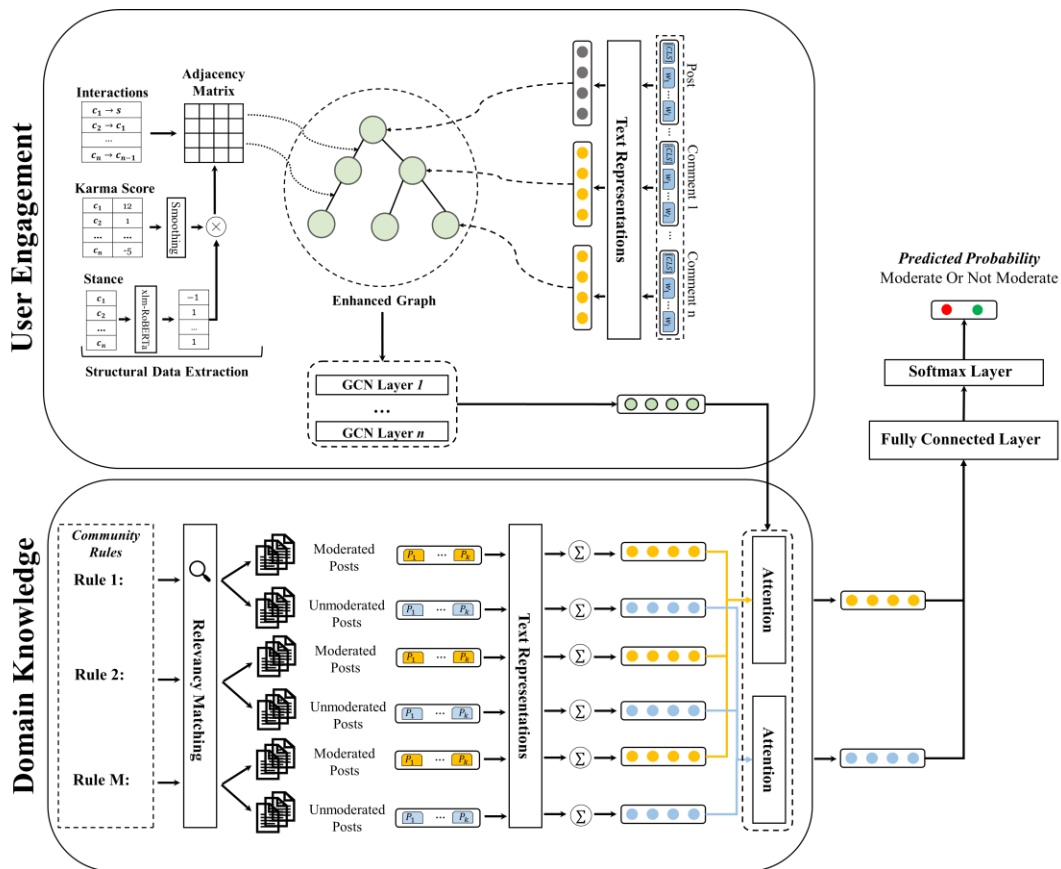


Figure 12: The Framework of UMCollab for Content Moderation

In this section, I introduce the framework of UMCollab and its key design artifacts for content moderation, which is depicted in Figure 12. Specifically, the UMCollab framework utilizes a dual approach by integrating user engagement via graph learning and incorporating moderators' domain knowledge by embedding historical content moderation decisions, made per community rules.

#### 5.4.1 User Engagement

One of the primary objectives of this research is to incorporate graph-based information of user engagement based on users' comments along with their content creditability and stance to facilitate content moderation. To construct a graph, I first formulate user engagement with a UGC  $U_r$  as a graph,  $G = (V, E, W)$ , where the node set  $V$  comprises the post content (i.e., including post title and post body) and its associated comments; the edge set  $E$  represents the observed interactions among the nodes in  $V$ , and  $W$  represents the weights of edge set  $E$ . I define an interaction as a comment in response to the original post or another user's comment and leverage GCN [197] to learn post representations from a user engagement graph. Each component of the graph is described in detail as follows.

**Node Representation:** Given that the use of a word embedding matrix is massive and space-consuming, I add a special classification token (i.e.,  $[CLS]$ ) at the beginning of a word sequence and use the final hidden state corresponding to this token as the aggregate sequence representation for the entire word sequence. In addition, unlike the word embedding vectors, the vectors for the  $[CLS]$  tokens can be directly used to represent post content and/or user comments separately for a specific UGC.

**Edge Representation:** In view that different user interactions are not equally important to the original posts/comments, I introduce weights to improve the edge representations in the graph.



Specifically, I estimate the weights of the edges based on two user engagement factors of the involved nodes: creditability and stance.

*Creditability*: I assume that more credible comments have a stronger impact on user engagement within a specific UGC. In this study, I leverage the karma score, referring to the difference between upvotes and downvotes that each comment received, as a proxy of the creditability of the user comment, which is shown in Equation 6.

$$W_{credibility} = \lambda + S_{Karma} \quad (6)$$

Where  $S_{Karma}$  denotes the karma score of a user comment, and  $\lambda$  is a smoothing factor. I introduce the smoothing factor to address the issue of missing or null values. The value of  $\lambda$  is set to be 1 in this study.

*Stance*: I assume that the stances or opinions as reflected in user comments have impacts on user engagement within a specific UGC. In this study, I leverage the xlm-RoBERTa model [198], which is pre-trained with the SemEval-2016 dataset [199], to generate the stance score  $W_{stance}$  for an edge in Graph  $G$ , and the score ranges from -1 ('against') to 1 ('favor') with less likelihood resulting in neutral (i.e., 0).

User comments with more prominent positions should receive higher weights. To derive edge weights, I use multiplication to fuse the creditability score  $W_{credibility}$  and stance score  $W_{stance}$ , as shown in Equation 7.

$$W_{edge} = W_{credibility} \times W_{stance} \quad (7)$$

The advantage of multiplying creditability and stance is to use the karma score to correct the stance of user comments. For example, a child comment with a stance of 'favor' is positive towards its parent comment. If the child comment has a negative karma score, I can infer that the stance of the child comment is discrepant with the majority of community members and should be

corrected to be ‘against’, which is negative. By multiplying the creditability and stance values, I can maintain or correct the stance of user engagements at the same time for edge weight generation.

**User Engagement Aggregation:** By aggregating the information of neighboring nodes, I can update the representations of node  $i$  in graph  $G_r$  based on Equation 8,

$$h_i^{(k)} = W_1^{(k)} h_i^{(k-1)} + W_2^{(k)} \sum_{j \in \mathcal{N}(i)} W_{edge\ i,j} \cdot h_j^{(k-1)} + b^{(k)} \quad (8)$$

where  $W_{edge\ i,j}$  denotes the edge weight from node  $j$  to node  $i$ ;  $\mathcal{N}(i)$  denotes a set of neighbors of node  $i$ ; and  $h_i^{(k-1)}$  and  $h_j^{(k-1)}$  are the representations of node  $i$  and node  $j$  at the  $(k-1)^{th}$  GCN layer, respectively.  $W_1^{(k)}$  and  $W_2^{(k)}$  are the learnable weights for the  $k^{th}$  GCN layer related to node  $i$  and its neighbors, and  $b^{(k)}$  is the bias for the  $k^{th}$  GCN layer. It can be observed from Equation 8 that the representation of node  $i$  at the  $k^{th}$  GCN layer is derived as a weighted average of its own representation and the representations of its neighbors at the  $(k-1)^{th}$  layer. By applying the stacked GCN layers, I can enhance the post representation  $h_u$  through user engagement by propagating the information of user engagement that includes user comments and their associated karma scores and stances.

#### 5.4.2 Moderators’ Domain Knowledge

Another primary objective of this research is to learn the domain knowledge of moderators based on their prior content moderation decisions in accordance with community rules/policies.

**Community Rule Representation:** I employ community rules and the Best Match 25 (BM25) algorithm [200] to identify the top 10 most relevant moderated and unmoderated posts corresponding to each community rule. The BERT encoder [62] is utilized to obtain the embeddings of the identified posts. Given that BM25 relevance scores indicate the importance of the most relevant posts in representing community rules, I select the top 10 most relevant posts and apply a weighted sum to them. This weighting uses BM25 normalized relevance scores to

generate community rule representations  $r^0$  and  $r^1$ , respectively. The rationale behind this approach is threefold: first, social media posts provide a practical context in which community rules are applied, aiding models in understanding the real-world application of these rules, thus making them more tangible and relatable; second, social media platforms host a diverse array of content, ensuring that each community covers a broad spectrum of scenarios; third, social media posts are associated with a history of moderated and unmoderated interactions, offering valuable insights into the enforcement and effectiveness of these rules over time. Sample exemplary community rules and their corresponding moderated and unmoderated posts are illustrated in Appendix F – Table 20.

**Domain Knowledge Representation:** To determine which specific rule(s) dictate a UGC moderation status, I leverage the attention mechanism to learn two types of domain knowledge in relation to the post  $h_u$ : moderated domain knowledge  $M^0$  and unmoderated domain knowledge  $M^1$ . Specifically, the similarity between the post  $h_u$  and the domain knowledge  $M^0$  and  $M^1$  is calculated separately using the dot product. Subsequently, the SoftMax function is employed to normalize these similarity scores as shown in Equations 9 and 10.

$$a_i^0 = \frac{\exp(w_i[r_i^0 \cdot h_u] + b)}{\sum_{r_i^0 \in M^0} \exp(w_i[r_i^0 \cdot h_u] + b)} \quad (9)$$

$$a_i^1 = \frac{\exp(w_i[r_i^1 \cdot h_u] + b)}{\sum_{r_i^1 \in M^1} \exp(w_i[r_i^1 \cdot h_u] + b)} \quad (10)$$

The attention mechanism offers several advantages in learning domain knowledge. Firstly, it enhances the accuracy of similarity measurements by focusing more on community rules that are similar or relevant to the UGC and disregarding irrelevant ones. Secondly, unlike fixed-size windows or traditional pooling techniques, attention mechanisms can adjust their range of influence. In this study,  $w_i$  is a learnable weight that modulates the influence of similarity to

various community rules within the input sequence, assigning greater importance to similar UGC and lesser importance to dissimilar ones. This mechanism enables the model to capture more nuanced similarities. Finally, I obtain the UGC representations  $h_k^0, h_k^1$  that are enhanced by user engagement and domain knowledge, as shown in equations 11 and 12.

$$h_k^0 = \sum_1^{|M^0|} a_i^0 r_i^0 \quad (11)$$

$$h_k^1 = \sum_1^{|M^1|} a_i^1 r_i^1 \quad (12)$$

#### 5.4.3 Prediction of Moderation Decisions

I design a fully connected layer to generate a binary classification result to determine whether or not a UGC should be moderated based on two different aspects of UGC representation  $h_k^0$  and  $h_k^1$ , as shown in equation 13.

$$\hat{y} = \text{softmax}(W_o[h_k^0; h_k^1] + b_o) \quad (13)$$

Where  $\hat{y}$  is the classification result of a UGC being moderated or not;  $\text{softmax}(\cdot)$  is the softmax function;  $W_o \in \mathbb{R}^{2 \times 2d}$  denote the learnable weights;  $d$  is the dimension of the UGC representations  $h_k^0$  and  $h_k^1$ , and  $b_o$  is the bias for the fully connected layer.

I choose the binary cross-entropy as the loss function, where  $T$  denotes the set of training instances, and  $\mu\|\Delta\|$  is parameter-specific regulation hyper-parameters to prevent overfitting. By minimizing the loss value calculated by equation 14, we train the model to generate classification results.

$$\mathcal{L} = -\sum_{y \in T} y \log(\hat{y}) + \mu\|\Delta\| \quad (14)$$

## 5.5 Experiments

In this section, I present the evaluation procedure, which encompasses data collection and preprocessing, baseline model selection, and model performance comparisons along with statistical analyses.

### 5.5.1 Data Collection and Preprocessing

I utilized the same dataset as referenced in Chapter 4, Section 4.4.3. Given the wide variation in the number of comments associated with each post and the highly right-skewed distribution of comments, I performed further filtering by establishing a lower bound of 1 comment and an upper bound of 100 comments per post. This facilitates the extraction of graph-based user engagement information. After preprocessing the data with these comment number restrictions, the dataset was reduced to 75,792 posts. Considering that most UGC on social media platforms is not subject to moderation, I performed under-sampling on the data with unmoderated posts. The final dataset comprises 4,806 posts from the gaming domain, 1,152 from fashion, 3,460 from health, and 5,718 from sports. In each domain, the unmoderated and moderated samples are evenly distributed.

### 5.5.2 Baseline Models

Given that this research focuses on text-based UGC for content moderation, I selectively choose four state-of-the-art baseline models for a comparative analysis against the UMCollab framework. In particular, Classifier97 [201] trains a separate classifier for each community allowing for tailored content moderation that accounts for the unique characteristics of an online community. CB-BLSTM [112] uses BiLSTM architecture which is effective for text sequence as it can capture context from both directions (forward and backward) in a sequence, making it suitable for understanding the nuances in UGC. AbuDL [113] used RNN architecture is well-suited

for handling sequential data like text, allowing the model to maintain context over sequences. Despite that HATE-L2 [102] used a traditional text presentation method followed by logistic regression, it shows promising performance in the detection of hate speech, which is one of the major bans in online communities. I discuss each baseline model detail as follows.

- Classifier97 [201]: the model comprises 97 neural network binary classifiers, each trained on an individual online community. Each classifier features a four-layer neural network architecture, with an embedding layer and dense layers to make a binary prediction for content moderation.
- CB-BLSTM [112]: this model employs a BiLSTM architecture, which retains two separate input states to handle text sequences of posts. The BLSTM model consists of 200 neurons in the first layer and 400 neurons in the second layer. It includes three dense layers with 128, 64, and 32 neurons, respectively.
- AbuDL [113]: this deep learning architecture leverages available metadata (e.g., likes, favorites) and the raw text of user posts for abusive behavior detection. An RNN architecture is employed, specifically utilizing word-level RNNs for the raw text input. It is worth noting that the metadata was not taken into account in this study due to limitations in data availability.
- HATE-L2 [102]: this model leverages logistic regression with L2 regularization. The input features are pre-processed by lowercasing and stemming each post's content, followed by creating TF-IDF-based n-gram representations.

### 5.5.3 Performance Evaluation

I employ a set of widely used evaluation metrics to measure the predicted results, including accuracy, precision, recall, and F1 score. Additionally, I utilize 10-fold cross-validation to evaluate

model performance, with a 90/10 data split for training and testing. The learning rate is set to 0.001; the dimension of the BERT embedding is 768; and the dropout rate is 0.2 for all models. All models are optimized using the Adam optimizer.

To evaluate the effectiveness of the UMCollab framework in comparison to the baseline models, I conduct paired samples t-tests for model performance comparisons within each domain. Furthermore, I perform an ablation analysis on UMCollab by removing each component from the framework, including user engagement, domain knowledge, and the creditability and stance of user engagement, followed by paired samples t-tests for model comparisons. In addition, I also perform multiple comparisons to evaluate the performance deterioration among the ablated models within each domain.

## 5.6 Results

I report experiment results, including model performance comparisons and ablation analysis in this section.

