Conceptualizing Visual Analytic Interventions for Content Moderation

Sahaj Vaidya* NJIT
Jie Cai† NJIT
Soumyadeep Basu‡ NJIT
Azadeh Naderi§ NJIT
Donghee Yvette Wohn¶ NJIT
Aritra Dasgupta‖ NJIT

ABSTRACT

Modern social media platforms like Twitch, YouTube, etc., embody an open space for content creation and consumption. However, an unintended consequence of such content democratization is the proliferation of toxicity and abuse that content creators get subjected to. Commercial and volunteer content moderators play an indispensable role in identifying bad actors and minimizing the scale and degree of harmful content. Moderation tasks are often laborious, complex, and even if semi-automated, they involve high-consequence human decisions that affect the safety and popular perception of the platforms. In this paper, through an interdisciplinary collaboration among researchers from social science, human-computer interaction, and visualization, we present a systematic understanding of how visual analytics can help in human-in-the-loop content moderation. We contribute a characterization of the data-driven problems and needs for proactive moderation and present a mapping between the needs and visual analytic tasks through a task abstraction framework. We discuss how the task abstraction framework can be used for transparent moderation, design interventions for moderators' well-being, and ultimately, for creating futuristic human-machine interfaces for data-driven content moderation.

Keywords: Content Moderation, Social Media, Task Abstractions, Real-time Decision-Making

1 INTRODUCTION

Content moderation has emerged as a major challenge confronting the safety and acceptance of modern social media platforms, like Facebook, Twitter, YouTube, Twitch, etc. Companies are increasingly allocating valuable resources, in terms of building automated models [10, 21] and training or hiring human moderators [37, 40] to deal with the growing menace of negativity and toxicity online. Data-driven approaches, like those based on machine learning, have become necessary for automatically detecting content that violates community guidelines. However, these approaches remain opaque, unaccountable, and poorly understood [17]. Additionally, automated moderation is not sufficient due to the inherent complexity and ambiguity of moderation tasks [38]. In this paper, through interdisciplinary collaboration among researchers from social science, human-computer interaction, and visualization, we analyze how visual analytic interventions can empower content moderators with greater data-driven awareness about *who* to monitor, *what* kind of messages need attention, and *how* to ensure transparent implementation of rules and policies (Figure 1).

While the term "content" can be broadly interpreted, we focus our discussion on moderation activities in platforms that involve synchronous communication among users of live-streaming platforms like Twitch, YouTube, Discord, Clubhouse, etc. For moderators, the real-time interactions and the need to make consequential decisions with very limited lead time can often lead to high cognitive load [7] and take an emotional toll [43]. The conventional understanding is that moderation of online conversations in live-streaming platforms is inherently reactive, where moderators see and then react to content generated by users, typically by removing them. However, a significant portion of work performed by volunteer moderators is social and communicative in nature [40]: moderation decisions need to be transparently communicated to the users and there is a high consequence for decisions that can be perceived as unfair or incorrect. A shared vision among researchers in content moderation and visualization, who are co-authors of this paper, is that access to visual analytic techniques has a transformative potential on moderation activities in live-streaming platforms. Visual analytics tools and interfaces will allow moderators to summarize conversations, interpret and reason about why automated methods might have flagged certain messages, and ultimately, engage in a more *proactive, data-driven moderation process.*

To realize this vision, in this paper, we discuss the results from our six-month-long collaborative effort towards distilling the data-driven problems and corresponding visual analytic interventions for proactive content moderation. Following Munzner's nested model [30], we first analyze the content moderation goals and the associated data abstraction. Next, we contribute a visual analytic task abstraction framework for mapping the problems and challenges to concrete moderators' decision-making tasks. We also discuss the applications and implications of our framework for future research on data-driven, human-in-the-loop content moderation processes.

2 PROBLEM CHARACTERIZATION

Grimmelmann [18] defines moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." With the proliferation of online communities, the number of human moderators is vastly outnumbered by the user-generated content and the increased negativity, which is a concern when content creation is growing at an exponential speed and a core element of many of the major informational and social platforms today. To reduce online negativity, commercial platforms apply many techniques to filter abusive language, such as improving algorithms and applying automation tools [10, 21]. Though these automated tools can identify new instances of negativity such as harassment and hate speech with pattern matching, violators always seek ways to circumvent the algorithms and cheat the tools with variants [9]. To supplement algorithmic moderation, platforms also rely on human moderators to remove flagged content or review instances in context-sensitive situations.

