2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P) | 978-1-6654-6512-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/EuroSP57164.2023.00056

SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice

Mohit Singhal* Chen Ling† Pujan Paudel† Poojitha Thota* Nihal Kumarswamy*
Gianluca Stringhini† Shirin Nilizadeh*
* *The University of Texas at Arlington*
† *Boston University*
(mohit.singhal,poojitha.thota,nihal.kumarswamy)@mavs.uta.edu, shirin.nilizadeh@uta.edu
(ccling,ppaudel,gian)@bu.edu

***Abstract*—Social media platforms have been establishing content moderation guidelines and employing various moderation policies to counter hate speech and misinformation. The goal of this paper is to study these community guidelines and moderation practices, as well as the relevant research publications, to identify the research gaps, differences in moderation techniques, and challenges that should be tackled by the social media platforms and the research community. To this end, we study and analyze fourteen most popular social media content moderation guidelines and practices, and consolidate them. We then introduce three taxonomies drawn from this analysis as well as covering over two hundred interdisciplinary research papers about moderation strategies. We identify the differences between the content moderation employed in mainstream and fringe social media platforms. Finally, we have in-depth applied discussions on both research and practical challenges and solutions.**

1. Introduction

Social media and online communities allow individuals to freely express opinions, engage in interpersonal communication, and learn about new trends and new stories. However, these platforms also create spaces for uncivil behavior and misinformation. Uncivil behavior is defined as explicit language, derogatory, or disrespectful content, which has become native on online platforms [66], [159], [331]. Misinformation is defined as false or inaccurate information [319], which has become rampant on social media platforms.

Uncivil behaviors like online harassment have a severe impact on users; social media provides anonymity, which can lead to disinhibition, deindividuation, and a lack of accountability, that can lead to anxiety and depression; or even suicide [68], [81], [120], [294]. Misinformation significantly impacts users with undesirable consequences and wreaks havoc on wealth, democracy, health, and national security [152]. Misinformation, conspiracies, and coordinated misinformation campaigns were prevalent throughout the COVID-19 pandemic [212], [276]. Such low-quality posts can also drown out useful content and exhaust the limited attention of users [180].

Since the early incubation of online communities, scholars and community managers alike reminisce over how to best manage online content and how to enable constructive, civil conversations among the users [83], [108]. However, there is no unified method for content moderation among the different social media platforms. Some employ more restrictive rules, while others emerge promising no or minimal moderation. For example, fringe social media platforms, such as Parler and Gab, have very minimal restrictions and they rarely ban users [33], [35].

Content moderation consists of several levels, including community guidelines, techniques to detect violations, and then policy enforcement. On each platform, content moderation is not constant but evolves as new challenges emerge, or it becomes clear that the in-use methods are not sufficient to protect information integrity. For example, during the 2020 Presidential elections and COVID-19 with the emerge of huge amount of misinformation and fake news on Twitter, the platform started using warning labels on posts to counter such content [2], [3]. In this paper, we study and categorize the topics covered in content moderation research and investigate the current state of content moderation on several social media platforms, focusing on the enforcement of moderation policies and also the community guidelines and how platforms define and moderate different types of content. With this analysis, we aim to obtain a comprehensive vision of content moderation from the points of view of both the research community and real-world practices. In particular, we try to answer the following research questions: **RQ1**: In what aspects and how does the research community study content moderation? What are the research gaps that need to be filled? **RQ2**: How does content moderation work in practice? What content do different social media platforms try to moderate, and how are the content moderation policies defined, implemented, and enforced? What are the practical and research gaps that need to be filled?

Studying and investigating these two research questions helps us in understanding all major components of the content moderation framework. Systemizing publications studying these components, and identifying their overlaps and differences, can help understand the research gaps (RQ1) in each component. Furthermore, no other work has provided a systemization of content moderation which are practiced in social media platforms (RQ2). To answer these research questions, we collate more than two hundred plus research papers describing the ever-growing changes to the content moderation practices of social media platforms and their impact on the end-user, and also investigate fourteen social media platforms to understand the current state of content moderation