### 5.6.1 Model Performances

The performances of the UMCollab and baseline models for each domain are reported in Figure 13, and the results of paired samples t-tests for model performance comparisons are reported in Table 14. In the fashion domain, UMCollab outperforms all baseline models in terms of accuracy ( $p < .001$ ), precision ( $p < .001$ ), recall ( $p < .05$ ), and F1 score ( $p < .001$ ). In the game domain, UMCollab exceeds all baseline models in accuracy ( $p < .001$ ), precision ( $p < .001$ ), recall ( $p < .05$ ), and F1 score ( $p < .01$ ). In the health domain, UMCollab demonstrates superior performance over all baseline models in accuracy ( $p < .001$ ), precision ( $p < .001$ ), recall ( $p < .05$ ), and F1 score ( $p < .01$ ). In the sports domain, UMCollab surpasses all baseline models in

accuracy ( $p < .001$ ), precision ( $p < .01$ ), and F1 score ( $p < .01$ ), with a marginal improvement in recall ( $p < .1$ ).

### 5.6.2 Ablation Study

The results of an ablation study are presented in Figure 14, and the results of the paired sample t-tests and multiple comparisons of model performance are reported in Tables 15 and 16 respectively. The findings indicate that each of the model components significantly (at least  $p < .05$ ) enhances model performance in content moderation, as evidenced by the marked decrease in performance across all four metrics when removing any individual component from the UMCollab model. Among the three model components, the removal of user engagement results in the most significant performance decline, followed by domain knowledge, and then creditability and stance.

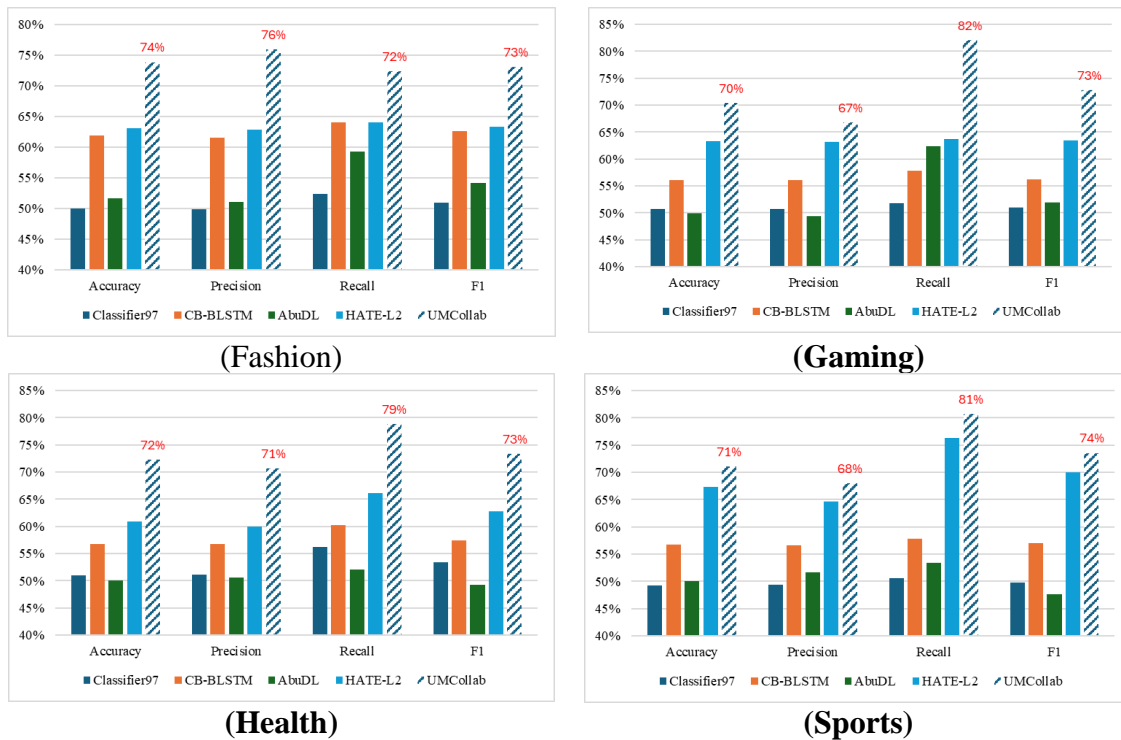


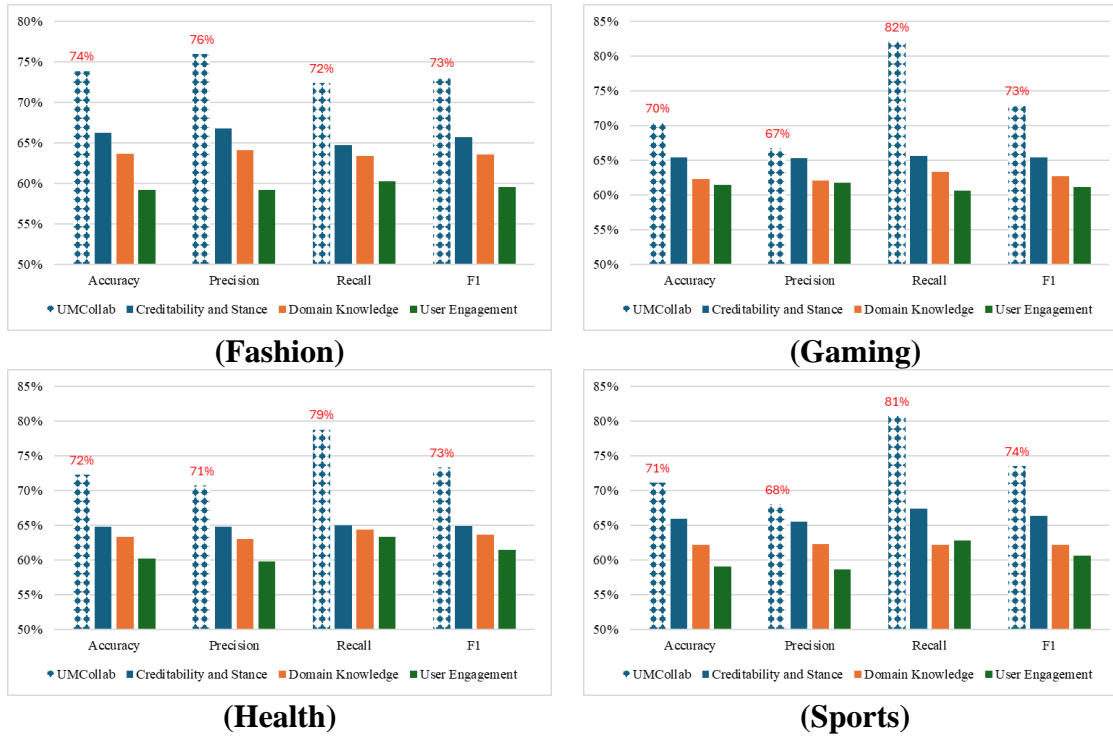
Figure 13: Model Performances of the UMCollab and Baseline Models



**Table 14: Performance Comparisons between the UMCollab and the Baseline Models**

Domain	Baseline Models	Accuracy	Precision	Recall	F1
Fashion	Classifier97	(-12.850)***	(-10.297)***	(-4.731)***	(-8.719)***
	CB-BLSTM	(-13.417)***	(-5.596)***	(-1.911)*	(-7.112)***
	AbuDL	(-16.600)***	(-12.120)***	(-2.733)*	(-7.714)***
	HATE-L2	(-9.205)***	(-4.624)***	(-2.234)*	(-6.642)***
Game	Classifier97	(-18.072)***	(-18.971)***	(-6.634)***	(-11.338)***
	CB-BLSTM	(-10.594)***	(-13.238)***	(-4.343)***	(-6.075)***
	AbuDL	(-22.080)***	(-18.427)***	(-2.313)*	(-4.638)***
	HATE-L2	(-5.919)***	(-5.814)***	(-3.617)**	(-4.031)**
Health	Classifier97	(12.867)***	(-7.796)***	(-4.594)***	(-9.355)***
	CB-BLSTM	(-10.761)***	(-6.419)***	(-3.220)**	(-7.487)***
	AbuDL	(-9.924)***	(-7.834)***	(-4.047)**	(-6.125)***
	HATE-L2	(-6.128)***	(-4.719)***	(-2.510)*	(-3.867)**
Sports	Classifier97	(-32.078)***	(-16.448)***	(-7.591)***	(-17.746)***
	CB-BLSTM	(-44.921)***	(-12.195)***	(-7.881)***	(-15.226)***
	AbuDL	(-102.021)***	(-8.451)***	(-2.978)**	(-4.297)**
	HATE-L2	(-6.185)***	(-3.218)**	(-1.783)†	(-4.907)***

\*\*\*:  $p < .001$ ; \*\*:  $p < .01$ ; \*:  $p < .05$ ; †:  $< .1$ ; each numeric value within parentheses represents the mean difference between the UMCollab and the baseline models.

**Figure 14: Performances Comparisons of the Ablated Models vs. the UMCollab Model**

**Table 15: Performance Comparisons between the UMCollab Model and Ablated Models**

Domain	Ablated Models	Accuracy	Precision	Recall	F1
Fashion	Creditability and Stance	(-.076)***	(-.091)**	(-.077)*	(-.074)***
	Domain Knowledge	(-.102)***	(-.119)**	(-.091)*	(-.095)***
	User Engagement	(-.147)***	(-.168)***	(-.122)**	(-.0135)***
Game	Creditability and Stance	(-.051)**	(-.014)*	(-.164)**	(-.074)**
	Domain Knowledge	(-.081)***	(-.046)***	(-.187)**	(-.102)**
	User Engagement	(-.090)***	(-.050)***	(-.214)**	(-.117)***
Health	Creditability and Stance	(-.075)***	(-.059)**	(-.138)*	(-.084)**
	Domain Knowledge	(-.090)***	(-.076)**	(-.144)*	(-.096)**
	User Engagement	(-.121)***	(-.109)***	(-.154)**	(-.118)***
Sports	Creditability and Stance	(-.052)***	(-.025)*	(-.134)***	(-.072)***
	Domain Knowledge	(-.089)***	(-.057)***	(-.186)***	(-.114)***
	User Engagement	(-.120)***	(-.093)***	(-.179)***	(-.129)***

\*\*\*:  $p < .001$ ; \*\*:  $p < .01$ ; \*:  $p < .05$ ; each numeric value within parentheses represents the mean difference between an ablated model and the full UMCollab model.

**Table 16: Multiple Comparisons of Performance Deterioration of the Ablated Models**

Domain	Ablated Models		Accuracy	Precision	Recall	F1
	(I)	(J)	(I-J) <sup>a</sup>	(I-J) <sup>a</sup>	(I-J) <sup>a</sup>	(I-J) <sup>a</sup>
Fashion	UE	CS	(-.070)***	(-.076)***	(-.045)*	(-.061)***
		DK	(-.044)**	(-.049)*	(-.031)†	(-.039)**
	DK	CS	(-.026)*	(-.027)*	(-.014)	(-.022)*
Game	UE	CS	(-.039)***	(-.036)**	(-.049)***	(-.043)***
		DK	(-.009)	(-.004)	(-.027)†	(-.016)†
	DK	CS	(-.031)***	(-.032)**	(-.023)*	(-.028)***
Health	UE	CS	(-.046)***	(-.049)**	(-.017)*	(-.034)***
		DK	(-.031)*	(-.032)*	(-.010)	(-.022)*
	DK	CS	(-.015)†	(-.017)†	(-.006)	(-.012)†
Sports	UE	CS	(-.068)***	(-.068)***	(-.045)**	(-.058)***
		DK	(-.031)*	(-.036)*	(.007)	(-.016)†
	DK	CS	(-.037)**	(-.032)**	(-.052)**	(-.042)***

<sup>a</sup>: mean difference; \*\*\*:  $p < .001$ ; \*\*:  $p < .01$ ; \*:  $p < .05$ ; †:  $< .1$ ; UE: user engagement; CS: creditability and stance; DK: domain knowledge

## 5.7 Discussion

As the information of UGC becomes richer, the method for automated content moderation should leverage multi-aspect of online interactive information to capture a broader range of tactics for content moderation interventions. This research underscores the significant benefits of integrating the dynamics of user engagement and the domain knowledge of moderators into deep learning models to improve the effectiveness of content moderation.

To answer RQ 3, the research findings indicate that the UMCollab framework significantly outperforms the baseline models across all evaluation metrics across different domains. To answer RQ 3.1, this research shows that user engagement, domain knowledge, and credibility and stance greatly contribute to the improvement of model performance in content moderation, with user engagement contributing the most.

The research implications of a user-moderator collaborative approach for content moderation are extensive and transformative. This approach promises significant advancements in AI, enhancing our understanding of online behavior and social norms within online communities. It also aims to improve moderation efficiency and fairness, as well as inform policy and governance. Specifically, the collaborative framework provides empirical evidence on the effectiveness of hybrid intelligence systems in content moderation. This can lead to the development of more sophisticated machine and deep learning models capable of understanding and processing the nuances of UGC. Moreover, the collaborative framework offers deeper insights into online user interactions and reactions, which can be used to identify patterns of behavior that lead to illegitimate content and to assess the impact of various moderation strategies on users' online interactive behavior. Furthermore, this research demonstrates the feasibility of designing effective interfaces for collaboration between users and moderators, which can enhance the

usability and effectiveness of content moderation tools. By encouraging constructive user participation in moderation and balancing the workload between automated systems and human input, this approach fosters a more engaged and responsible community. Last but not least, this area of research encourages collaboration among computer scientists, sociologists, psychologists, and legal experts, fostering interdisciplinary innovation on a large scale. A comprehensive collaborative model allows content moderation systems to be assessed from technical, social, and ethical dimensions, ensuring a holistic approach to managing online communities.

The practical implications of a deep learning-based user-moderator collaboration model for content moderation are far-reaching. This collaborative framework offers dynamic monitoring capabilities for UGC, enhancing both the efficiency and accuracy of content moderation. The framework is adaptable to new types of UGC and emerging challenges, maintaining ongoing relevance and effectiveness by continuously learning from user feedback and moderator decisions over time. Additionally, the framework provides scalability to handle large volumes of UGC, making it feasible to moderate content with multidimensional information derived from online communities. By automating routine moderation tasks, human moderators can concentrate on more complex cases, optimizing human resource utilization and reducing operational costs for social media platforms. Furthermore, the involvement of users in the moderation process through reporting mechanisms and feedback loops empowers them to contribute to the community's health, fostering a sense of fairness and trust among users.

## CHAPTER 6: CONCLUSIONS

This study introduces a user-moderator collaborative framework, UMCollab, which leverages user engagement dynamics and domain knowledge to enhance deep learning models for automated content moderation. It draws on an extensive investigation of perceived biases in content moderation, examining review information comprehensiveness, user roles, and content familiarity. The study also empirically explores the effectiveness and efficiency of user engagement in content moderation. The findings contribute both theoretically and methodologically to the field of content moderation research and offer practical implications for online users, moderators, and social media platforms.