*e-mail: ssv47@njit.edu.
†e-mail: jie.cai@njit.edu.
‡email: sb2356@njit.edu.
§email: azadeh.naderi7@gmail.com.
¶email: donghee.y.wohn@njit.edu.
‖e-mail: aritra.dasgupta@njit.edu.

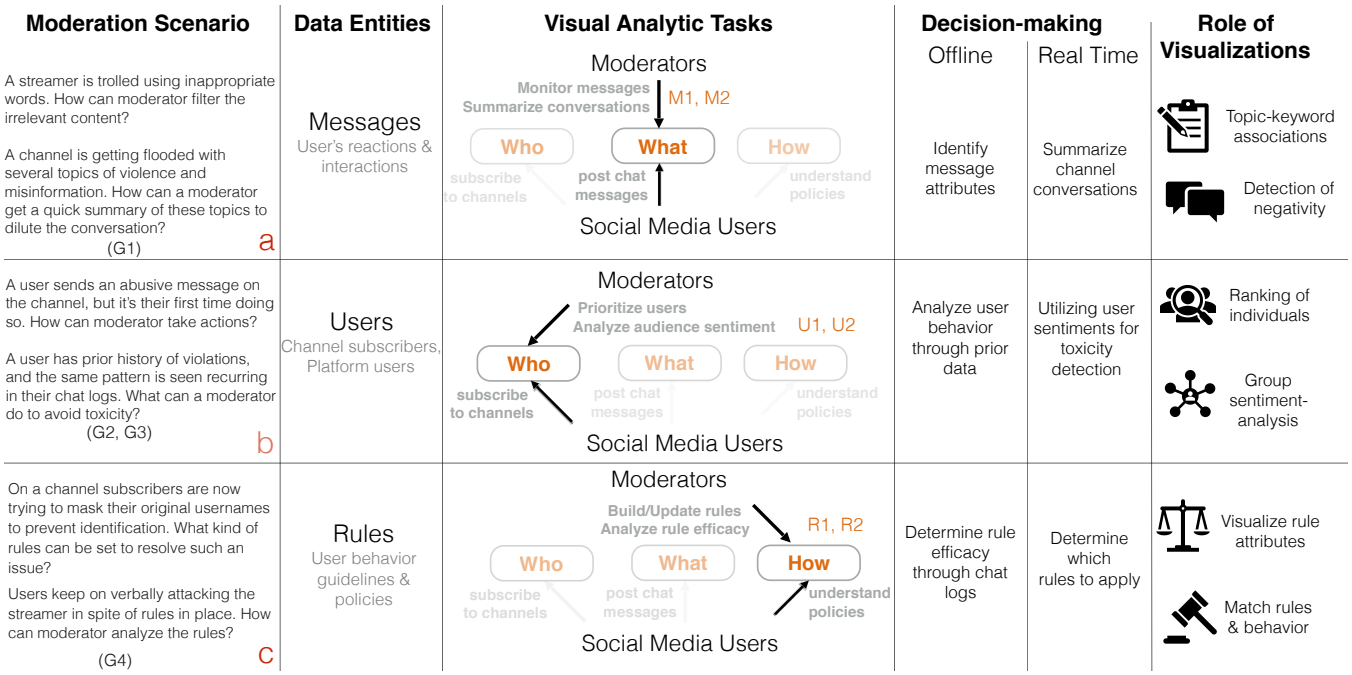

Figure 1: **Mapping between content moderation goals and visual analytic tasks.** Scenarios illustrating how moderators can leverage the expressive power of visualizations for making offline and real-time decisions about the *who*, *what*, and *how* dimensions of data-driven moderation.