DOI 10.1109/EuroSP57164.2023.00056

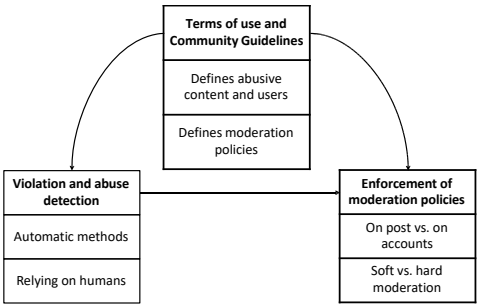

Figure 1: Content Moderation Framework

practices. With these analyses, we create three granular taxonomies, which explain and categorize the moderation policies, the types of content that is moderated, and the comprehensiveness of the provided community guidelines. To the best of our knowledge, ours is the first paper that systematizes content moderation policies in social media platforms. Papers studying social media as a business (i.e., platforms as exploitative data-gathering/ surveillance systems for advertising) are out of scope of this paper.

In our analysis, we found that there are still inconsistencies in the way social media platforms moderate content and how they define the guidelines. For example, while certain categories of abuse are banned across all the platforms, there is no consensus among platforms on what content to moderate and what not to. Hence, moderation happens arbitrarily. From our analysis of previous literature on *misinformation* and *hate speech* detection and the effectiveness of guidelines and enforcement methods, we identified and discussed several research gaps. We believe that our findings will help not only the computer research community but also the social media companies, to create a more inclusive and transparent moderation process.

2. Content Moderation

Moderation is the governance mechanism that structures participation in an online community to facilitate cooperation and prevent abuse [145]. It determines which posts and users are allowed to stay online and which are removed or suspended, how prominently the allowed posts are displayed, and which actions accompany the content removals, i.e. chance to appeal the decision [134]. Figure 1 shows the content moderation framework, which we generated based on an extensive review of previous works and community guidelines. This framework includes three interconnected components: (1) terms of use and community guidelines which define abusive content and behavior, as well as moderation policies, (2) violation and abuse detection methods, and (3) enforcement of the moderation policies. Terms of use and community guidelines show how the platforms detect violations and abuse, and dictate the policies that the platform tries to enforce. Throughout this paper, we discuss the content moderation and literature with regards to each of the components.

2.1. Terms of Use and Community Guidelines

Social media platforms define their content moderation policies alongside their privacy policies, copyright etc., in their *terms of service* and *community guidelines*. Typically, whenever a user joins a social media platform, they come across *terms of service* and they are required to provide affirmative agreement to register an account and use the platform. Terms of service are not particularly useful as an education tool as they are likely written in *legal* terms and previous works have found that users rarely read them [293]. In order for the users to better understand the rules, these platforms provide them as *community guidelines/ standards*, which are established in layman's terms. These terms and guidelines specify the types of content that are prohibited on the platform, and the actions that will be taken once a violation is detected (i.e., moderation policies). Some examples of such content are *child sexual exploitation, terrorism, and pornography*, which by law are required to be removed. Other types of prohibited content are *malicious content*, which includes spam, malware, and phishing URLs purposefully spread on social media platforms to gain more victims [84], [286], [287], [334]. While there are numerous works on the moderation of such content, these works are out of the scope of this paper. Instead, we investigate research that focus on content, such as *fake news*, *misinformation*, and *offensive and hate speech*. These are emerging and gray areas, with no universally accepted definitions, which contribute to controversial discussions on the trade-off between moderation and freedom of speech.

2.2. Violation and Abuse Detection

To detect violations as well as abusive content and behavior, social media platforms rely on a combination of human moderators and automated algorithms, which include heuristic-based and rule-based techniques as well as sophisticated machine learning-based models.

Human-based Moderation: Human moderators can be paid workers or volunteers who not only identify and vet abusive content, but may also enforce the moderation policies, e.g., by directly removing such content. Some platforms, like Twitter, Facebook, etc., employ a large group of freelancers, who work on contracts [11], [249], while others, for example, Reddit, rely on volunteer moderators, that are typically selected from the most actively involved users in the community [165], [209]. Reddit also gives moderators the power to create and enforce local rules that sit below the broader rule set of the platform. 4chan is similar to Reddit, where there are additional rules specific to each board. Some platforms, such as Facebook, Youtube, Snapchat, etc., also rely on users, to flag content that they deem unfit for the readers or they violate the community guidelines [99]. These platforms may then use heuristics or models to employ moderation policies based on these flags, e.g., they might have a simple rule that a post is deleted if enough users flag it, or they might send such flagged content to their human moderators, who closely vet and enforce the policies.