### 6.1 Summary

Chapter 3 investigates the effectiveness and efficiency of user engagement in content moderation and presents three primary findings in content moderation research. Firstly, integrating the directness of user engagement produces the most favorable outcomes in content moderation, closely followed by temporal engagement. These findings underscore the importance of immediacy in user engagement for assessing the quality and relevance of social media posts accurately, thereby enhancing the overall content moderation effectiveness. Secondly, the study reveals an incremental improvement in model performance from the first to the third user engagement, with no clear trend observed in subsequent user engagement. This suggests that utilizing three user engagements suffices to achieve satisfactory content moderation performance. Thirdly, the study provides insights into the efficiency of user engagement for timely content moderation, ensuring robust performance under time constraints.

Chapter 4 examines the effect of review information comprehensiveness on content moderation, which is further moderated by user roles and content familiarity. The results present

empirical evidence for perception biases inherent in this process. The study underscores the importance of incorporating comprehensive review information to enhance transparency and ethical decision-making in the moderation of UGC. Specifically, it shows that users reviewing UGC with comments experience longer review time and higher mental workloads compared to moderators. Both users and moderators perceive a higher expectation for content moderation interventions from their peers and systems. Additionally, the dissertation reveals that both users and moderators reviewing unfamiliar UGC show a stronger perception of compliance and expectations for content moderation interventions directed toward their peers and systems compared with reviewing familiar UGC.

Chapter 5 introduces a novel user-moderator collaborative framework to facilitate automated content moderation and emphasizes the significant benefits of incorporating user engagement dynamics and domain knowledge into deep learning models to enhance the effectiveness of content moderation. The research findings demonstrate that the model based on the UMCollab framework outperforms existing baseline models. Moreover, the study underscores that factors such as user engagement, domain knowledge, and credibility and stance are crucial for improving model performance in content moderation, with user engagement being the most significant contributor to these improvements.

## **6.2 Research Contributions**

This research makes multifaceted contributions to the information system literature.

Firstly, this study provides pioneering empirical evidence on the impact of user engagement characteristics on content moderation in social media, addressing a significant gap in the existing literature. By incrementally examining the number of comments on the effectiveness of content moderation for the first time, this research sheds light on the degree of user engagement

in moderating UGC. Additionally, the study offers valuable insights into the impact of user engagement on the efficiency of content moderation, deepening our understanding of how user interactions influence the timeliness of content moderation decisions. These insights not only enhance the theoretical development for the incorporation of user engagement for content moderation research but also inform practical strategies for optimizing user involvement in upholding online community rules.

Secondly, this study addresses a significant gap in the literature by advancing the theoretical framework of perceived biases in content moderation and guiding the development of fairer and more effective moderation practices. It provides empirical evidence of perceived biases in content moderation by examining review information comprehensiveness, user roles, and content familiarity. By integrating these diverse aspects, the research offers a thorough understanding of how various factors impact perceived bias in content moderation—an area that is notably underexplored in the existing literature.

Thirdly, this study introduces a novel framework that integrates user engagement dynamics and domain knowledge into deep learning techniques to enhance the effectiveness of content moderation, going beyond the previous focus on either element alone. The findings provide concrete evidence of the importance of user engagement, such as credibility and stance, as well as domain knowledge in improving model performance. This innovative approach not only advances the technical capabilities for content moderation but also demonstrates the practical value of combining human and machine insights to achieve more accurate and efficient moderation outcomes.

Fourthly, this study enhances the rigor of content moderation research by contributing new data collection methods. The methodology used can serve as a testbed for future research, enabling

further advancements in data gathering and analysis. These methods can be adopted and refined by subsequent studies, promoting continued progress in the field.

### **6.3 Research Implications**

The research artifacts and associated findings have implications for improving the efficiency and effectiveness of content moderation processes.

This dissertation implies significant advancements in the utilization of NLP models to streamline content moderation by automating the initial review process of UGC. This automation reduces the workload on human moderators and enables quicker responses to potentially harmful content. NLP models are scalable, and capable of handling large volumes of content, thereby improving the scalability of AI-driven content moderation systems. This enhancement in understanding the interplay between technology and user behavior is pivotal. Additionally, the research provides valuable insights for policymakers and regulators, highlighting the capabilities of NLP in content moderation. These insights can inform the development of regulations that ensure the responsible use of AI in managing online content.

In addition, this research has the potential to lead to the development of integrative models that combine elements of review information comprehensiveness, user roles, and content familiarity, offering a holistic view of content moderation processes. Investigating the impact of comprehensive review information can contribute to the development of new theoretical frameworks that explain how context and detail influence decision-making processes in content moderation. Examining the impact of user roles on content moderation efficiency and perceptions can expand role theory by integrating it into online moderation contexts, providing new perspectives on role dynamics in digital environments. The research can also offer evidence for the role of heuristics and biases in content moderation, demonstrating how familiarity with content



can lead to more heuristic-driven and efficient decision-making. Insights from this research can significantly contribute to theory building in the field of digital communication, offering a nuanced understanding of how various factors influence content moderation. Moreover, the findings encourage cross-disciplinary research, uniting scholars from psychology, information science, communication, and sociology to explore the multifaceted nature of content moderation. The research can also inform policy and regulation studies, providing empirical evidence on how different factors influence the effectiveness and fairness of content moderation practices.

Last but not least, this research can lead to the development of hybrid models that integrate human judgment and machine efficiency. Understanding how collaboration between users and moderators can enhance automated systems can inform the creation of more robust and adaptive content moderation frameworks. The study can provide insights into how collaborative frameworks impact user experience. Investigating user-moderator collaboration can highlight the importance of trust and transparency in content moderation. Findings can inform the development of transparent systems that clearly communicate how moderation decisions are made, fostering greater trust among users. Additionally, insights from user-moderator collaboration can be utilized to improve the algorithms used in automated content moderation. Continuous feedback from human moderators and users can help refine deep learning models, making them more accurate and context-aware.

## **6.4 Practical Implications**

The findings of this study not only guide the development of content moderation techniques but also have practical implications for designing online community policies and optimizing moderation strategies. This study advocates for a collaborative moderation approach, ensuring comprehensive UGC review from multiple perspectives. It underscores the importance of

effectively managing workload and expectations among users and moderators. Clear communication, consistent application of moderation policies, and supportive organizational structures are crucial to navigating these dynamics and promoting effective moderation practices.

Platforms may consider the increased time and cognitive effort required for reviewing comprehensive UGC. This consideration may necessitate additional staffing, more robust training programs, and possibly the development of specialized roles focused on handling complex cases. As UGC becomes richer, content moderation methods should leverage multiple aspects of online users' interactive information to capture a broader range of tactics for content moderation interventions. To mitigate the increased cognitive load, platforms may invest in advanced technological aids, such as AI-powered tools that can pre-filter or flag content needing detailed human review. These tools can handle preliminary assessments, allowing moderators to focus on more complex, nuanced decisions. Additionally, technology can assist in organizing and presenting diverse information in a more digestible format, reducing the cognitive burden on reviewers. Platforms can also use insights from UGC to refine their content policies. Clear guidelines on what constitutes a violation would lead to better alignment between community expectations and platform enforcement practices. Moreover, content moderation systems can be fine-tuned to prioritize content with extensive information for initial automated review, leveraging detailed information to make preliminary assessments. This can streamline the moderation process and ensure that complex cases are addressed efficiently.

## **6.5 Limitations and Future Work**

This research has several limitations and opens avenues for future investigation, including:

- Proactive and Pre-Moderation: while this study primarily addresses post- and reactive-moderation strategies, exploring pre-moderation and proactive moderation approaches is

equally essential. Future research should investigate how evolving community rules can be preemptively integrated into content moderation models to prevent the dissemination of harmful content before it becomes public. This proactive approach requires models that continuously adapt to reflect changing community norms and values, enhancing the anticipatory capabilities of content moderation systems.

- Large Language Models: a notable limitation of this research is the restriction on the number of comments used in developing the deep learning models for content moderation supported by RoBERTa. With the increasing power of other large language models (e.g., GPT, Gemini, and LLaMA), future studies should explore the potential of these models to further improve the scalability and effectiveness of content moderation.
- Hybrid Systems Combining Human- and AI-based Moderation: this research integrates user engagement and domain knowledge into a deep learning model for automated content moderation. However, future research should delve into hybrid systems that effectively combine the strengths of human moderators and AI models. Such systems can leverage the nuanced understanding and contextual awareness of human moderators alongside the speed and reliability of AI, potentially leading to superior moderation performance. Investigating the optimal balance and interaction between human and machine moderation can provide deeper insights into creating more efficient and reliable models for content moderation.
- Generalizability: the current study selects several specific domains, including healthcare, gaming, fashion, and sports. While the proposed framework is envisioned to be generalizable to various domains and online communities, it is crucial to conduct extensive examinations across a broader range of communities and domains to validate its

applicability. Future research should explore diverse online communities to gain a more complete understanding of the adaptability and effectiveness of the moderation framework.

- **Moderation via Negative User Expressions:** this study examines the characteristics of user engagement in content moderation, primarily focusing on temporality, credibility, and orientation. Negative expressions in user comments likely signal that a post may be controversial, offensive, or harmful. These comments offer valuable context that aids moderators in understanding the broader implications of the post. Additionally, negative comments often include toxic language, such as insults, threats, or hate speech, reflecting user sentiment and reactions to the post. By concentrating on posts that attract negative comments, moderators can prioritize content that is more likely to violate community guidelines.
- **Multimodal Content and Comprehensive Review:** this study investigates the effects of review information comprehensiveness, user roles, and content familiarity on the decision-making process in content moderation, primarily focusing on textual information such as posts, user comments, and community rules. However, future research should consider the growing prevalence of multimodal social media content, including images, videos, and audio. Additionally, the clarity and interpretation of community rules in the context of these varied content types should be examined. By incorporating multimodal content, researchers can develop more holistic and robust moderation models that address the full spectrum of online content, ensuring more comprehensive and accurate moderation practices.
- **Ethical and Societal Implications:** as content moderation increasingly relies on AI and automated systems, it is imperative to consider the ethical and societal implications of these

technologies. Future research should investigate the impact of automated content moderation on user privacy, trust, transparency, and the potential for algorithmic biases. Addressing these ethical concerns is critical to developing responsible and fair content moderation practices that encourage productive user engagement while maintaining community standards.

## REFERENCES

- [1] E. C. Malthouse, B. J. Calder, S. J. Kim, and M. Vandenbosch, "Evidence that user-generated content that produces engagement increases purchase behaviours," *Journal of Marketing Management*, vol. 32, no. 5–6, pp. 427–444, 2016.
- [2] H. Allcott, M. Gentzkow, and C. Yu, "Trends in the diffusion of misinformation on social media," *Research & Politics*, vol. 6, no. 2, p. 2053168019848554, Apr. 2019, doi: 10.1177/2053168019848554.
- [3] J. Grimmelmann, "The virtues of moderation," *Yale JL & Tech.*, vol. 17, p. 42, 2015.
- [4] T. Gillespie, "Content moderation, AI, and the question of scale," *Big Data & Society*, vol. 7, no. 2, p. 2053951720943234, Jul. 2020, doi: 10.1177/2053951720943234.
- [5] "Global Internet Forum to Counter Terrorism | About." Accessed: Jul. 19, 2023. [Online]. Available: <https://perma.cc/44V5-554U>
- [6] N. J. Stroud, J. M. Scacco, A. Muddiman, and A. L. Curry, "Changing deliberative norms on news organizations' Facebook sites," *Journal of Computer-Mediated Communication*, vol. 20, no. 2, pp. 188–203, 2015.
- [7] A. Sukumaran, S. Vezich, M. McHugh, and C. Nass, "Normative influences on thoughtful online participation," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 3401–3410.
- [8] A. A. Anderson, D. Brossard, D. A. Scheufele, M. A. Xenos, and P. Ladwig, "The 'nasty effect': Online incivility and risk perceptions of emerging technologies," *Journal of computer-mediated communication*, vol. 19, no. 3, pp. 373–387, 2014.
- [9] J. Feigenbaum, A. D. Jaggard, and R. N. Wright, "Towards a formal model of accountability," in *Proceedings of the 2011 New Security Paradigms Workshop*, in NSPW '11. New York, NY, USA: Association for Computing Machinery, Sep. 2011, pp. 45–56. doi: 10.1145/2073276.2073282.
- [10] E. Chandrasekharan, S. Jhaver, A. Bruckman, and E. Gilbert, "Quarantined! Examining the effects of a community-wide moderation intervention on Reddit," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 29, no. 4, pp. 1–26, 2022.
- [11] C. Lampe, P. Zube, J. Lee, C. H. Park, and E. Johnston, "Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums," *Government Information Quarterly*, vol. 31, no. 2, pp. 317–326, 2014.
- [12] C. Lampe and P. Resnick, "Slash(dot) and burn: distributed moderation in a large online conversation space," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, in CHI '04. New York, NY, USA: Association for Computing Machinery, Apr. 2004, pp. 543–550. doi: 10.1145/985692.985761.
- [13] J. Seering, "Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–28, 2020.
- [14] D. Karabulut, C. Ozcinar, and G. Anbarjafari, "Automatic content moderation on social media," *Multimed Tools Appl*, vol. 82, no. 3, pp. 4439–4463, Jan. 2023, doi: 10.1007/s11042-022-11968-3.
- [15] S. Myers West, "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms," *New Media & Society*, vol. 20, no. 11, pp. 4366–4383, Nov. 2018, doi: 10.1177/1461444818773059.

- [16] A. Seyam, A. Bou Nassif, M. Abu Talib, Q. Nasir, and B. Al Blooshi, "Deep Learning Models to Detect Online False Information: A Systematic Literature Review," in *The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research*, in ArabWIC 2021. New York, NY, USA: Association for Computing Machinery, Aug. 2021, pp. 1–5. doi: 10.1145/3485557.3485580.
- [17] H. Habib, M. B. Musa, F. Zaffar, and R. Nithyanand, "To Act or React: Investigating Proactive Strategies For Online Community Moderation," Jun. 27, 2019, *arXiv: arXiv:1906.11932*. Accessed: Aug. 18, 2022. [Online]. Available: <http://arxiv.org/abs/1906.11932>
- [18] C. Schluger, J. P. Chang, C. Danescu-Niculescu-Mizil, and K. Levy, "Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–27, 2022.
- [19] S. Taylor, C. A. Landry, M. M. Paluszczek, R. Groenewoud, G. S. Rachor, and G. J. G. Asmundson, "A Proactive Approach for Managing COVID-19: The Importance of Understanding the Motivational Roots of Vaccination Hesitancy for SARS-CoV2," *Frontiers in Psychology*, vol. 11, p. 2890, 2020, doi: 10.3389/fpsyg.2020.575950.
- [20] L. Zhang, J. Yang, W. Chu, and B. Tseng, "A machine-learned proactive moderation system for auction fraud detection," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2501–2504.
- [21] L. Zhang, J. Yang, and B. Tseng, "Online modeling of proactive moderation system for auction fraud detection," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 669–678.
- [22] P. Coutinho and R. José, "Moderation techniques for user-generated content in place-based communication," in *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*, Jun. 2017, pp. 1–6. doi: 10.23919/CISTI.2017.7975786.
- [23] G. De Gregorio, "Democratising online content moderation: A constitutional framework," *Computer Law & Security Review*, vol. 36, p. 105374, 2020.
- [24] S. Kamara *et al.*, "Outside looking in: Approaches to content moderation in end-to-end encrypted systems," *arXiv preprint arXiv:2202.04617*, 2022.
- [25] C. A. Pan, S. Yakhmi, T. P. Iyer, E. Strasnick, A. X. Zhang, and M. S. Bernstein, "Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–31, 2022.
- [26] M. Alizadeh, F. Gilardi, E. Hoes, K. J. Klüser, M. Kubli, and N. Marchal, "Content Moderation As a Political Issue: The Twitter Discourse Around Trump's Ban," *Journal of Quantitative Description: Digital Media*, vol. 2, 2022.
- [27] C. Clune and E. McDaid, "Content moderation on social media: constructing accountability in the digital space," *Accounting, Auditing & Accountability Journal*, 2023.
- [28] E. J. Llansó, "No amount of 'AI' in content moderation will solve filtering's prior-restraint problem," *Big Data & Society*, vol. 7, no. 1, p. 2053951720920686, Jan. 2020, doi: 10.1177/2053951720920686.
- [29] S. Jhaver, A. Bruckman, and E. Gilbert, "Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, p. 150:1-150:27, Nov. 2019, doi: 10.1145/3359252.