2.1 Moderation Goals and Challenges

The moderation process involves how human moderators govern both content and community members and the standards development for the governance. There are mainly two threads of research about human moderators handling offensive content and users: proactively preventing mechanism and reactive punishing mechanism. A thread of research focuses on proactively preventing offensive behaviors via norm-setting such as setting a good example in the chatroom to influence other viewers in live streaming chat [7], or engaging in rule developments [40]. Another thread of research focuses on reactively removing content and punishing users, such as deleting content and banning users [11] and explaining and communicating rules to violators [7]. This thread of research also explores how moderators collaborate with automated tools [6, 21]. According to empirical research about content moderation and Grimmelmann's moderation goal to create a productive, open, and accessible online community [18], the moderation goals are summarized as follows:
G1: Get rid of harmful messages/comments and users, at the same time, curate valuable information in the community [11].
G2: Retain newcomers and foster the community via interaction and engagement [7].
G3: Distinguish between good and bad actors and punish the latter but avoid excessive punishment towards unintentional violators or first-time violators [4, 8].
G4: Develop and clarify the moderation guideline and maintain the transparency of moderation [5, 23].

Our focus is to explore the influence of data-driven methods on human moderator's decision-making process. We address questions such as: How do moderators use aggregate information about users and their messages to guide their decision-making? How can visualization help moderators to facilitate intervention in context-sensitive situations?

2.2 Data Abstraction

To address the goals and research questions, we first describe the specific data entities that be considered as the building blocks of algorithmic moderation tools and that can be used to develop human-in-the-loop moderation tools. The moderation process comprises three main data entities: Messages, User Profiles, and Rules.

Messages (M): Messages encode the response of the users towards the actual content and their interactions with other users of a channel. Moderators can leverage text-based analysis of messages to analyze and monitor the conversations on the channel. This monitoring of chat helps to flag messages and detect violations of established rules or signals of abusive content. In live streaming environments such as Twitch [7], this is cognitively demanding as a large volume of messages is posted in a short span of time making it difficult for moderators to make timely decisions. Platforms often employ crowdsourced moderation strategies in the form of flagging tools that allow users to express concerns about potentially offensive content [25]. This strategy does not perform effectively in the context of real-time moderation because of the time gap between reporting bad content and reviewing it [43].

User Profiles (U): Users of social media platforms are central to the moderation process. The goal is to encourage user participation in online communities by providing them value-based content. Moderators can characterize the users based on their engagement in online activities. On the other hand, moderators can also punish those users who do not abide by the norms. Online communities do not share the users' information of each micro community with customized community guidelines As for live voice moderation, it is even more challenging to collect voice information for moderators to make decisions [22] such as Discord Voice chat and Clubhouse. The history of a user's prior behavior is obtained from archival data and does not change dynamically with time.

Rules (R): Rules define the code of conduct regarding a user's online behavior. Moderators take data-driven decisions matching user profiles with rules that are set for a particular stream. The severity of punishment varies based on the user profile and the importance of the rule [5]. A key challenge faced by moderators is to go over real-time messages and fine-tune their mental model for applying chat rules by assessing the severity of the violation [6]. Similar to user profiles, rules defining online behavior are mostly static and do not evolve in real-time.

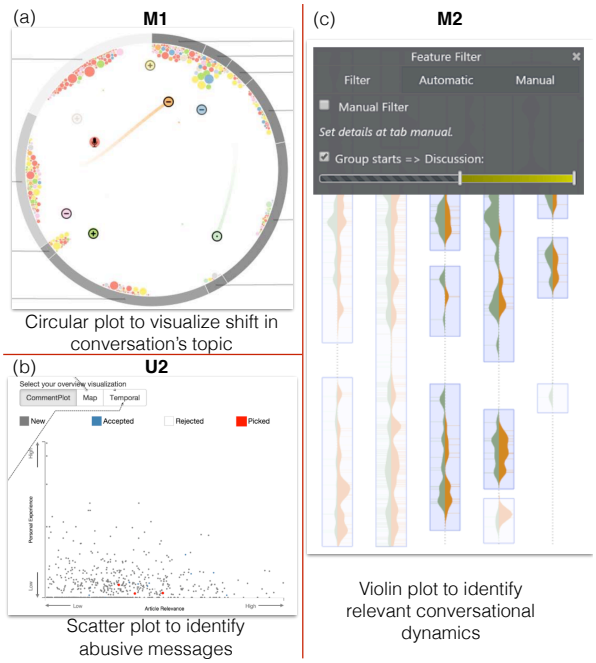

Figure 2: **Examples of techniques for visualizing conversations.** (a) ConToVi identifies the shifts in conversation topics for navigating the online discussions [15], (b) Park et al. describes a user-centric design approach to select flagged comments with the help of comment analytic scores which can detect only a small set of messages because of keyword limitations [36], (c) Seebacher et al. displays relevant conversational dynamics while fading out the non-relevant ones [39].