Algorithmic-based Moderation: Human-based moderation though effective comes at a cost of time, labor, and liability for the host company, and also at the emotional cost of the workers since they are at the frontlines [99], [249]. To minimize such trade-offs, social media companies are increasingly deploying rule-based techniques and machine learning models to automate the process of

identifying problematic material [19]. For detecting different types of violation and abusive content, social media platforms develop a different set of algorithms and methods. For example, YouTube relies heavily on its automated flagging systems to remove offensive comments [46]. Perceptual hashing [226] is used to automatically detect pornographic images and videos, adult nudity, etc., which involves fingerprinting some perpetually salient feature set of the content. Combating terrorism content is challenging and requires instantaneous removal from the platform. Companies such as Facebook, Google, Twitter, and Microsoft created a group called Global Internet Forum to Counter Terrorism (GIFCT), which maintains a Shared Industry Hashed Database (SIHD) [15], [22]. To detect misinformation in images, Meta has deployed SimSearchNet++, which can detect variations in an image with high precision. Pre-trained NLP language models such as BERT [61], [221], [256], RoBERTa [197] & XLM-R [98] have been recently proposed to detect hate speech and are used by Facebook [31], [32]. In academia, most hate speech and toxicity detection studies use Google's Perspective API to detect toxicity in text [141]. The API uses machine learning techniques and a manually curated dataset of text to identify toxicity. Most social media platforms do not publicly share the details of their tools used to detect misinformation or offensive content, however, in the literature, AI-based models are proposed to detect specific types of misinformation about COVID-19 and vaccinations [186], [273], security tools [287], hate speech, and offensive comments [101], [333], fake reviews [207], [222], etc.

Human-in-the-loop moderation: Relying solely on algorithms to moderate also has some limitations, e.g., the performance of AI-based models can significantly drop for out-of-distribution examples that differ from the training data [73], [169]. Hence, some studies have proposed using human-in-the-loop moderation [189], [203], which refers to interactive training paradigm where the AI receives input from the human to improve its performance. Moreover, some legislation, like *The European Union General Data Protection Regulation (GDPR)* has mandated human-in-the-loop moderation. Article 22 & recital 71 of GDPR guarantees that a decision should not be based solely on automatic systems, if the decision is challenged, it would be subjected to human review [12], [13], [139].

2.3. Content Moderation Policy Enforcement

Terms of service and community guidelines may also define the policies that the platforms enforce when violations of service or abusive behavior are detected. These policies can take place at various levels, e.g., at the post level vs. account level [191], [225], and also specify various consequences, e.g., showing a warning vs. removing the abusive account. We found two popular moderation policies that platforms use: *hard* and *soft* moderation.

Hard Moderation: Hard moderation is the most severe way a platform enforces its policies, which removes the content or entities from the platform [90], [155], [257], when a violation of community guidelines occur. Then, other users cannot access the content or connect with the abusive account. Platforms provide an option for the affected users to appeal for the content that might be wrongly removed by the platform [47], [51].

Soft Moderation: Nowadays, soft moderation is the platforms' first line of defense to counter content that violates their content guidelines. In soft moderation, platforms do not remove the content. However, they aim to inform the users about potential issues with the content by adding a warning label, substantiating the post with labels in order to inform and educate the users, or limiting the reach of the doubtful content by putting it in quarantine [130], [213], [330]. Soft moderation has recently received a lot of attention, both in academic circles and by social media platforms, because of the broader effects and effectiveness of this approach. For example, since the invasion of Ukraine by Russian forces, Twitter started adding labels to tweets from Russian & Belarusian state-affiliated media websites [74]. Reddit started *quarantining* r/NoNewNormal, a subreddit that is generally antimask, anti-vax, and is against any governmental COVID-19 restrictions across the world [43]. In quarantine, users are shown a warning page, and they have to make a deliberate choice to view the content. Facebook, Instagram, and Twitter have a *strike system* in place to discourage users from posting misleading false content [3], [44]. For example, if a user gets two or three strikes, then they will not be able to access their account for 12 hours. These steps are still considered as soft moderation, as the user is not permanently banned from the platform.