- [30] T. Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.
- [31] N. P. Suzor, S. M. West, A. Quodling, and J. York, “What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation,” *International Journal of Communication*, vol. 13, no. 0, Art. no. 0, Mar. 2019.
- [32] S. T. Roberts, *Behind the screen*. Yale University Press, 2019. Accessed: Jul. 16, 2024. [Online]. Available: [https://books.google.com/books?hl=en&lr=&id=3-aaDwAAQBAJ&oi=fnd&pg=PP1&dq=Behind+the+Screen:+Content+Moderation+in+the+Shadows+of+Social+Media.+&ots=CPVBCjTnIw&sig=u\\_0e3Ftb0jL\\_-x-W9Z9LtTE\\_h9Q](https://books.google.com/books?hl=en&lr=&id=3-aaDwAAQBAJ&oi=fnd&pg=PP1&dq=Behind+the+Screen:+Content+Moderation+in+the+Shadows+of+Social+Media.+&ots=CPVBCjTnIw&sig=u_0e3Ftb0jL_-x-W9Z9LtTE_h9Q)
- [33] R. Gorwa, R. Binns, and C. Katzenbach, “Algorithmic content moderation: Technical and political challenges in the automation of platform governance,” *Big Data & Society*, vol. 7, no. 1, p. 2053951719897945, Jan. 2020, doi: 10.1177/2053951719897945.
- [34] J. N. Matias, “The Civic Labor of Volunteer Moderators Online,” *Social Media + Society*, vol. 5, no. 2, p. 205630511983677, Apr. 2019, doi: 10.1177/2056305119836778.
- [35] K. Klonick, “The new governors: The people, rules, and processes governing online speech,” *Harv. L. Rev.*, vol. 131, p. 1598, 2017.
- [36] R. Jiménez Durán, “The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter,” Nov. 01, 2021, *Rochester, NY*: 4044098. doi: 10.2139/ssrn.4044098.
- [37] Y. Liu, P. Yildirim, and Z. J. Zhang, “Implications of Revenue Models and Technology for Content Moderation Strategies,” *Marketing Science*, vol. 41, no. 4, pp. 831–847, Jul. 2022, doi: 10.1287/mksc.2022.1361.
- [38] H. Shahbaznezhad, R. Dolan, and M. Rashidirad, “The Role of Social Media Content Format and Platform in Users’ Engagement Behavior,” *Journal of Interactive Marketing*, vol. 53, pp. 47–65, Feb. 2021, doi: 10.1016/j.intmar.2020.05.001.
- [39] L. A. Kappelman, “Measuring user involvement: a diffusion of innovation perspective,” *SIGMIS Database*, vol. 26, no. 2–3, pp. 65–86, May 1995, doi: 10.1145/217278.217286.
- [40] K. Z. Zhang, C. M. Cheung, and M. K. Lee, “Examining the moderating effect of inconsistent reviews and its gender differences on consumers’ online shopping decision,” *International Journal of Information Management*, vol. 34, no. 2, pp. 89–98, 2014.
- [41] J. A. Jiang, P. Nie, J. R. Brubaker, and C. Fiesler, “A trade-off-centered framework of content moderation,” *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 1, pp. 1–34, 2023.
- [42] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan, “Human-ai collaboration via conditional delegation: A case study of content moderation,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–18.
- [43] J. Seering, T. Wang, J. Yoon, and G. Kaufman, “Moderator engagement and community development in the age of algorithms,” *New Media & Society*, vol. 21, no. 7, pp. 1417–1443, Jul. 2019, doi: 10.1177/1461444818821316.
- [44] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.
- [45] E. Wulczyn, N. Thain, and L. Dixon, “Ex machina: Personal attacks seen at scale,” in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1391–1399.



- [46] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [47] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.
- [48] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, “Tweeteval: Unified benchmark and comparative evaluation for tweet classification,” *arXiv preprint arXiv:2010.12421*, 2020.
- [49] C. Guo, J. Cao, X. Zhang, K. Shu, and M. Yu, “Exploiting Emotions for Fake News Detection on Social Media,” Mar. 2019.
- [50] R. Yang, J. Ma, H. Lin, and W. Gao, “A Weakly Supervised Propagation Model for Rumor Verification and Stance Detection with Multiple Instance Learning,” in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR ’22. New York, NY, USA: Association for Computing Machinery, Jul. 2022, pp. 1761–1772. doi: 10.1145/3477495.3531930.
- [51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
- [52] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Dec. 2019, Accessed: Feb. 24, 2023. [Online]. Available: <https://openreview.net/forum?id=SyxS0T4tvS>
- [53] J. C. Medina Serrano, O. Papakyriakopoulos, and S. Hegelich, “NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube,” in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online: Association for Computational Linguistics, Jul. 2020. Accessed: Aug. 19, 2022. [Online]. Available: <https://aclanthology.org/2020.nlpCOVID19-acl.17>
- [54] S. Raza and C. Ding, “Fake news detection based on news content and social contexts: a transformer-based approach,” *Int J Data Sci Anal*, vol. 13, no. 4, pp. 335–362, May 2022, doi: 10.1007/s41060-021-00302-z.
- [55] R. K. Kaliyar, P. Kumar, M. Kumar, M. Narkhede, S. Namboodiri, and S. Mishra, “DeepNet: An Efficient Neural Network for Fake News Detection using News-User Engagements,” in *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, Oct. 2020, pp. 1–6. doi: 10.1109/ICCCS49678.2020.9277353.
- [56] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, “Rumor detection with hierarchical social attention network,” in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 943–951.
- [57] E. Chandrasekharan, C. Gandhi, M. W. Mustelier, and E. Gilbert, “Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators,” *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, p. 174:1-174:30, Nov. 2019, doi: 10.1145/3359276.
- [58] S. Tirunillai and G. J. Tellis, “Does chatter really matter? Dynamics of user-generated content and stock performance,” *Marketing Science*, vol. 31, no. 2, pp. 198–215, 2012.

- [59] B. R. Chakravarthi, “HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion,” in *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, 2020, pp. 41–53.
- [60] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.
- [61] P. Roma and D. Aloini, “How does brand-related user-generated content differ across social media? Evidence reloaded,” *Journal of Business Research*, vol. 96, pp. 322–339, 2019.
- [62] W. Iqbal, G. Tyson, and I. Castro, “Looking on Efficiency of Content Moderation Systems from the Lens of Reddit’s Content Moderation Experience During COVID-19,” *Available at SSRN 4007864*, 2022.
- [63] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, “Misinformation in Social Media: Definition, Manipulation, and Detection,” *SIGKDD Explor. Newsl.*, vol. 21, no. 2, pp. 80–90, Nov. 2019, doi: 10.1145/3373464.3373475.
- [64] D. A. Broniatowski *et al.*, “Twitter and Facebook posts about COVID-19 are less likely to spread misinformation compared to other health topics,” *PLOS ONE*, vol. 17, no. 1, p. e0261768, Jan. 2022, doi: 10.1371/journal.pone.0261768.
- [65] E. Chen, H. Chang, A. Rao, K. Lerman, G. Cowan, and E. Ferrara, “COVID-19 misinformation and the 2020 US presidential election,” *The Harvard Kennedy School Misinformation Review*, 2021, doi: 10.37016/mr-2020-57.
- [66] R. Dolan, J. Conduit, C. Frethey-Bentham, J. Fahy, and S. Goodman, “Social media engagement behavior: A framework for engaging customers through social media content,” *European Journal of Marketing*, vol. 53, no. 10, pp. 2213–2243, 2019.
- [67] N. P. Cechetti, E. A. Bellei, D. Biduski, J. P. M. Rodriguez, M. K. Roman, and A. C. B. De Marchi, “Developing and implementing a gamification method to improve user engagement: A case study with an m-Health application for hypertension monitoring,” *Telematics and Informatics*, vol. 41, pp. 126–138, 2019.
- [68] L. Hong and M. Lalmas, “Tutorial on online user engagement: Metrics and optimization,” in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 1303–1305.
- [69] R. Jaakonmäki, O. Müller, and J. Vom Brocke, “The impact of content, context, and creator on user engagement in social media marketing,” in *Proceedings of the Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Press, 2017, pp. 1152–1160.
- [70] T. Dias Oliva, “Content moderation technologies: Applying human rights standards to protect freedom of expression,” *Human Rights Law Review*, vol. 20, no. 4, pp. 607–640, 2020.
- [71] G. Morrow, B. Swire-Thompson, J. M. Polny, M. Kopec, and J. P. Wihbey, “The emerging science of content labeling: Contextualizing social media content moderation,” *Journal of the Association for Information Science and Technology*, vol. 73, no. 10, pp. 1365–1386, 2022, doi: 10.1002/asi.24637.
- [72] J. C. Bertot, P. T. Jaeger, S. Munson, and T. Glaisyer, “Social media technology and government transparency,” *Computer*, vol. 43, no. 11, pp. 53–59, 2010.
- [73] S. M. C. Loureiro and J. Lopes, “How corporate social responsibility initiatives in social media affect awareness and customer engagement,” *Journal of Promotion Management*, vol. 25, no. 3, pp. 419–438, 2019.

- [74] M. D. Molina and S. S. Sundar, "Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation," *New Media & Society*, p. 14614448221103534, 2022.
- [75] J. Zeng and D. B. V. Kaye, "From content moderation to visibility moderation: A case study of platform governance on TikTok," *Policy & Internet*, vol. 14, no. 1, pp. 79–95, 2022.
- [76] B. Sander, "Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation," *Fordham Int'l LJ*, vol. 43, p. 939, 2019.
- [77] S. Karunakaran and R. Ramakrishnan, "Testing stylistic interventions to reduce emotional impact of content moderation workers," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2019, pp. 50–58.
- [78] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, and M. Lease, "The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–14.
- [79] K. Vaccaro, Z. Xiao, K. Hamilton, and K. Karahalios, "Contestability For Content Moderation," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, p. 318:1-318:28, Oct. 2021, doi: 10.1145/3476059.
- [80] L. Wang and H. Zhu, "How are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 824–838.
- [81] S. Roberts, "Commercial Content Moderation: Digital Laborers' Dirty Work," *Media Studies Publications*, Jan. 2016, [Online]. Available: <https://ir.lib.uwo.ca/commpub/12>
- [82] K. Crawford and T. Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint," *New Media & Society*, vol. 18, no. 3, pp. 410–428, 2016.
- [83] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman, "Human-machine collaboration for content regulation: The case of reddit automoderator," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 26, no. 5, pp. 1–35, 2019.
- [84] H. Bloch-Wehba, "Automation in moderation," *Cornell Int'l LJ*, vol. 53, p. 41, 2020.
- [85] M. Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process," *Meta*. Accessed: Jul. 18, 2023. [Online]. Available: <https://about.fb.com/news/2018/04/comprehensive-community-standards/>
- [86] "YouTube Community Guidelines enforcement – Google Transparency Report." Accessed: Jul. 18, 2023. [Online]. Available: <https://transparencyreport.google.com/youtube-policy/removals>
- [87] "Evolving our Twitter Transparency Report: expanded data and insights." Accessed: Jul. 18, 2023. [Online]. Available: [https://blog.twitter.com/en\\_us/topics/company/2018/evolving-our-twitter-transparency-report](https://blog.twitter.com/en_us/topics/company/2018/evolving-our-twitter-transparency-report)
- [88] H. Sun and W. Ni, "Design and Application of an AI-Based Text Content Moderation System," *Scientific Programming*, vol. 2022, p. e2576535, Feb. 2022, doi: 10.1155/2022/2576535.
- [89] E. Spertus, "Smokey: Automatic recognition of hostile messages," in *Aaai/iaai*, 1997, pp. 1058–1065.