3 Visual Analytic Task Abstraction

In this section, we map the moderation goals to entity-level visual analytic tasks, focusing on message analysis (M1, M2), user profiling (U1, U2), and rule building (R1, R2). We discuss the role of analytical methods and visualization for addressing moderation goals using examples from the visual analytics literature (*a detailed list included in the supplemental material*) and also highlight key gaps and challenges.

3.1 Message Analysis Tasks

M1: Reasoning about Violations: The real-time nature of the streaming data requires the moderator to maintain the pace of processing the continuous data and analyze it. This task aims to achieve the goal of filtering out abusive messages and provide users with qualitative content (**G1**). The task of determining violations involves two components: *monitoring messages to identify anomalies* and *identifying message attributes*. If we look at the two scenarios in Figure 1a, monitoring helps to flag messages based on their content. Identifying message characteristics is another way to detect patterns in the chat streams [2]. Annotating and deleting spam messages [32] through user intervention can be helpful to recognize signatures of messages for flagging to identify change.

The dynamic nature of streaming data makes it difficult to analyze the chats for offensive content and make timely decisions. Therefore, platforms are increasingly turning to automated systems to detect abusive content within a shorter duration [35]. Moderators can leverage the visual analytics methods to review contextual information. Several visual analytic approaches provide support to analyze real-time content using interactivity for anomaly detection in the message streams [15, 16]. The sedimentation view [15] shown in Figure 2a is an example of representing only the relevant pieces of communication from the entire conversation. T-Cal [16] is a timeline-based approach that highlights areas with high information density. This provides a visual cue to the moderator to monitor those highlighted regions closely.

The challenge for visualizing the dynamic of chat streams lies in the automatic identification of appropriate cues from the message dynamics. Additionally, because of various nuances in vocabulary and language, the process of automated content moderation suffers from the limitation of deriving contextual insights from the messages.

M2: Summarizing Real-time Conversations: Communication via stream chat involves interaction between multiple users containing a large volume of messages. Because of this information density, simplification is required. The topic summary identified using this task help moderators to set the tone of the conversation and maintain the regulations to provide a positive atmosphere for online discussion(**G2**).

Generating a summary of conversations involves two components: *text summarization* and *topic identification*. Automatic summarization of messages in a channel is valuable to the moderators but it has certain limitations. The summarization of conversation necessitates addressing the trade-off between information loss (e.g., leaving out potentially relevant information) and abstraction of key topical patterns so that harmful content can be quickly detected [1, 29]. Visual analytic interfaces (Figure 2c) can help identify the shifts in conversation topics for navigating online discussions. Approaches like trains of thoughts [41] and conversation clusters [3] group messages of the same theme together. These approaches can allow moderators to have a better understanding of the topics of conversation.

Using visual analytic systems to explore the conversations based on topics is helpful to extract relevant linguistic features from the chat. With all the approaches discussed above, scalability and adaptation of visualizations to changes in dynamic conversation streams [13] remain a challenge. This challenge needs to be handled by assessing the perceptual limitations of the alternative designs in communicating the number, frequency, and degree of changes in conversation streams.

3.2 User Profiling Tasks

U1: Ranking user profiles using prior history: This task aims to analyze data about users' past online behavior. This includes *analyzing users' historical data* and *ranking users based on their profiles*. Studying user's online behavior helps moderators identify the type of users they need to pay special attention to (**G3**). Consequently, this task can help moderators to foster a healthy community of users and retain their participation (**G2**).

The collection of user's historical data incorporates the study of their characteristics, interests, ratings, usage patterns, and chat logs to recognize behavioral patterns. Scoring profiles based on recently opened accounts and user activities [6] helps understand the punishment based on the context and weight of the violation. This further helps to determine the type of punishment for the user when situations arise as described in Figure 1b. An example of this is the work by Oliva et al. which ranks user profiles based on the toxicity level [34]. This can be useful for a moderator to monitor highly sensitive users based on their profile toxicity scores.

For the methods described above, this task is often limited by the ability of automated programs to process the numerous amounts of user's archival data and the algorithm used for ranking.