3. Content Moderation in Literature

To answer our first research question, we examined the last five years of research from the top security conferences (1.45%), such as IEEE Security and Privacy, USENIX Security, NDSS, ACM CCS, IEEE Euro S&P, top data mining conferences (4.5%), such as KDD, WSDM, CIKM, and ICDM, top machine learning conferences (4%), such as AAAI AI, IEEE/CVF, top linguistic conference (13.5%), such as ACL Anthology, and also from the top HCI and social science conferences and journals (35.7%), such as CHI, CSCW, ICWSM, The Web Conference, New Media & Society, Political Behaviour, etc. While this distribution shows that content moderation is an interdisciplinary topic, our paper can spark further research from the security community. We list the distribution of papers in Table 4 in Section B. In our study, we focus on publications investigating topics related to content moderation practices for countering online hate speech, online harassment, trolling, cyberbully or in general online toxicity as well as misinformation, disinformation and fake news. We used the snowballing approach [140] for finding relevant papers. We started our search by a set of keywords (i.e., content moderation, hate speech, misinformation, detection, etc.), and expanded our search, by studying their references and related work sections. The overall set of keywords are: *content moderation* + (social media, tools, effectiveness, engagement, support, removal, removal comprehension, bias), (soft, hard) moderation, and (hate speech, misinformation) (detection, identification, system, automatic, methods, tools, NLP).

To create categories for papers about detection of hate speech and misinformation as well as the moderation studies, three authors classified the papers using the open

coding process [135]. The three authors, independently studied the papers to determine themes, sub-themes, and the approaches and methods that are employed to address a specific task. To find the agreement score, we gave a value of 1, if three authors had a perfect agreement on the categories, otherwise, we divided the number of matched categories by the number of possible categories. Using this methodology, we found substantial agreements of 75% and 70% for groupings of detection methods and moderation studies, respectively. In Section 3.1, we list the identified categories for the publications that are about the detection techniques focusing on *Hate Speech* and *Misinformation*, and in Section 3.2, we list the identified categories for publications regarding moderation policies. If a study fits in multiple categories e.g., using content based features and DNNs, then we label that study as hybrid approach.

3.1. Detection Techniques

Some recent surveys [122], [148], [161], [265], [327], [339] have reviewed detection techniques with respect to either hate speech or misinformation. However, these works focused on specific methods (such as NLP, graph-based, etc.). Our work, in contrast, studies methods detecting both types of abuse on social media and, more importantly, studies these methods considering the entire moderation ecosystem. We systemize the detection techniques based on their methods, grouping them by similarities and differences between the underlying methods (e.g., propagation-based techniques, lexicon-based techniques, etc.) under a common taxonomy. We identified five and seven broad categories for hate speech detection and misinformation detection, respectively, which is shown in Table 6 in Section C. We found that both hate speech detection and misinformation detection systems equally use features derived from the textual part of the social media posts, such as TF-IDF, Part of Speech (POS) tags, etc. [101], [104]. Lexicon-based methods are more common in hate speech detection [72], [307]. Propagation structure, crowd intelligence-based methods, and knowledge-based methods are popular in misinformation detection due to their spread on social media platforms [241], [262]. We found that both hate speech and misinformation detection systems have leveraged the advances of Deep Neural Networks (DNNS) advances, eliminating the need for feature engineering and domain expertise [217], [318]. Misinformation and hate speech have evolved across modalities, including images and videos. We found methods combine these different modalities to identify misinformation (e.g., fauxtography) and hate speech (e.g., hateful memes) [239], [315].