- [90] S. Sood, J. Antin, and E. Churchill, “Profanity use in online communities,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2012, pp. 1481–1490.
- [91] *Our List of Dirty, Naughty, Obscene, and Otherwise Bad Words*. (Jul. 19, 2023). LDNOOBW. Accessed: Jul. 19, 2023. [Online]. Available: <https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words>
- [92] S. Brody and N. Diakopoulos, “Coooooooooooooooooolllllllllll!!!!!! using word lengthening to detect sentiment in microblogs,” in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 562–570.
- [93] S. Chancellor, J. A. Pater, T. Clear, E. Gilbert, and M. De Choudhury, “#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, in CSCW ’16. New York, NY, USA: Association for Computing Machinery, Feb. 2016, pp. 1201–1213. doi: 10.1145/2818048.2819963.
- [94] B. Fishman, “Crossroads: Counter-terrorism and the Internet (February 2019),” *Texas National Security Review*, 2019.
- [95] D. Androćec, “Machine learning methods for toxic comment classification: a systematic review,” *Acta Universitatis Sapientiae, Informatica*, vol. 12, no. 2, pp. 205–216, Nov. 2020, doi: 10.2478/ausi-2020-0012.
- [96] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, “Detecting offensive language in social media to protect adolescent online safety,” in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, IEEE, 2012, pp. 71–80.
- [97] T. De Smedt, G. De Pauw, and P. Van Ostaeyen, “Automatic detection of online jihadist hate speech,” *arXiv preprint arXiv:1803.04596*, 2018.
- [98] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, “Detection of harassment on web 2.0,” *Proceedings of the Content Analysis in the WEB*, vol. 2, no. 0, pp. 1–7, 2009.
- [99] J. Salminen *et al.*, “Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2018.
- [100] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, “Offensive language detection using multi-level classification,” in *Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, May 31–June 2, 2010. Proceedings 23*, Springer, 2010, pp. 16–27.
- [101] R. Binns, M. Veale, M. Van Kleek, and N. Shadbolt, “Like trainer, like bot? Inheritance of bias in algorithmic content moderation,” in *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, Springer, 2017, pp. 405–415.
- [102] T. Davidson, D. Warmesley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, Art. no. 1, May 2017, doi: 10.1609/icwsml.v11i1.14955.
- [103] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, “Inducing a lexicon of abusive words—a feature-based approach,” in *Proceedings of the 2018 Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1046–1056.
- [104] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 29–30.
  - [105] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, “Deeper Attention to Abusive User Content Moderation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1125–1135. doi: 10.18653/v1/D17-1117.
  - [106] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “Hatexplain: A benchmark dataset for explainable hate speech detection,” in *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 14867–14875.
  - [107] F. Tan, Y. Hu, C. Hu, K. Li, and K. Yen, “TNT: Text Normalization based Pre-training of Transformers for Content Moderation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 4735–4741. doi: 10.18653/v1/2020.emnlp-main.383.
  - [108] Q. He, Y. Hong, and T. S. Raghu, “The effects of machine-powered platform governance: An empirical study of content moderation,” *Available at SSRN 3767680*, 2021.
  - [109] W. Wang *et al.*, “MTTM: metamorphic testing for textual content moderation software,” in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, IEEE, 2023, pp. 2387–2399.
  - [110] Y. Ye, T. Le, and D. Lee, “NoisyHate: Benchmarking Content Moderation Machine Learning Models with Human-Written Perturbations Online,” *arXiv preprint arXiv:2303.10430*, 2023.
  - [111] “Jigsaw,” Jigsaw. Accessed: Jul. 21, 2023. [Online]. Available: <https://jigsaw.google.com/>
  - [112] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, “Cyberbullying detection solutions based on deep learning architectures,” *Multimedia Systems*, vol. 29, no. 3, pp. 1839–1852, Jun. 2023, doi: 10.1007/s00530-020-00701-5.
  - [113] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, “A Unified Deep Learning Architecture for Abuse Detection,” in *Proceedings of the 10th ACM Conference on Web Science*, in WebSci ’19. New York, NY, USA: Association for Computing Machinery, Jun. 2019, pp. 105–114. doi: 10.1145/3292522.3326028.
  - [114] K. Wang, Z. Fu, L. Zhou, and Y. Zhu, “Content Moderation in Social Media: The Characteristics, Degree, and Efficiency of User Engagement,” in *2022 3rd Asia Symposium on Signal Processing (ASSP)*, Dec. 2022, pp. 86–91. doi: 10.1109/ASSP57481.2022.00022.
  - [115] “Perspective API.” Accessed: Jul. 21, 2023. [Online]. Available: <https://perspectiveapi.com/>
  - [116] “敏感词检测\_文本审核\_敏感词过滤-百度 AI 开放平台.” Accessed: Jul. 25, 2023. [Online]. Available: <https://ai.baidu.com/tech/textcensoring>
  - [117] “文本内容审核\_文本审核\_文字审核-华为云.” Accessed: Jul. 25, 2023. [Online]. Available: <https://www.huaweicloud.com/product/textmoderation.html>
  - [118] S. Udupa, A. Maronikolakis, and A. Wisiorek, “Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence,” *Big Data & Society*, vol. 10, no. 1, p. 20539517231172424, 2023.

- [119] B. Etim, “The Times Sharply Increases Articles Open for Comments, Using Google’s Technology,” *The New York Times*, Jun. 13, 2017. Accessed: Jul. 18, 2023. [Online]. Available: <https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html>
- [120] “Automoderator - reddit.com,” reddit. Accessed: Jul. 18, 2023. [Online]. Available: <https://www.reddit.com/wiki/automoderator/>
- [121] P. Juneja, D. Rama Subramanian, and T. Mitra, “Through the looking glass: Study of transparency in Reddit’s moderation practices,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. GROUP, pp. 1–35, 2020.
- [122] R. Ma and Y. Kou, ““How advertiser-friendly is my video?”: YouTuber’s Socioeconomic Interactions with Algorithmic Content Moderation,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–25, 2021.
- [123] S. Jhaver, D. S. Appling, E. Gilbert, and A. Bruckman, “Did you suspect the post would be removed?”: User reactions to content removals on reddit,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, 2018.
- [124] M. Eslami *et al.*, “First I ‘like’ it, then I hide it: Folk Theories of Social Feeds,” in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 2371–2382.
- [125] D. Kaye, “Speech police: The global struggle to govern the Internet,” 2019.
- [126] S. T. Roberts, “Digital detritus: ‘Error’ and the logic of opacity in social media content moderation,” *First Monday*, 2018.
- [127] N. P. Suzor, *Lawless: The secret rules that govern our digital lives*. Cambridge University Press, 2019.
- [128] T. Gillespie, “Do not recommend? Reduction as a form of content moderation,” *Social Media+ Society*, vol. 8, no. 3, p. 20563051221117552, 2022.
- [129] S. A. Baker, M. Wade, and M. J. Walsh, “The challenges of responding to misinformation during a pandemic: content moderation and the limitations of the concept of harm,” *Media International Australia*, vol. 177, no. 1, pp. 103–107, Nov. 2020, doi: 10.1177/1329878X20951301.
- [130] R. Krishna, “Tumblr Launched An Algorithm To Flag Porn And So Far It’s Just Caused Chaos,” BuzzFeed News. Accessed: Jul. 18, 2023. [Online]. Available: <https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban>
- [131] L. Parks, “Dirty data: content moderation, regulatory outsourcing, and the cleaners,” *Film Quarterly*, vol. 73, no. 1, pp. 11–18, 2019.
- [132] M. Soha and Z. J. McDowell, “Monetizing a meme: YouTube, content ID, and the Harlem Shake,” *Social Media+ Society*, vol. 2, no. 1, p. 2056305115623801, 2016.
- [133] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on twitter using a convolution-gru based deep neural network,” in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, Springer, 2018, pp. 745–760.
- [134] C. Wang, “Interpreting neural network hate speech classifiers,” in *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 2018, pp. 86–92.
- [135] D. L. Burk and J. E. Cohen, “Fair use infrastructure for rights management systems,” *Harv. JL Tech*, vol. 15, p. 41, 2001.
- [136] S. Bar-Ziv and N. Elkin-Koren, “Behind the scenes of online copyright enforcement: Empirical evidence on notice & takedown,” *Conn. L. Rev.*, vol. 50, p. 339, 2018.

- [137] K. Erickson and M. Kretschmer, “This video is unavailable,” *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, vol. 9, p. 75, 2018.
- [138] J. M. Urban, J. Karaganis, and B. Schofield, “Notice and takedown in everyday practice,” *UC Berkeley Public Law Research Paper*, no. 2755628, 2017.
- [139] N. Diakopoulos and M. Naaman, “Towards quality discourse in online news comments,” in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 2011, pp. 133–142.
- [140] T. Poell, D. B. Nieborg, and B. E. Duffy, *Platforms and cultural production*. John Wiley & Sons, 2021.
- [141] B. Marshall, “Algorithmic misogyny in content moderation practice,” *Heinrich-Böll-Stiftung European Union*, 2021.
- [142] N. Suzor, T. Van Geelen, and S. Myers West, “Evaluating the legitimacy of platform governance: A review of research and a shared research agenda,” *International Communication Gazette*, vol. 80, no. 4, pp. 385–400, Jun. 2018, doi: 10.1177/1748048518757142.
- [143] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, “Grand challenges for HCI researchers,” *interactions*, vol. 23, no. 5, pp. 24–25, Aug. 2016, doi: 10.1145/2977645.
- [144] D. Wang *et al.*, “From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people,” in *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–6.
- [145] K. Vaccaro, C. Sandvig, and K. Karahalios, “‘At the End of the Day Facebook Does What It Wants’ How Users Experience Contesting Algorithmic Content Moderation,” *Proceedings of the ACM on human-computer interaction*, vol. 4, no. CSCW2, pp. 1–22, 2020.
- [146] Y. Kou and X. Gui, “Mediating community-AI interaction through situated explanation: the case of AI-Led moderation,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–27, 2020.
- [147] J. Cobbe, “Algorithmic Censorship by Social Platforms: Power and Resistance,” *Philos. Technol.*, vol. 34, no. 4, pp. 739–766, Dec. 2021, doi: 10.1007/s13347-020-00429-0.
- [148] Y. Gerrard, “Beyond the hashtag: Circumventing content moderation on social media,” *New Media & Society*, vol. 20, no. 12, pp. 4492–4511, Dec. 2018, doi: 10.1177/1461444818776611.
- [149] Y. Zhu *et al.*, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27. Accessed: Jul. 09, 2024. [Online]. Available: [https://www.cv-foundation.org/openaccess/content\\_iccv\\_2015/html/Zhu\\_Aligning\\_Books\\_and\\_ICCV\\_2015\\_paper.html](https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html)
- [150] Y. You, J. Li, J. Hseu, X. Song, J. Demmel, and C.-J. Hsieh, “Reducing BERT pre-training time from 3 days to 76 minutes,” *arXiv preprint arXiv:1904.00962*, vol. 12, p. 2, 2019.
- [151] J. Ma, W. Gao, and K.-F. Wong, “Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 708–717. doi: 10.18653/v1/P17-1066.

- [152] E. Chandrasekharan *et al.*, “The Internet’s Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales,” *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, p. 32:1-32:25, Nov. 2018, doi: 10.1145/3274301.
- [153] Z. Lin, N. Salehi, B. Yao, Y. Chen, and M. Bernstein, “Better When It Was Smaller? Community Content and Behavior After Massive Growth,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, Art. no. 1, May 2017.
- [154] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, “You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech,” *Proceedings of the ACM on human-computer interaction*, vol. 1, no. CSCW, pp. 1–22, 2017.
- [155] L. Lessig, *Code: And other laws of cyberspace*. ReadHowYouWant. com, 2009.
- [156] A. Schmidt and M. Wiegand, “A Survey on Hate Speech Detection using Natural Language Processing,” in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1–10. doi: 10.18653/v1/W17-1101.
- [157] J. Farkas and J. Schou, “Fake News as a Floating Signifier: Hegemony, Antagonism and the Politics of Falsehood,” *Javnost - The Public*, vol. 25, no. 3, pp. 298–314, Jul. 2018, doi: 10.1080/13183222.2018.1463047.
- [158] R. Caplan and D. Boyd, “Isomorphism through algorithms: Institutional dependencies in the case of Facebook,” *Big Data & Society*, vol. 5, no. 1, p. 205395171875725, Jan. 2018, doi: 10.1177/2053951718757253.
- [159] T. B. Ksiazek, “Civil interactivity: How news organizations’ commenting policies explain civility and hostility in user comments,” *Journal of Broadcasting & Electronic Media*, vol. 59, no. 4, pp. 556–573, 2015.
- [160] M. Zuckerberg, “A Blueprint for Content Governance and Enforcement.” Accessed: Jul. 21, 2023. [Online]. Available: <https://www.facebook.com/notes/751449002072082/>
- [161] M. A. Arbib, “Schema theory,” *The encyclopedia of artificial intelligence*, vol. 2, pp. 1427–1443, 1992.
- [162] S.-K. Thiel, M. Reisinger, K. Röderer, and P. Fröhlich, “Playing (with) democracy: A review of gamified participation approaches,” *JeDEM-eJournal of eDemocracy and Open Government*, vol. 8, no. 3, pp. 32–60, 2016.
- [163] J. Stromer-Galley, “Diversity of political conversation on the Internet: Users’ perspectives,” *Journal of Computer-Mediated Communication*, vol. 8, no. 3, p. JCMC836, 2003.
- [164] A. W. Woolley, I. Aggarwal, and T. W. Malone, “Collective intelligence and group performance,” *Current Directions in Psychological Science*, vol. 24, no. 6, pp. 420–424, 2015.
- [165] M. Brugnach and H. Ingram, “Ambiguity: the challenge of knowing and deciding together,” *Environmental science & policy*, vol. 15, no. 1, pp. 60–71, 2012.
- [166] H. A. Simon, “Information-processing theory of human problem solving,” *Handbook of learning and cognitive processes*, vol. 5, pp. 271–295, 1978.
- [167] J. Sweller, “CHAPTER TWO - Cognitive Load Theory,” in *Psychology of Learning and Motivation*, vol. 55, J. P. Mestre and B. H. Ross, Eds., Academic Press, 2011, pp. 37–76. doi: 10.1016/B978-0-12-387691-1.00002-8.
- [168] C. Grevet, L. G. Terveen, and E. Gilbert, “Managing political differences in social media,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 1400–1408.



- [169] N. Marangunić and A. Granić, “Technology acceptance model: a literature review from 1986 to 2013,” *Universal access in the information society*, vol. 14, pp. 81–95, 2015.
- [170] J. R. Hackman and L. W. Porter, “Expectancy theory predictions of work effectiveness,” *Organizational behavior and human performance*, vol. 3, no. 4, pp. 417–426, 1968.
- [171] J. E. Maddux, “Self-Efficacy Theory,” in *Self-Efficacy, Adaptation, and Adjustment*, J. E. Maddux, Ed., in The Plenum Series in Social/Clinical Psychology. , Boston, MA: Springer US, 1995, pp. 3–33. doi: 10.1007/978-1-4419-6868-5\_1.
- [172] S. Roy, “Theory of social proof and legal compliance: a socio-cognitive explanation for regulatory (non) compliance,” *German Law Journal*, vol. 22, no. 2, pp. 238–255, 2021.
- [173] G. G. Whitchurch and L. L. Constantine, “Systems Theory,” in *Sourcebook of Family Theories and Methods*, P. Boss, W. J. Doherty, R. LaRossa, W. R. Schumm, and S. K. Steinmetz, Eds., Boston, MA: Springer US, 1993, pp. 325–355. doi: 10.1007/978-0-387-85764-0\_14.
- [174] L. Hasher and R. T. Zacks, “Automatic processing of fundamental information: the case of frequency of occurrence,” *American psychologist*, vol. 39, no. 12, p. 1372, 1984.
- [175] A. Bandura, *Self-efficacy: The exercise of control*. Macmillan, 1997. Accessed: Jul. 16, 2024. [Online]. Available: [https://books.google.com/books?hl=en&lr=&id=eJ-PN9g\\_o-EC&oi=fnd&pg=PA116&dq=Self-Efficacy:+The+Exercise+of+Control.+&ots=zAIJGYib0m&sig=9C\\_uk7nGQ5Mh7M2Ms zomgAXC740](https://books.google.com/books?hl=en&lr=&id=eJ-PN9g_o-EC&oi=fnd&pg=PA116&dq=Self-Efficacy:+The+Exercise+of+Control.+&ots=zAIJGYib0m&sig=9C_uk7nGQ5Mh7M2Ms zomgAXC740)
- [176] A. Tversky and D. Kahneman, “Availability: A heuristic for judging frequency and probability,” *Cognitive psychology*, vol. 5, no. 2, pp. 207–232, 1973.
- [177] P. Checkland, “Systems thinking, systems practice,” 1999, Accessed: Jul. 16, 2024. [Online]. Available: <http://evidence.thinkportal.org/handle/123456789/25702>
- [178] Y. Majima, K. Nishiyama, A. Nishihara, and R. Hata, “Conducting Online Behavioral Research Using Crowdsourcing Services in Japan,” *Frontiers in Psychology*, vol. 8, 2017, Accessed: Mar. 10, 2022. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fpsyg.2017.00378>
- [179] M. Chmielewski and S. C. Kucker, “An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results,” *Social Psychological and Personality Science*, vol. 11, no. 4, pp. 464–473, May 2020, doi: 10.1177/1948550619875149.
- [180] S. Cohen, T. Kamarck, and R. Mermelstein, “Perceived stress scale,” *Measuring stress: A guide for health and social scientists*, vol. 10, no. 2, pp. 1–2, 1994.
- [181] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,” in *Advances in Psychology*, vol. 52, P. A. Hancock and N. Meshkati, Eds., in Human Mental Workload, vol. 52. , North-Holland, 1988, pp. 139–183. doi: 10.1016/S0166-4115(08)62386-9.
- [182] T. Ammari, S. Schoenebeck, and D. M. Romero, “Pseudonymous Parents: Comparing Parenting Roles and Identities on the Mommit and Daddit Subreddits,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–13. Accessed: Jan. 13, 2022. [Online]. Available: <https://doi.org/10.1145/3173574.3174063>
- [183] A. Field, “Discovering statistics using IBM SPSS statistics.” sage London, 2013.
- [184] J. Sweller, “Cognitive load during problem solving: Effects on learning,” *Cognitive science*, vol. 12, no. 2, pp. 257–285, 1988.