U2: Reasoning about audience sentiment: Research in natural language processing (NLP) has extensively investigated the problem of sentiment analysis, where sentiment is generally classified as positive, negative, or neutral. The task of reasoning about audience sentiment can help moderators *determine the level of toxicity* in a channel. Models for sentiment analysis can be used to detect the reasons for negative audience sentiment and help moderators better understand the semantics of each message to avoid excessive punishment (**G3**).

Moderators often face challenges when detecting abusive content

from online communications. They try to mitigate the problem by implementing refined filters [20]. But these systems often fail due to a lack of correlation between the semantic space and user sentiments. Several authors have proposed solutions for semi-automatic detection of toxicity. For example, the interface CommentIQ shown in Figure 2b enables flagging of messages based on keywords [36]. However, this approach can detect only a small set of messages because of keyword limitations. Nobata et al. [33] trained a machine learning model to identify hate speech using a custom-built lexicon. All these lexicons have drawbacks that arise from the limited set of vocabulary. Chatzakou et al. [12] considered sentiment as an input to their neural network but did not discuss the impact on user perception. Visual analytic techniques can enable moderators to draw inferences based on group sentiment within their audiences.

3.3 Rule-Building Tasks

R1: Augmenting the Rule Book: Rules are made to educate the platform users about norms for expected behavior. These rules include respecting others in the community, following the guidelines made by the community, etc. The task of augmenting the set of rules includes *building rules* and *modifying rules* based on a user's behavior. This task aligns with **G4** at the broad level, helping the community moderators understand how rules match with violations and add community-specific rules based on the streamer's requirements.

Modifying the rules can be grounded in assessing users' relative standing in the community. This includes analyzing the history of past rule-breaking cases and the severity of the rules that have been broken [5, 7, 43]. Using a set of rules allows the creation of automatic filters that remove the unwanted content by comparing it with existing rules. It is important to visualize the user involvement before and after posting the rules periodically. The distribution of messages shown per participant before and after posting rules by a chatbot in Kim et al.'s work [24] is an example of this visualization. However, a shortcoming of these methods is that they cannot detect the dynamic reactions to the rules and thus can hinder real-time filtering and decision-making.

R2: Determining Rule Efficacy: The task of determining the effectiveness of rules fulfills the purpose of developing moderation guidelines to maintain transparency of content moderation (**G4**), by comparing the existing rules with violations and identifying the effective rules and the missing parts. For this, rule-based techniques help moderators to detect abusive content and filter out those messages. It helps moderators to revise the guideline and regulate situations like Figure 1c. This task composes of *inspecting rule accuracy* and *categorizing rules based on the severity*.

Most of these rules are designed manually. Kontostathis et al. proposed a rule-based system to automatically detect harmful messages in relay chat [26] using an existing set of rules. Visualizing the numerical profile scores and the rule-breaking severity scores of the users will help the moderators understand the similarity and differences among "good" or "bad" rules. It will be beneficial in both cases - a popular channel crowded with users and also newer channels where the moderator lacks prior experience. Maintaining and modifying the rules is a time-consuming process. There may be a message containing conflicting keywords in an appropriate context, but it can be marked as offensive based on the rules. Whereas in other cases, a message containing abusive content may still be accepted and marked as appropriate. Visual analytic interventions can help detect and fill these gaps by enabling provenance-based retrieval and validation of rules.

4 Applications of Task Abstraction Framework

In this section, we discuss how our task abstraction framework can be applied in practice to addressing open problems in visualization design and human-machine interface development.

Ensuring moderation transparency: Using the visual analytic tasks, moderators can examine the rules and criteria through the lens of transparency. Many content moderation systems on social media sites are black-box in nature; users have to figure out on their own about why content is removed [31]. This lack of transparency can create barriers for user engagement for volunteer moderators who need to proactively communicate to users about guidelines and action consequences. In such a high-consequence setting, tasks like M1, R1, U2 can allow moderators to achieve a balance between preserving the safety of their communities and mitigating the effects of negative responses. Visual analytic interventions can help achieve this balance by using evidence-based communication [42] of moderation actions between moderators and platform users.