Hate Speech Detection Previous works have extensively used Machine Learning (ML) techniques to detect hate speech, offensive language and toxicity online. Some works [126], [310] have proposed detection methods for specific types of hate speech such as misogyny, sexism, Islamophobia, etc. Other works [101], [237] have proposed methods for detecting offensive language. In this work, we refer to all of them as hate speech detection. We identified four categories of detection studies, based on the features and machine learning algorithms that they employ: **Traditional Machine Learning:** Many works have proposed using traditional ML algorithms such as Support Vector Machines [101], [126], [258], Logistic Regression [255], etc. We characterized these studies based on the proposed features into *content* and *lexicon* based. **Content Based:** includes works that have obtained features solely from the text to detect hate speech, toxicity and offensive language, i.e., TF-IDF, n-grams, POS tags, BoWV etc. [101], [102], [255], [258]. **Lexicon Based:** Scholars have extensively used syntactic and semantic orientations of the existing lexicons for detecting hate speech because keyword-based approaches show high false positive rates, mostly due to ignoring the context [72], [87], [115], [126], [205], [263], [297], [307], [310], [321]. **Deep Neural Based:** With the advancements in the Deep Neural Networks, a majority of existing literature has used neural networks and deep neural networks to detect hate speech [59], [65], [95], [110], [127], [160], [237], [328], [336], [337], [341] and toxic comments [114], [132], [292], [305]. Note that there are methods that use DNN based architectures together with the content, and lexical based features, which we list separately in the subsequent category of *hybrid approaches*. The methods categorized as Deep Neural based use CNN, LSTM, Transformers to detect hate speech. CNNs have been used in a variety of works for hate speech detection tasks [175], [177], [233], [247], [253], [317]. Sequential models such as LSTM have been effective in detecting hate speech on social media [58], [71], [192], [217], [229], [242], [245], [308]. Following the success of transformer networks trained on large amount of data (or Large Language Models), pre-trained models such as BERT [106], GPT [82] and RoBERTa [197] have been used successfully in hate speech detection [61], [88], [220], [221], [256], [322].

Hybrid Approaches: Previous works have effectively combined features derived from the content, lexicon and deep neural network, and several works have proposed hybrid approaches, using an ensemble of classifiers using different sets of features [63], [70], [94], [103], [124], [128], [146], [206], [208], [235], [259], [340]. Grondahl et al. [146] evaluated four state-of-the-art hate speech detection models trained on different datasets and found that the models only work well within their respective trained datasets, failing to generalizing across datasets.

Multi-modal Based: Hate speech can span across multiple modalities such as images, videos, memes, etc. Hence it is imperative to detect this type of content [137]. Works have used text and images together to detect hate speech [285], [325] text and socio-cultural information [311] and memes [137], [179], [194], [239], [309].

Misinformation Detection Scholarships have also extensively used ML techniques to detect misinformation/ fake news. We identified four categories, based on the features and the machine learning algorithms that they use: **Traditional Machine Learning:** Many works have proposed using traditional ML algorithms such as Random Forrest [126], [287], [338], Logistic Regression [104], etc. We characterized these studies based on the proposed features into *content based*, *propagation based*, *hybrid* and *crowd intelligence*. **Content Based:** features such as number of content words (nouns, verbs, adjective), and the presence and frequency of specific POS patterns, TD-IDF, have been employed, to detect misinformation, in [104], [119], [150], [154], [156], [238], [243], [287], [313],

[314], [338]. **Propagation Based:** Different to methods extracting features from the textual content, there are multiple works using the propagation structure, or spreading networks of content on social media to detect misinformation [262], [318]. These works leverage the peculiar traits in underlying network structure of the dissemination of misinformation [281], classifying news propagation paths for early detection [195], and identifying higher-order mutual attention paths in the propagation structures [216]. Scholarships have also used user based features such as number of friends, followers etc, to identify misinformation/ fake news [282], [284], [287]. **Hybrid Approaches:** Combining the features derived from both the content and propagation structure of content on social networks, several works have proposed hybrid approaches to detect misinformation [153], [170], [176], [254], [283]. **Crowd Intelligence:** Wisdom of the crowd can be effectively used to detect misinformation. Multiple works have leveraged implicit crowd intelligence as the content spreads on a social media for early detection of misinformation [171], [181], [202], [241], [280], [301], [312].

Deep Neural Networks: Recent advent of Deep Learning algorithms have allowed researchers to build detection systems without feature engineering. Multiple works have used advanced deep learning techniques such as self-attention [64], adversarial learning [246], graph neural networks [76], [329], recursive neural networks [202], and other deep learning architectures [174], [196], [198], [219], [254], [280], [315], [318] to detect misinformation. **Knowledge Based:** Another line of research uses Knowledge Graph built through relational knowledge extracted from the Open Web and evaluate the truthfulness of contents by verifying them against the "gold" Knowledge Base [100], [112], [157], [185], [231]. Scholarships have used a combination of Knowledge Base and deep neural networks to find misinformation [100], [211], [291], [324]. **Multi-modal Based:** The scope of detecting misinformation expands beyond just detecting text based misinformation but to identifying multi-modal misinformation. Researchers have proposed multiple methods to tackle this problem leveraging the relationship between different modalities of content on social media [288], learning shared representations of text and images [178], learning the shared representations of audio and visual information [274], learning transferable multi-modal feature representations [315], and fusion based methods [240], [290], [320].