- [185] K. Hyland, “Stance and engagement: a model of interaction in academic discourse,” *Discourse Studies*, vol. 7, no. 2, pp. 173–192, May 2005, doi: 10.1177/1461445605050365.
- [186] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Antisocial Behavior in Online Discussion Communities,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, no. 1, Art. no. 1, 2015, doi: 10.1609/icwsm.v9i1.14583.
- [187] C. Clarke *et al.*, “Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 364–376. doi: 10.18653/v1/2023.acl-long.22.
- [188] D. Kumar, Y. A. AbuHashem, and Z. Durumeric, “Watch Your Language: Investigating Content Moderation with Large Language Models,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, pp. 865–878, May 2024, doi: 10.1609/icwsm.v18i1.31358.
- [189] N. Duarte, E. Llanso, and A. Loup, “Mixed messages? The limits of automated social media content analysis,” 2017.
- [190] J. W. Dearing and J. G. Cox, “Diffusion of innovations theory, principles, and practice,” *Health affairs*, vol. 37, no. 2, pp. 183–190, 2018.
- [191] J. Suler, “The online disinhibition effect,” *Cyberpsychol Behav*, vol. 7, no. 3, pp. 321–326, Jun. 2004, doi: 10.1089/1094931041291295.
- [192] L. R. Anderson and C. A. Holt, “Information cascades in the laboratory,” *The American economic review*, pp. 847–862, 1997.
- [193] A. M. Williams and P. R. Ford, “Expertise and expert performance in sport,” *International Review of Sport and Exercise Psychology*, vol. 1, no. 1, pp. 4–18, 2008.
- [194] J. St. B. T. Evans and K. E. Stanovich, “Dual-Process Theories of Higher Cognition: Advancing the Debate,” *Perspect Psychol Sci*, vol. 8, no. 3, pp. 223–241, May 2013, doi: 10.1177/1745691612460685.
- [195] P. M. Groves and R. F. Thompson, “Habituation: a dual-process theory.,” *Psychological review*, vol. 77, no. 5, p. 419, 1970.
- [196] V. Crosset and B. Dupont, “Cognitive assemblages: The entangled nature of algorithmic content moderation,” *Big Data & Society*, vol. 9, no. 2, p. 20539517221143361, 2022.
- [197] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” Feb. 22, 2017, *arXiv*: arXiv:1609.02907. doi: 10.48550/arXiv.1609.02907.
- [198] A. Conneau *et al.*, “Unsupervised Cross-lingual Representation Learning at Scale,” Apr. 07, 2020, *arXiv*: arXiv:1911.02116. doi: 10.48550/arXiv.1911.02116.
- [199] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, “SemEval-2016 Task 6: Detecting Stance in Tweets,” in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 31–41. doi: 10.18653/v1/S16-1003.
- [200] A. Trotman, A. Puurula, and B. Burgess, “Improvements to BM25 and Language Models Examined,” in *Proceedings of the 19th Australasian Document Computing Symposium*, in ADCS ’14. New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 58–65. doi: 10.1145/2682862.2682863.
- [201] J. S. Park, J. Seering, and M. S. Bernstein, “Measuring the Prevalence of Anti-Social Behavior in Online Communities,” Aug. 27, 2022, *arXiv*: arXiv:2208.13094. doi: 10.48550/arXiv.2208.13094.

## APPENDIX A: IRB APPROVAL



<b>To:</b>	Kanlun Wang Graduate Programs
<b>From:</b>	IRB
<b>Approval Date:</b>	17-Apr-2024
<b>Expiration Date of Approval:</b>	No Date of Expiration - No End Date
<b>RE:</b>	Notice of IRB Approval by Expedited Review (under 45 CFR 46.110)
<b>Submission Type:</b>	Initial Application
<b>Expedited Category:</b>	7
<b>Study #:</b>	IRB-24-0623
<b>Study Title:</b>	Content Moderation in Online Communities

This submission has been approved by the IRB. It has been determined that the risk involved in this research is no more than minimal. The approval has no expiration or end date and is not subject to an annual continuing review. However, you are required to obtain approval for all changes to any aspect of this study before they can be implemented and to comply with the Investigator Responsibilities detailed below. This includes submitting a progress report (Administrative Check-In) at requested time points. Carefully review the Investigator Responsibilities listed below.

Your approved consent forms and other documents are available online at [Submission Page](#).

### Investigator's Responsibilities:

1. Amendments **must** be submitted for review and the amendment must be approved before implementing the amendment. This includes changes to study procedures, study materials, personnel, etc.
2. Researchers must adhere to all site-specific requirements mandated by the study site (e.g., face mask, access requirements and/or restrictions, etc.).
3. Data security procedures must follow procedures as approved in the protocol and in accordance with [UNCIT Guidelines for Data Handling](#).
4. Promptly notify the IRB ([uncc-irb@charlotte.edu](mailto:uncc-irb@charlotte.edu)) of any adverse events or unanticipated risks to participants or others.
5. Three years (3) following this approval/determination, you must complete the Admin-Check In form via Niner Research to provide a study status update.
6. Be aware that this study is included in the Office of Research Protections and Integrity (ORPI) Post-Approval Monitoring program and may be selected for post-review monitoring at some point in the future.
7. Reply to the ORPI post-review monitoring and administrative check-ins that will be conducted periodically to update ORPI as to the status of the study.
8. Complete the Closure eform via Niner Research once the study is complete.

Please be aware that approval may still be required from other relevant authorities or "gatekeepers" (e.g., school principals, facility directors, custodians of records).

## APPENDIX B: PRE-SCREENING SURVEYS

### Users with Moderation Experience

#### Q1: Age

- Under 18
- 18-25
- 26-30
- 31-35
- 36-40
- 41-45
- 46-50
- 51-55
- 56-60
- 61-65
- 66 and above

*If “Under 18” is not selected, proceed to Q2. Otherwise disqualified from the study.*

**Q2: In the past three months, how frequently have you engaged in online activities (e.g., sharing a post, commenting/replying to a user's post or comment, or moderating user-generated content)?**

- Never
- A few times a quarter
- A few times a month
- A few times a week

- A few times a day
- More than a few times a day

*If “Never” is not selected, proceed to Q3. Otherwise disqualified from the study.*

**Q3: Have you served as a moderator in an online community or a social media platform?**

- Yes
- No

*If “No” is selected proceed to Q4; Otherwise disqualified from the study.*

**Q4: Have you encountered situations where the content you generated was moderated (e.g., content removal or deletion) by social media platforms or online communities?**

- Yes
- No

*If “Yes” is selected proceed to Q5, and Q6; Otherwise disqualified from the study.*

**Q5: Which one of the following domains (i.e., the topic of discussion in online communities) are you most familiar with?**

- Health
- Sports
- Fashion
- Gaming

**Q6: To what extent are you familiar with XXX domain? (XXX: Based on participants’ selections on Q5)**

- Extremely familiar
- Somewhat familiar
- Slightly familiar

- Neither familiar nor unfamiliar
- Slightly unfamiliar
- Somewhat unfamiliar
- Extremely unfamiliar

*If “Slightly familiar”, “Somewhat familiar”, or “Extremely familiar” is selected proceed to Q7 and Q8. Otherwise disqualified from the study.*

**Q7: Which one of the following domains (i.e., the topic of discussion in online communities) are you least familiar with? (Eliminate one of the following options based on the participant’s selection on Q5)**

- Health
- Sports
- Fashion
- Gaming

**Q8: To what extent are you familiar with the XXX domain? (XXX: Based on participants’ selections on Q7)**

- Extremely unfamiliar
- Somewhat unfamiliar
- Slightly unfamiliar
- Neither familiar nor unfamiliar
- Slightly familiar
- Somewhat familiar
- Extremely familiar

*If “Slightly unfamiliar”, “Somewhat unfamiliar”, or “Extremely unfamiliar” is selected, qualified. Otherwise disqualified from the study.*

*After completing the pre-screening, qualified regular user participants with moderation experience will proceed to the regular user with moderation experience study participation consent.*

## Users without Moderation Experience

### Q1: Age

- Under 18
- 18-25
- 26-30
- 31-35
- 36-40
- 41-45
- 46-50
- 51-55
- 56-60
- 61-65
- 66 and above

*If “Under 18” is not selected, proceed to Q2. Otherwise disqualified from the study.*

**Q2: In the past three months, how frequently have you engaged in online activities (e.g., sharing a post, commenting/replying to a user's post or comment, or moderating user-generated content)?**

- Never
- A few times a quarter
- A few times a month
- A few times a week
- A few times a day
- More than a few times a day



*If “Never” is not selected, proceed to Q3. Otherwise disqualified from the study.*

**Q3: Have you served as a moderator in an online community or a social media platform?**

- Yes
- No

*If “No” is selected proceed to Q4; Otherwise disqualified from the study.*

**Q4: Have you encountered situations where the content you generated was moderated (e.g., content removal or deletion) by social media platforms or online communities?**

- Yes
- No

*If “No” is selected proceed to Q5, and Q6; Otherwise disqualified from the study.*

**Q5: Which one of the following domains (i.e., the topic of discussion in online communities) are you most familiar with?**

- Health
- Sports
- Fashion
- Gaming

**Q6: To what extent are you familiar with XXX domain? (XXX: Based on participants’ selections on Q5)**

- Extremely familiar
- Somewhat familiar
- Slightly familiar
- Neither familiar nor unfamiliar
- Slightly unfamiliar

- Somewhat unfamiliar
- Extremely unfamiliar

*If “Slightly familiar”, “Somewhat familiar”, or “Extremely familiar” is selected proceed to Q7 and Q8. Otherwise disqualified from the study.*

**Q7: Which one of the following domains (i.e., the topic of discussion in online communities) are you least familiar with? (Eliminate one of the following options based on the participant’s selection on Q5)**

- Health
- Sports
- Fashion
- Gaming

**Q8: To what extent are you familiar with the XXX domain? (XXX: Based on participants’ selections on Q7)**

- Extremely unfamiliar
- Somewhat unfamiliar
- Slightly unfamiliar
- Neither familiar nor unfamiliar
- Slightly familiar
- Somewhat familiar
- Extremely familiar

*If “Slightly unfamiliar”, “Somewhat unfamiliar”, or “Extremely unfamiliar” is selected, disqualified. Otherwise disqualified from the study.*

*After completing the pre-screening, qualified regular user participants without moderation experience will proceed to the regular user without moderation experience study participation consent.*

## Moderators

### Q1: Age

- Under 18
- 18-25
- 26-30
- 31-35
- 36-40
- 41-45
- 46-50
- 51-55
- 56-60
- 61-65
- 66 and above

*If “Under 18” is not selected, proceed to Q2. Otherwise disqualified from the study.*

**Q2: In the past three months, how frequently have you engaged in online activities (e.g., sharing a post, commenting/replying to a user's post or comment, or moderating user-generated content)?**

- Never
- A few times a quarter
- A few times a month
- A few times a week
- A few times a day
- More than a few times a day

*If “Never” is not selected, proceed to Q3. Otherwise disqualified from the study.*

**Q3: Have you served as a moderator in an online community or a social media platform?**

- Yes
- No

*If “Yes” is selected proceed to Q4 and Q5. Otherwise disqualified from the study.*

**Q4: Which one of the following domains (i.e., the topic of discussion in online communities) have you served as a primary moderator?**

- Health
- Sports
- Fashion
- Gaming

**Q5: To what extent are you familiar with XXX domain? (XXX: Based on participants’ selections in Q4)**

- Extremely familiar
- Somewhat familiar
- Slightly familiar
- Neither familiar nor unfamiliar
- Slightly unfamiliar
- Somewhat unfamiliar
- Extremely unfamiliar

*If “Slightly familiar”, “Somewhat familiar”, or “Extremely familiar” is selected proceed to Q6 and Q7. Otherwise disqualified from the study.*

**Q6: Which one of the following domains (i.e., the topic of discussion in online communities) are you least familiar with? (Eliminate one of the following options based on the participant's selection in Q4)**

- Health
- Sports
- Fashion
- Gaming

**Q7: To what extent are you familiar with the XXX domain? (XXX: Based on participants' selections on Q6)**

- Extremely unfamiliar
- Somewhat unfamiliar
- Slightly unfamiliar
- Neither familiar nor unfamiliar
- Slightly familiar
- Somewhat familiar
- Extremely familiar

*If “Slightly unfamiliar”, “Somewhat unfamiliar”, or “Extremely unfamiliar” is selected, qualified. Otherwise disqualified from the study.*

*After completing the pre-screening, qualified moderator participants will proceed to the moderator study participation consent.*

## **APPENDIX C: USER STUDY CONSENT FORM**

### **Consent to Participate in a Research Study**

Title of the Project: Content Moderation in Online Communities

Principal Investigator: [Kanlun Wang, Ph.D. Candidate, BISOM, UNC Charlotte]

Faculty Advisor: [Lina Zhou, PhD, BISOM, UNC Charlotte]

You are invited to participate in a research study. Participation in this research study is voluntary.

The information provided is to give you key information to help you decide whether or not to participate.

### **Important Information You Need to Know**

- The purpose of this study is to examine the process and decision of content moderation in online communities.
- There is no restriction on sex, race, ethnicity, origin, religion, or social or economic qualifications. However, qualified participants must meet the following qualification criteria to participate in this study:
  - You must be age 18 or older and be residing in the U.S. to participate in this study.
  - You must finish more than 500 HITs on Amazon Mechanical Turk, with a more than 95% approval rate.
  - You must have had experience with engaging in online activities (e.g., sharing posts, replying, and/or commenting on other users' posts and/or comments ) in the past three months.

- You must be a moderator who is currently or has previously been responsible for overseeing an online community (e.g., reviewing user-generated posts, and/or setting up online community rules).
- You must declare your level of familiarity and unfamiliarity with one of the following domains: gaming, fashion, health, or sports.
- The entire study takes approximately 45 minutes to complete. You have qualified to participate in this study based on your response to the pre-screening. If you consent to participate in this study, you will be asked to respond to a pre-experiment survey that asks for your demographic information and experience with social media online engagement. Then, you will be asked to participate in our formal study that requires you to review a series of user-generated content. A post-review survey will be administered after each review, which asks about your perception of content moderation based on your review of user-generated content.
- If you successfully complete the entire study and we approve your completed work (i.e., responding to all required surveys and reviewing all required user-generated content along with your perceptions), you will receive \$4 as a reward within 30 days after you submit your completed work on Amazon Mechanical Turk. It is important to know that you will be disqualified from the study, or your work will be disapproved if you fail to follow the study instructions and/or provide invalid responses (e.g., incorrectly answer the attention check questions).
- A potential minor risk is that participants may feel eye, arm, and/or wrist soreness due to interactions with a PC. To minimize this risk, participants can take a break when they feel necessary during the experiment session. In addition, there may be risks that are unknown,



such as reviewing sensitive information (e.g., phishing, cyberbullying, and hate speech) in user-generated content.