Facilitating social and communicative moderation: Though automated moderation tools can potentially detect signals of violation within a large volume of text stream, moderators are still irreplaceable, To foster and grow online communities, volunteer moderators play multiple roles with social and communicative attributes [7, 43] and are related to tasks U1 and U2. Our framework can guide designers to develop visualization tools to meet the needs of different communities of volunteer and commercial content moderators. For example, volunteer moderators have more flexible guidelines for their communities while commercial moderators have to follow the universal platform policy. This implies that volunteer moderators are in greater need of tools for mining users' behavior (M1, U2) and adapting their rules (R1) accordingly. On the other hand, commercial moderators can benefit from rule evaluation tasks (R2) for data-driven validation of their policies.

Designing for moderators' well-being: Along with reducing the cognitive load of moderators, realizing tasks like M1 and M2 enables exploration of the visualization design space for addressing psychological implications of content moderation. Decision-making about negative content often leads to psychological and emotional distress. Though reducing distress is not the primary goal of moderation, it can be embedded in the visualization design space. Visualization design strategies that optimize emotional impact [19] can reduce moderators' exposure to problematic content and can work as interventions to mitigate distress [14, 27].

Instantiating human-machine moderation interfaces: Mainstream moderation tools list violators and violations with limited explanations, and more importantly, lack proactive moderation capabilities. Our task abstraction framework can be applied for instantiating human-machine collaboration interfaces, where human and machine efforts are complementary, leading to optimal task performance as a team [28]. Moderators can ground their exploration process based on facets of interest (person, topic, region, flagged content, etc.), flag particular users or sensitive topics, while a machine learning model can be trained for learning from their interactions and suggesting corrective actions.

5 Conclusion and Future Work

Our work introduces a visual analytic task abstraction framework for addressing data-driven problems in proactive content moderation. We discuss the implications of the visual analytics framework for influencing the future of transparent and communicative moderation practices. As a next step, we plan to realize our proposed visual analytic tasks within existing content moderation workflows. We will conduct empirical studies to evaluate how visual analytic interventions and the resulting human-machine interfaces help reduce the cognitive load and emotional toll of content moderators.

6 Acknowledgement

This work was funded by the National Science Foundation (award number 1928627).

REFERENCES

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Twitcident: fighting fire with information from social web streams. In *Proc. International conference on world wide web*, pp. 305–308, 2012.

[2] P. H. Adams and C. H. Martell. Topic detection and extraction in chat. In *2008 IEEE international conference on Semantic computing*, pp. 581–588. IEEE, 2008.

[3] T. Bergstrom and K. Karahalios. Conversation clusters: grouping conversation topics through human-computer dialog. In *Proc. Conference on Human Factors in Computing Systems (CHI)*, pp. 2349–2352, 2009.

[4] L. Blackwell, M. Handel, S. T. Roberts, A. Bruckman, and K. Voll. Understanding" bad actors" online. In *Extended Abstracts, Conference on Human Factors in Computing Systems*, pp. 1–7, 2018.

[5] J. Cai, C. Guanlao, and D. Y. Wohn. Understanding rules in live streaming micro communities on twitch. In *Proceedings of ACM International Conference on Interactive Media Experiences (IMX'21)*, 2021.

[6] J. Cai and D. Y. Wohn. Categorizing live streaming moderation tools: An analysis of twitch. *International Journal of Interactive Communication Systems and Technologies (IJICST)*, 9(2):36–50, 2019.

[7] J. Cai, D. Y. Wohn, and M. Almoqbel. Moderation visibility: Mapping the strategies of volunteer moderators in live streaming micro communities. In *Proc. ACM Conference on Interactive Media Experiences (IMX)*, 2021.

[8] J. Cai and D. Yvette Wohn. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2):25, 2021. doi: 10.1145/3479554

[9] S. Chancellor, J. A. Pater, T. Clear, et al. thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proc. Computer-Supported Cooperative Work & Social Computing*, pp. 1201–1213, 2016.

[10] 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.*, 3(CSCW):1–30, 2019.

[11] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, et al. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.

[12] D. Chatzakou, N. Kourtellis, J. Blackburn, et al. Mean birds: Detecting aggression and bullying on twitter. In *Proc. 2017 ACM on web science conference*, pp. 13–22, 2017.

[13] A. Dasgupta, D. L. Arendt, L. R. Franklin, et al. Human factors in streaming data analysis: Challenges and opportunities for information visualization. In *CGF*, vol. 37, pp. 254–272. Wiley, 2018.