Research Gaps. Bozarth and Budak [80] evaluated five representative classifier architectures for misinformation detection, and found that the performance of detection systems vary across datasets, hence prompting a need for building comprehensive evaluation systems. Similarly, the evaluation of hate speech detection methods by [146] suggests that traditional machine learning based methods can be as good as Deep Neural Networks based methods, and the focus of researchers should be more on designing richer, robust datasets. The sizes of existing datasets for hate speech detection range from 6K comments [103] to 100K tweets [125]. However, Madukew et al. [204] identified multiple limitations with the existing datasets for hate speech detection: i) conflating class labels, ii) varying definitions of hate speech across manual annotations, and iii) class imbalance issues, thus highlighting the need for benchmark datasets. Fairness remains a key challenge when building these detection tools, as the underlying biases in the groundtruth dataset or lack of various dialects can reflect in the classification results [79], [264]. Scholarships should focus on building standard benchmarks for evaluating hate speech detection systems across different domains, different languages, and across different nuances of hate speech such as implicit hate speech [67], [218], [327]. Misinformation detection methods are topic specific, e.g., they detect misinformation about presidential election or COVID-19, etc. However, misinformation about other topics are under-worked and under-studied. For example, Singhal et al. [287] found that misinformation regarding security and privacy threats are also prevalent on social media. Despite the promise of zero shot, cross-lingual language models, such as multilingual BERT [106], they are limited for cross-lingual hate speech detection [227]. Future works should focus on building efficient and scalable cross-lingual hate speech detection methods. Scholarships should focus on detecting domain independent multi-modal misinformation [57], and generalizable multi-modal hate speech capturing the complexities of hate speech embedded in memes [163].

3.2. Moderation Policies

We identified six broad categories from our analysis of papers that are shown in Table 5 and Table 7 in the Section C: **Consumption of (fake/ misinformation) news:** Some research studies have tried to understand how the users consume the news online, and what methods are employed by social media to verify the information shared on these platforms. Geeng et al. [130] focused on the effect of warning labels that were added on Twitter, Instagram, and Facebook on posts related to COVID-19 misinformation. They found that most of the survey takers had a positive attitude however, a majority of participants discovered or corrected misinformation by using other means, most commonly web searches. Zhang et al. [335] found that most participants determined the credibility of news regarding COVID-19 using other heuristics such as web searchers. This research corroborates with other research, where authors found that people use multiple heuristics on and off social media to determine the credibility of information [75], [121], [131], [151], [162], [199], [200], [279], [302], [306]. **Research Gaps:** While these research studies give direction on how users investigate fake news and employ warnings, there are some research gaps that the research community should investigate. The study demographics concentrate in the US and had a large percentage of young, tech-savvy participants (18–25 years old) as shown in Table 5. The study also uses textual data, such as articles and some images. Studies with more diverse participants, and also those which focus on specific demographics, e.g., certain age, gender, ethnicity can help our understanding of the acceptance of different content moderation policies in different communities. Also, most of these works are simulation-based, where, for example, participants interacted with simulated web pages that show fake news. Users might show a different behavior, or use different approaches for investigating statements given in images, memes, and videos, therefore more observatory studies are needed to understand the problem.