- You will not benefit personally by participating in this study, but what we learn about content moderation decision-making may be beneficial to other online users at large.

Your privacy will be protected, and confidentiality will be maintained to the extent possible. To protect your privacy, your identifying information will not be recorded. We will protect the confidentiality of the research data by storing the data in a UNC Charlotte subscribed survey platform (i.e., Qualtrics) and by separating your identifiable information (i.e., IP Addresses) from the research data. Other people may need to see the information we collect about you, to make sure that we are conducting this study appropriately and safely, including people who work for UNC Charlotte and other agencies as required by law or allowed by federal regulations. After this study is complete, study data may be shared with other researchers for use in other studies without asking for your consent again. The data we share will NOT include information that could identify you.

Participation is voluntary. You may choose not to take part in the study. You may start participating and change your mind or stop participating at any time.

If you have any questions, please feel free to contact us at [kwang17@charlotte.edu](mailto:kwang17@charlotte.edu). If you have questions about your rights as a research participant or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Office of Research Protections and Integrity at [uncc-irb@charlotte.edu](mailto:uncc-irb@charlotte.edu).

I am 18 years of age or older and I have read and understand the information provided. I understand that I may contact the researcher listed above if I have any questions. I freely consent to participate in the study.

- I agree
- I do not agree

Note: The consent form above is for the moderator user group only. The consent forms for users with and without content moderation experience user groups are different according to their different social media engagement experiences.

**APPENDIX D: PRE-EXPERIMENT SURVEYS****Users****Gender**

- Male
- Female
- Non-binary/third gender
- Prefer not to say

**What is your highest education level?**

- Less than a high school diploma
- High school degree or equivalent
- Some college but without a degree
- Associate degree
- Bachelor's degree
- Master's degree
- Doctorate
- Other (please specify)

**What is your occupation?**

- Full-time employee
- Part-time employee
- Self-employed
- Retired
- Student

- Unemployed
- Disabled, not able to work
- Other (please specify)

**What is your ethnic origin?**

- White
- Black/African American
- Native American (e.g., American Indian and Alaska Native)
- Native Hawaiian or Other Pacific Islander
- Asian
- Hispanic/Latino
- Middle Eastern or North African
- Other (please specify)

**Please rate your IT (Information Technology) expertise & experience**

- None
- Beginner (Little Experience)
- Intermediate (Working Knowledge)
- Proficient
- Expert

**Mental Stress: Please rate each of the following statements, ranging from never to very often.**

- In the last month, how often have you been upset because of something that happened unexpectedly?
- In the last month, how often have you felt that you were unable to control the important things in your life?

- In the last month, how often have you felt nervous and “stressed”?
- In the last month, how often have you felt confident about your ability to handle your personal problems?
- Please select “Sometimes” for this question (Attention Check)
- In the last month, how often have you felt that things were going your way?
- In the last month, how often have you found that you could not cope with all the things that you had to do?
- In the last month, how often have you been able to control irritations in your life?
- In the last month, how often have you felt that you were on top of things?
- In the last month, how often have you been angered because of things that were outside your control?
- In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

**Which of the following social media platform(s) have you seen user-generated content being removed or deleted by online communities or social media platforms in the past three months? (Select all that apply)**

- Meta(Facebook)
- Twitter
- Reddit
- Instagram
- LinkedIn
- YouTube
- Snapchat

- TikTok
- Pinterest
- Other (Please specify)
- None of the above

**How long have you been active on social media platform(s)?**

- Less than 6 months
- Less than 1 year
- 1-2 years
- 3-5 years
- 6-10 years
- More than 10 years

**In the xxx domain (XXX: Based on participants' selections on the pre-screening survey), what types of following violations that could subject your posted content to moderation?**

**(Select all that apply)**

- Hate speech and discrimination
- Harassment and cyberbullying
- Porn
- Copyright infringement
- Spam and phishing
- Misinformation and fake news
- Illegal activities (e.g., drug trafficking, terrorism, or other criminal behavior)
- Impersonation and identity theft (i.e., impersonating others or attempting to steal their identity)

- Irrelevant content (i.e., some topics that are out of the scope of discussion in the community)
- Other (please specify)

**How can you ensure that your post is not subject to removal or banning within online communities? (Select all that apply)**

- Following platform guidelines
- Reviewing similar posts
- Feedback from friends or peers
- Using content moderation tools
- Other (Please specify)

**Do you carefully read the community rules or policies before making a post on social media platforms?**

- Yes
- No

*If “Yes” is selected, proceed to the next question. Otherwise, skip the next question.*

**Are community rules or policies on social media platforms easy to understand?**

- Extremely easy
- Somewhat easy
- Slightly easy
- Neither easy nor difficult
- Slightly difficult
- Somewhat difficult
- Extremely difficult

**Do you carefully review some sample posts within the community that you intend to post before making a post on social media platform(s)?**

- Yes
- No

If “Yes” is selected, proceed to the next question. Otherwise, skip the next question.

**After reviewing sample posts within the platform or community, to what extent are you confident that your content aligns with the community rules or policies?**

- Extremely confident
- Somewhat confident
- Slightly confident
- Neither confident nor unconfident
- Slightly unconfident
- Somewhat unconfident
- Extremely unconfident

**Please provide an estimated number of posts that have been moderated (i.e., removed or deleted) by social media platforms or online communities.**

- Never
- 1
- 2~5
- 6~10
- 11~20
- 21~50
- More than 50



**Have you ever received an early warning before your content was banned or removed by moderators, online communities, or social media platforms?**

- Yes
- No

If “Yes” is selected, proceed to the next question. Otherwise, skip the next question.

**Why did you receive an early warning? (Please select all that apply)**

- Hate speech and discrimination
- Harassment and cyberbullying
- Porn
- Copyright infringement
- Spam and phishing
- Misinformation and fake news
- Illegal activities (e.g., drug trafficking, terrorism, or other criminal behavior)
- Impersonation and identity theft (i.e., impersonating others or attempting to steal their identity)
- Irrelevant content (i.e., some topics that are out of the scope of discussion in the community)
- Other (please specify)

**Have you ever received follow-up explanations for why your content was banned and/or removed by moderators, online communities, or social media platforms?**

- Yes
- No

If “Yes” is selected, proceed to the next question. Otherwise, skip the next question.

**Did the explanations explicitly state that your content contained one or more of the following violations after your content was banned and/or removed? (Please select all that apply)**

- Hate speech and discrimination
- Harassment and cyberbullying
- Porn
- Copyright infringement
- Spam and phishing
- Misinformation and fake news
- Illegal activities (e.g., drug trafficking, terrorism, or other criminal behavior)
- Impersonation and identity theft (i.e., impersonating others or attempting to steal their identity)
- Irrelevant content (i.e., some topics that are out of the scope of discussion in the community)
- Other (please specify)

**Have you ever appealed any content moderation decisions?**

- Yes
- No

If “Yes” is selected, proceed to the next two questions. Otherwise, skip the next two questions.

**Which of the following methods have you used for appeal? (Please select all that apply)**

- In-app/platform reporting (i.e., users can submit an appeal through the platform's reporting or feedback system)

- Appeal forms (i.e., users can fill out to provide more information about their content and request a review)
- Email support (i.e., users may be able to send an email to the platform's support team to appeal the moderation decision)
- Transparent communication (i.e., platforms may provide clear communication about the appeal process and the reasons for their content moderation decisions)
- Other (please specify)

**Please rate the effectiveness of the content moderation appeal process.**

- Extremely effective
- Somewhat effective
- Slightly effective
- Neither effective nor ineffective
- Slightly ineffective
- Somewhat ineffective
- Extremely ineffective

**Please rate your level of satisfaction with the content moderation appeal outcome.**

- Extremely satisfactory
- Somewhat satisfactory
- Slightly satisfactory
- Neither satisfactory nor unsatisfactory
- Slightly unsatisfactory
- Somewhat unsatisfactory

- Extremely unsatisfactory

## **Moderators**

### **Gender**

- Male
- Female
- Non-binary/third gender
- Prefer not to say

### **what is your highest education level?**

- Less than a high school diploma
- High school degree or equivalent
- Some college, no degree
- Associate degree
- Bachelor's degree
- Master's degree
- Doctorate
- Other (please specify)

### **What is your occupation?**

- Full-time employee
- Part-time employee
- Self-employed
- Retired
- Student
- Unemployed
- Disabled, not able to work

- Other (please specify)

**What is your ethnic origin?**

- White
- Black/African American
- Native American (e.g., American Indian and Alaska Native)
- Native Hawaiian or Other Pacific Islander
- Asian
- Hispanic/Latino
- Middle Eastern or North African
- Other (please specify)

**Please rate your IT (Information Technology) expertise & experience**

- None
- Beginner (Little Experience)
- Intermediate (Working Knowledge)
- Proficient
- Expert

**Mental Stress: Please rate each of the following statements, ranging from never to very often.**

- In the last month, how often have you been upset because of something that happened unexpectedly?
- In the last month, how often have you felt that you were unable to control the important things in your life?
- In the last month, how often have you felt nervous and “stressed”?

- In the last month, how often have you felt confident about your ability to handle your personal problems?
- Please select “Sometimes” for this question (Attention Check)
- In the last month, how often have you felt that things were going your way?
- In the last month, how often have you found that you could not cope with all the things that you had to do?
- In the last month, how often have you been able to control irritations in your life?
- In the last month, how often have you felt that you were on top of things?
- In the last month, how often have you been angered because of things that were outside your control?
- In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

**Which of the following social media platform(s) have you served as a moderator? (Select all that apply)**

- Meta(Facebook)
- Twitter
- Reddit
- Instagram
- LinkedIn
- YouTube
- Snapchat
- TikTok
- Pinterest

- Other (Please specify)
- None of the above

**How long have you been active on social media platform(s)?**

- Less than 6 months
- Less than 1 year
- 1-2 years
- 3-5 years
- 6-10 years
- More than 10 years

**How long have you been a moderator for online communities?**

- Less than 6 months
- Less than 1 year
- 1-2 years
- 2-5 years
- 5-10 years
- More than 10 years

**How many online communities have you served as a moderator?**

- 1
- 2
- 3
- 4
- 5
- More than 5



**How did you become a moderator? (Select all that apply)**

- Friend, family member, or social networking connection
- Recognized from other moderating experience
- Stand-out member of the community
- Availability at important times of day
- Volunteered or applied to be a moderator
- Support to the community (such as design, technical, financial, or other ways)
- Other (Please specify)

**What type(s) of training did you go through to become a moderator? (Select all that apply)**

- Platform-based instructions (e.g., tutorials or explanations of moderation resources)
- Community-based understanding of being in a community
- Advice from the head moderator or other members of the moderation team
- No training required
- Other (Please specify)

**Which of the following task(s) takes most of your time as a moderator? (Select all that apply)**

- Approving new members
- Contributing to community discussion
- Managing disruptive behaviors, general incivility, and targeted attacks
- Contributing to rule or guideline development for the community
- Developing filtering or banning systems for the community
- Reviewing reported or flagged user-generated content
- Warning offenders
- Providing explanations to users why they were punished

- Other (Please specify)

**What percentage of your moderation job requires a manual review of user-generated content?**

- 0~10% (0~10 out of 100 posts)
- 10~20%
- 20~30%
- 30~40%
- 40~50%
- More than 50 %

**What are some technical tools that you have used to make your moderation job easier?**

**(Select all that apply)**

- Community flagging and reporting (i.e., allowing users to flag or report content they find offensive or violating the guidelines)
- Filtering (i.e., using predefined lists of words or phrases to automatically flag and remove content)
- Black-listing (i.e., creating a list of items based on user reputation and trust systems)
- Hash databases (i.e., leveraging unique identifiers of known illegal or harmful content)
- Time-based restriction (i.e., restricting users from posting in certain events or situations when they might lead to an increased likelihood of rule violations)
- Machine learning and AI models (i.e., developing a customized model(s) in analyzing content and understanding the context, sentiment, and potential violations present in the language used.)

- Platform API or community-based API (i.e., a well-established model for automatic content moderation)
- Other (please specify)

**Which of the following types of content do you consider as violation(s) of your community rules or policies? (Select all that apply)**

- Hate speech and discrimination
- Harassment and cyberbullying
- Porn
- Copyright infringement
- Spam and phishing
- Misinformation and fake news
- Illegal activities (e.g., drug trafficking, terrorism, or other criminal behavior)
- Impersonation and identity theft (i.e., impersonating others or attempting to steal their identity)
- Irrelevant content (i.e., some topics that are out of the scope of discussion in the community)
- Other (please specify)

**Have you ever given users an early warning before banning or removing user-generated content?**

- Yes
- No

If “Yes” is selected, proceed to the next two questions. Otherwise, skip the next two questions.

**Did you use a standardized warning template or a customized warning message to send warnings to individual users?**

- Please provide a copy of the warning template here
- Please provide a sample of a customized warning message here

**What types of following content were considered as violations that would trigger warning(s) before banning or removing user-generated content? (select all that apply)**

- Hate speech and discrimination
- Harassment and cyberbullying
- Porn
- Copyright infringement
- Spam and phishing
- Misinformation and fake news
- Illegal activities (e.g., drug trafficking, terrorism, or other criminal behavior)
- Impersonation and identity theft (i.e., impersonating others or attempting to steal their identity)
- Irrelevant content (i.e., some topics that are out of the scope of discussion in the community)
- Other (please specify)

**Have you ever provided explanations to corresponding authors whose content is subject to removal and banning?**

- Yes
- No

If “Yes” is selected, proceed to the next two questions. Otherwise, skip the next two questions.