[14] L. Derick, G. Sedrakyan, P. J. Munoz-Merino, et al. Evaluating emotion visualizations using affectvis, an affect-aware dashboard for students. *Journal of Research in Innovative Teaching & Learning*, 2017.

[15] M. El-Assady, V. Gold, C. Acevedo, C. Collins, and D. Keim. Contovi: Multi-party conversation exploration using topic-space views. In *Computer Graphics Forum*, vol. 35, pp. 431–440. Wiley Online Library, 2016.

[16] S. Fu, J. Zhao, H. F. Cheng, H. Zhu, and J. Marlow. T-cal: Understanding team conversational data with calendar-based visualization. In *In Proc. Conference on Human Factors in Computing Systems*, pp. 1–13, 2018.

[17] R. Gorwa, R. Binns, and C. Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2020.

[18] J. Grimmelmann. The virtues of moderation. *Yale JL & Tech.*, 17:42, 2015.

[19] L. Harrison, R. Chang, and A. Lu. Exploring the impact of emotion on visual judgement. In *In Proc. IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 227–228. IEEE, 2012.

[20] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

[21] 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*, 26(5):1–35, 2019.

[22] J. A. Jiang, C. Kiene, S. Middler, J. R. Brubaker, and C. Fiesler. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[23] 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*, 4(GROUP):1–35, 2020.

[24] S. Kim, J. Eun, C. Oh, et al. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proc. Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[25] K. Klonick. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598, 2017.

[26] A. Kontostathis. Chatcoder: Toward the tracking and categorization of internet predators. In *Proceedings of Text Mining Workshop 2009 Held in Conjunction with the Ninth Siam international Conference on Data Mining (SDM 2009). SPARKS, NV. MAY 2009.* Citeseer, 2009.

[27] J. Liem, C. Perin, and J. Wood. Structure and empathy in visual data storytelling: Evaluating their influence on attitude. In *Computer Graphics Forum*, vol. 39, pp. 277–289. Wiley Online Library, 2020.

[28] C. Lyn Paul, L. M. Blaha, et al. Opportunities and challenges for human-machine teaming in cybersecurity operations. In *In Proc. Human Factors and Ergonomics Society*, vol. 63, pp. 442–446. SAGE Publications, 2019.

[29] A. Marcus, M. S. Bernstein, O. Badar, et al. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proc. CHI conference on Human factors in computing systems*, pp. 227–236, 2011.

[30] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[31] S. Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, 2018.

[32] T. K. Naab, A. Kalch, and T. G. Meitz. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, 20(2):777–795, 2018.

[33] C. Nobata, J. Tetreault, A. Thomas, et al. Abusive language detection in online user content. In *Proc. WWW*, pp. 145–153, 2016.

[34] T. D. Oliva, D. M. Antonialli, and A. Gomes. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732, 2021.

[35] E. Papegnies, V. Labatut, R. Dufour, and G. Linares. Graph-based features for automatic online abuse detection. In *International conference on statistical language and speech processing*, pp. 70–81. Springer, 2017.

[36] D. Park, S. Sachar, N. Diakopoulos, and N. Elmqvist. Supporting comment moderators in identifying high quality online news comments. In *Proc. Conference on Human Factors in Computing Systems*, pp. 1114–1125, 2016.

[37] S. T. Roberts. Commercial content moderation: Digital laborers' dirty work. 2016.

[38] S. T. Roberts. Digital detritus:'error'and the logic of opacity in social media content moderation. *First Monday*, 2018.

[39] D. Seebacher, M. T. Fischer, R. Sevastjanova, et al. Visual analytics of conversational dynamics. *arXiv preprint arXiv:2105.04897*, 2021.

[40] J. Seering, T. Wang, J. Yoon, and G. Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.

[41] D. Shahaf, C. Guestrin, and E. Horvitz. Trains of thought: Generating information maps. In *Proc. WWW*, pp. 899–908, 2012.

[42] S. Vaidya and A. Dasgupta. Knowing what to look for: A fact-evidence reasoning framework for decoding communicative visualization. In *2020 IEEE Visualization Conference (VIS)*, pp. 231–235. IEEE, 2020.

[43] D. Y. Wohn. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proc. Conference on human factors in computing systems (CHI)*, pp. 1–13, 2019.