**Did you use a standardized explanation template or a customized explanation message to send explanations to individual users?**

- Please provide a copy of the explanation template here
- Please provide a sample of a customized explanation message here

**Which of the following content would trigger you to provide explanations to authors after user-generated content banning or removal? (Select all that apply)**

- Hate speech and discrimination
- Harassment and cyberbullying
- Porn
- Copyright infringement
- Spam and phishing
- Misinformation and fake news
- Illegal activities (e.g., drug trafficking, terrorism, or other criminal behavior)
- Impersonation and identity theft (i.e., impersonating others or attempting to steal their identity)
- Irrelevant content (i.e., some topics that are out of the scope of discussion in the community)
- Other (please specify)

**How often have the rules in your community changed over time?**

- Every month
- Every 6 months – 1 year
- Every 1-2 years
- Every 2-5 years

- Every 5-10 years
- More than 10 years

**Please describe a significant or memorable moderation experience.**

## **APPENDIX E: POST-REVIEW SURVEY**

**According to the community rules you just reviewed, are the community rules or policies easy to understand?**

- Extremely easy
- Easy
- Somewhat easy
- Neither easy nor difficult
- Somewhat difficult
- Difficult
- Extremely difficult

**According to the community rules you just reviewed, does the user-generated post violate any of those rules?**

- Yes
- No

**Which types of violations have you identified in the post? (Select all that apply)**

- None
- Hate speech and discrimination
- Harassment and cyberbullying
- Porn
- Copyright infringement
- Spam and phishing
- Misinformation and fake news
- Illegal activities (e.g., drug trafficking, terrorism, or other criminal behavior)

- Impersonation and identity theft (i.e., impersonating others or attempting to steal their identity)
- Irrelevant content (i.e., some topics that are out of the scope of discussion in the community)
- Other (please specify)

**Based on your review of the user-generated content, which moderation decision do you suggest?**

- Be Moderated
- Not Moderated

**Please rate the level of confidence in your moderation decision.**

- Extremely confident
- Confident
- Somewhat confident
- Neither confident nor unconfident
- Somewhat unconfident
- Unconfident
- Extremely unconfident

**Please rate each of the following statements and indicate how strongly you agree or disagree with each of the statements, with 1 indicating strongly disagree and 7 indicating strongly agree.**

- I think this content complied with the community rules/guidelines.
- I anticipate that online users will feel the necessity for moderation of this content.
- I anticipate that community moderators will moderate this content.



- Please select “Agree” for this question (Attention Check)
- I am confident that the community moderation system will moderate this content.
- Mental demand: I felt that the task was mentally demanding.
- Feelings of success (Performance): I felt successful accomplishing what I was asked to do.
- Negative emotions (Frustration): I felt irritated, stressed, and annoyed versus content, relaxed, and complacent during the task.
- I believe that external knowledge (e.g., user discussion, user profile, moderator’s experience) would facilitate me in making content moderation decisions.

## APPENDIX F: SUPPLEMENTARY TABLES

**Table 17: Selected UGC for the User Study**

Domain	Post - Title	Post - Body	Comments
Sports	GameDay vs Hurricane Ian	Okay so... do I brave Hurricane/Tropical Storm Ian and drive to Clemson for GameDay tomorrow or stay in Charlotte and prep for the storm? (That's apparently going to hit us? We're under tropical storm warning, I'm so confused.)	It will be wet and windy, but I don't see how a storm could screw up I-85 any worse than South Carolina's department of transportation... Everything should be pretty cleared out by tomorrow morning. This afternoon through the evening will be the worst of it. Current forecast is clear. Come on down.
	Advice for a newb	Looking to get in shape and join a gym. Have my eyes on a boxing gym and a bjj (i.e., Brazilian jiu-jitsu) place. Any advice/tips for a first sit through? Going to pick just one, have been wanting to get involved for years and finally pulling the trigger.	Just turning up is the main thing. If you go once or twice and like it, get your own gloves etc sooner rather than later. Do they do a free first session? I'd try both and see which you prefer. I recommend bjj before boxing if your going to compete. Good luck man
	What are the best YouTube channels to learn BJJ (i.e., Brazilian jiu-jitsu)?	I've been training about 2-3 times a week while also trying to learn different techniques and strategies through YouTube channels like knight jiu jitsu and Chewjitsu. What are some other channels you'd recommend for supplemental learning?	I'm a big fan of Jon Thomas' stuff on Youtube. Can't recommend enough. He's also active on here. Shoutout to /u/Macarrao09 ! Thanks man! I really like JonThomasbjj and Marcos Tinoco bjj. Andre Wiltse is putting out some great free material. William Tackett too.
	Easton Synergy Remake on sale today only on the Bauer site!	\$200 remade with newer upgraded material. Yes, this is real life.	Just so everyone understands, this is just a reskin and it doesn't have the exact same specs as the OG. From the product

---

		<p>page: "Built with updated materials for improved feel &amp; performance but holds true to the iconic look with the most iconic patterns that Synergy was known for." To be clear, it looks dope as fuck, but this is a mid-level Bauer stick with weapons-grade nostalgia bait attached to make us want it.</p> <p>I love my Easton sticks. So sad they discontinued the curve I like.</p> <p>You can all thank Gepetto at <a href="http://prostockhockeysticks.com">prostockhockeysticks.com</a> for forcing the issue here. Dude is the absolute man. His Easton remakes are probably more accurate also.</p>
Health	<p>How unhealthy would eating this be?</p> <p>How can I lose weight quickly?</p>	<p>Let's say for breakfast and lunch you eat 100% healthy....such as only salads, fish (like wild alaskan salmon), rice and lentils etc. but in the evening you eat 20 piece of chicken mcnuggets. You do this everyday. How unhealthy would that be?</p> <p>Basically today my boyfriend tried to lift me like bridal style, and then he told me I was really heavy and he could even hold me for more than a few seconds. He asked how much I weighed, so I told him I was around 130 lbs.</p>
		<p>It's suboptimal</p> <p>I liked the Dr. Greger sentiment of, healthy/unhealthy vs what? That diet would be healthier vs eating nuggets every meal... But vs eating clean every meal, probably not.</p> <p>Half your daily calorie intake would be coming from low nutritional deep fried chicken paste. So not great.</p> <p>This is ridiculous. At 5'9, 130lbs you are at the lighter end of a healthy bmi. Might be your boyfriend that needs a bit of strength training.</p> <p>Troll post or you two are just extremely immature. If true story your</p>

---

---

	<p>He said he was shocked I weigh this much, and that it's not healthy for a woman to weigh much more than 100 lbs. I feel so embarrassed and I know he prefers skinnier girls, so how can I lose weight quickly?</p>	<p>boyfriend might wanna hit the gym because lifting 130lbs is kinda the bare minimum expectation for most men.</p> <p>127 is on the lower end of normal for your height, bordering on underweight. Your boyfriend is just ridiculously weak.</p>
What type of vegans	<p>So feel like this will cause a lot of controversy (depending on what type of animal rights activist type you are you) but how do you we feel about PoC (i.e., person of color) vegans being highlight instead of mainstream white vegans? I feel like a lot of people love earthling Ed and are a big fan of his. But this type of veganism wouldn't be labeled intersectional. What do you feel your type of veganism is like?</p>	<p>I just don't want to hurt animals. Not sure how you'd label me?</p> <p>Veganism has roots in non white cultures so there's no reason we should be the poster child for it. We should absolutely promote PoC to dispell the myth that this is a white privilege movement.</p> <p>Vegans are vegans regardless of race, there is no white vegans vs poc vegans, that's crazy.</p>
ideas to help studio with unpleasant smell	<p>I typically attend a evening class same time each day. About 2-3 months ago we got a new member that became a frequent participant of that class time. The issue is the person just walking in has a very strong odor and it just gets worse as class goes on. The front desk know, they have done online reminders on facebook/ Instagram and even made signs to try and make it a reminder to try and keep your smells down. It doesn't help. My class time has started to shrink my guess since I</p>	<p>This is when the studio management needs to have that hard conversation with the individual. The other members pay too much to put up with that and if people are dropping the class then it's a business issue.</p> <p>If the class is literally losing participants, the studio manager or head coach needs to step up and talk to that person.</p> <p>Studio Manager just needs to straight up talk to the person, especially if they're losing business. It will be an awkward conversation but can be</p>

---

	<p>have seen someone walk in look at the board and said oh okay I will do rowing I am not near them. I do feel bad to be the person that is a negative person about this but on tornado days the whole gym becomes the smell. Any idea of how to help the studio deal with this issue. I don't want them to make them feel unwelcome everyone deserves to be there but it affects many people that attend the studio.</p>	<p>totally respectful and gentle.</p>
Fashion	<p>Sellers who have 1000s of listings... where do they get the stuff?</p> <p>I'm just curious - I've been noticing more sellers with thousands of brand new, tag-on items from brands like Reformation. What's going on? Clearly people aren't stocking their closets with thrifted items.</p>	<p>If a seller has a business license and storage for large pallet orders, they can buy wholesale and liquidation. There's too many clothes and too little time for it to be on the shelves to be purchased so manufacturers and stores try find other ways to get rid of it quickly.</p> <p>I have a closet like this. NWT/NWOT and popular brands. I get them directly from the retailer. They are overstock, samples or defective items. Most of them are fine to sell, some do have noticeable flaws and they are listed and photographed. All of them have slightly sticky zippers or a loose button. I fix and disclose. For most I can't even find any defect and are good as new obviously if a buyer finds something they can open a case but that's only happened to me twice in</p>

		<p>thousands of sales. These companies make sure that their items are free of any defects so any tiny detail gets tossed to sell for a cheaper amount and that's what I buy.</p> <p>Retail arbitrage is one way.</p>
Fall Sweaters	Where can I find good quality sweaters with interesting cuts or patterns? Ideally for about \$50, but that might be wishful thinking.	<p>Thrifting. I got several sweaters thrifting from local goodwills for fall quite recently. Overall I got around 8 items for fall for around \$30.</p> <p>Interesting styles and good quality, I'd suggest Madewell or Everlane. They are a little pricier but worth it.</p> <p>I just bought one from Etsy.</p>
Lands End 50% off Entire Order	I've never posted on here before so I'm sorry if this is against the rules. I just ordered from lands end and they have a 50% off your entire order sale going on right now. Got 3 flannels for \$48.	<p>Just FYI for anyone who isn't familiar with this brand, but this is "normal" prices for Land's End since they do the perpetual sale thing.</p> <p>Up to 50% off.</p> <p>After finally buying some stuff they're like a higher quality J crew. Lands end is so fucking amazing price wise wise and quality wise but a chunk of there stuff isn't that good looking or more for a older market. The only thing j crew beats them on is the amount of selection and how they look. If they got some designers from j crew and kept the sane quality and prices I'd never get anything anywhere else.</p>

---

	<p>How do I set the correct month on my watch?</p>	<p>I just got a new Invicta Aviator 38415 (i.e., a watch brand and model) and as I was trying to set the date I noticed something, the last day of the month is 31, so I keep moving the crown and again it has 31, and again, to no end... The current month is September which has 30 days, so I'm really confused as to how to set the date to have 30 days. Thanks!</p>	<p>You don't. The watch doesn't know what month it is. Think of it as a day counter. Set it to the day it is now, then if its a month with 31 days you are fine, any less you will need to adjust it at the end of the month.</p> <p>I am trying not to laugh at the thought how many witty comments will this post gather :D. To be frank though, this (changing the date at the end of months shorter than 31 days) is something you will have to live with. If you cannot be arsed, you can opt for a digital watch or gonna have to look a bit more for a watch with perpetual calendar. There are some fairly inexpensive quartz and mechanical choices from Seiko, Tissot and some.</p> <p>You set it to the correct date and then when the 1st of October rolls around change it to the 1st.</p>
Game	<p>Toxic players I simply cannot believe that there are toxic players in a game like Warframe (playerbase of 50k+, a free-to-play action role-playing third-person shooter multiplayer online game) which has so</p>	<p>"Just met this one toxic player out of my 9,99999999 hours of playing. I can count the amount of toxic players with one hand, I can't count past five. It was a Wukong player (99.9% of playtime on Wukong) and he said to us ""Poopoo Peepee"" while we were opening relics.</p>	<p>Dude why is this all allowed, he should be permabanned for that. "poopoo peepee"! How offensive! That player deserves punishment How dare! Dude, he should be banned, locked up and punished for at least 5 years for disturbing the peace and harmony of our fragile community. Now I am offended at intergalactic levels and I</p>

many nice players. Just wanted to share this since I figure someone else has probably dealt with these types of players before and will find this amusing."

Is this worth trying out?

Basically I got this game from prime gaming and I was wondering if it's worth trying it out? What should I expect if I do?

Best government type in 2022 and why do you think so?

I'm curious to see what others think the best government type is in 2022, maybe different ones for different reasons? Lets see your opinions!

need to go to the psychologist.

It's a great game. You really should give it an hour at least. I'm on hour 850.

In my unprofessional opinion it is worth playing you should expect lost of nuke drops and most likely kind people (there is some trolls not many though).

Yes. Online Fallout.

Never been a fan of republics but I did a Florence run and some of the stuff in there is so busted I didn't bother swapping when I had the opportunity.

It's mostly republics in the 2022 start date, so it's not much of a competition.

Monarchy, unlimited gov capacity due to centralizing mechanic.

Collect all the duct tape and fans you come across.

Hoard everything you find.

Play the game and don't worry about whether you "did it right".

---



**Table 18: Selected Subreddits for the Four Domains**

<b>Fashion</b>	<b>Health</b>	<b>Sports</b>	<b>Gaming</b>
r/Watches	r/orangetheory	r/bjj	r/Warframe
r/femalefashionadvice	r/vegan	r/MMA	r/Fallout
r/frugalmalefashion	r/loseit	r/hockey	r/fo76
r/poshmark	r/nutrition	r/CFB	r/eu4

**Table 19: The Results of the Homogeneity Test for the User Study**

<b>Dependent Variables</b>	<b>Familiar</b>	<b>Unfamiliar</b>
Content Review Time	(16.641)<.001***	(19.490)<.001***
Mental Workload	(5.239).001**	(5.863)<.001***
Perceived Confidence	(1.502).212	(2.053).105
Perceived Compliance	(2.556).054†	(.372).774
Perceived Users Moderation	(5.925)<.001***	(6.236)<.001***
Perceived Moderators Moderation	(2.842).037*	(4.241).005**
Perceived Systems Moderation	(1.221).301	(1.054).367

Levene's test scores are reported in parentheses, \*\*\*:  $p < .001$ ; \*\*:  $p < .01$ ; \*:  $p < .05$ ; †:  $< .1$ .

**Table 20: Community Rule and Post Examples**

#### **A Community Rule from r/CFB:**

No joking or trash talk about sexual assault or violence.

Only serious discussion is allowed about serious crimes, injuries, and death so jokes and trash talk stemming from these subjects are prohibited. If you're not sure, err on the side of caution. This includes, but is not limited to:

- \* Victim blaming
- \* Sexual assault & rape
- \* Domestic violence & other violent crimes
- \* Wishing for and celebrating injuries or death

<b>Moderated Posts</b>	<b>Unmoderated Posts</b>
<p>Post 1: Do you watch the Formula 1® Esports Series? What do you think about this type of competition?</p> <p>Hello, I would like to ask you what you think about this matter. I've been watching tournaments on various games since I was little, but I have mixed feelings about the electronic version of the formula. Is it really that popular? Do viewers like watching this? For me personally, it is strange to watch such a competition. Tell me what you think, I'm interested!</p> <p>(Irrelevant post)</p>	<p>Post 1: What radio should I get? Well, my team is buying some radios but my question comes here: do i have to use the exact same model that they have or can I choose another?</p> <p>To contextualize better: they are looking to get the Baofeng BF-888S but I'm thinking about the Baofeng Uv-5r. My question is: Will I be able to communicate with them using a different model? and if yes, what model should I get?</p> <p>Post 2:</p>

Post 2: Tennessee gonna get hate for this, "How Much Do You Weigh?" I find this an  
but my god are you guys annoying. It incredibly rude question yet people seem to  
doesn't matter at this point if you are or have no problem asking it in the gym.  
aren't ranked 1. You don't need to bitch  
and a t like you're the best team ever.

---