Engagement of users: Some studies investigated how users interacted with moderated postings and the consequences on engagement when certain communities or influencers are deplatformed and moved to stand-alone or fringe websites. Zannettou [330] performed a data driven study on the soft moderation interventions employed by Twitter. He found that tweets that have warning labels tend to have a higher user engagement. This was also corroborated by researchers in [232]. However, [182], [193] found a contrasting results, where user engagement on content with warning labels on TikTok and YouTube was found to be less. Mena [213] conducted a user study to understand the effect of warning labels on the likelihood of sharing fake news on Facebook. He found that flagging fake news has a significant effect on users' sharing intentions; that is, users are less willing to share content with the labels. This was corroborated in [116], [168], [234], [236], [272], [326]. Chandrasekhran et al. [90] found that quarantine made it more difficult to recruit new members on r/TheRedPill and r/The_Donald, however they find that the existing members hateful rhetoric remained the same. Similarly, Shen and Rose [278] found that Reddit's quarantine was effective in decreasing the posting levels, however the toxicity of users remained the same. Trujillo et al. [300] found that quarantine did in fact reduced the activity of problematic users, however it also caused an increase in toxicity and led users to share more polarized and less factual news. A similar result is seen in the data driven study done by works of [62], [91], [155], [244], [299] which studied the effectiveness of deplatforming. Works such as [166], [257] found that deplatforming significantly decreased posting level, user engagement and toxic rhetoric of the users. **Research Gaps:** Most user studies have a skewed demographic (shown in Table 5), especially in terms of self-reported political views, as there were more liberals than conservatives, which could have affected their findings. More studies with diverse participants can fill the gap. Also, data-driven studies focusing on specific groups of users with different cultures and backgrounds can help understand the factors that affect user engagement in practice. While works have studied the effect of deplatforming, they only study on a few platform, including Reddit, Gab, and Twitter which can be seen in Table 7. Studies on other platforms, such as Facebook and Instagram, can provide more insights, as these platforms are more popular among different populations. Moreover, as content moderation policies on these platforms are constantly evolving, the impact of such changes on user engagement can be studied. Future scholarships could propose and study interventions that can be placed during sharing process, e.g., disabling sharing or displaying a splash warning screen on the shared content, similar to *quarantine*.

Effectiveness: Works have examined the effectiveness of both soft and hard moderation techniques on social media platforms. Some works investigated whether moderation can lead to users moving to less moderated platforms.

Soft Moderation Interventions: A 2018 Gallup survey found that more than 60% of U.S. adults were less prone to sharing stories from sites that were clearly labeled as unreliable [23]. Saltz et al. [260] found that participants had a different opinion regarding Facebook COVID-19 warning labels, some perceiving them necessary step to inform users whereas others saw them as politically biased and an act of censorship. Many studies [168], [173], [183], [213] found that interstitial covers, labels and flagging decrease the perceived accuracy of COVID-19 misinformation and fake news on Twitter [275] and Facebook [97], [213]. Previous research has also found that correcting or debunking fake news can significantly decrease users' gullibility to the story [89], [182], [199], [228], [232], [234], [272], [326]. Seo et al. [271] investigated users perceptions when they were exposed to fact checking warning labels and machine learning generated warning labels. They found that users tend to trust fact checking warning labels more than machine learning generated warning labels. However, previous works also demonstrated some fortuitous consequences from the use of warning labels. Pennycook et al. [236] found an *implied truth* effect, where the posts that included misinformation but were not accompanied by a warning label were considered credible by the users. Studies found that there can be an unintended *backfire effect*, where participants strengthen their support for the false political news that has a warning label or they distrust the source that fact checked it [129], [144]. A few studies [278], [299] found that Reddit's quarantine was ineffective and may also increase the polarization in political spaces.

Hard Moderation Interventions: Chandrasekharan et al. [92] studied the Reddit comments that were removed by moderators to find macro, meso, and micro norms enforced to remove problematic content such as hateful content. Chandrasekhar et al. [91] found that Reddit's ban on r/fatpeoplehate and r/CoonTown was effective, where users drastically decreased their hate speech usage. A similar result was seen in [257]. Schoenebeck et al. [266] showed that users prefer that the platforms remove harassing content. Thomas et al. [298] found that content creators felt platform policies and community guidelines were at least somewhat effective at keeping them safe from hate and harassment. In [223], the author found that when a user was blocked by a user who had a large number of followers, that user significantly reduced their use of a racist slur. However, Jhaver et al. [167] found that users who use blocklist on Twitter were not being adequately protected from harassment. Targets of online harassment expressed frustration with the lack of available support tools and the ineffectiveness of current hard moderation interventions of social media [77], [109], [164], [224], [261], [323]. Facebook, Twitter, Instagram, YouTube, and other platforms have all banned controversial influencers for spreading misinformation, conducting harassment, or violating other platform policies [16], [17], [20], [27]. With these social media banning or fact checking posts, many right-wing individuals, citing censorship, are flocking to communities with fewer restrictions such as Parler, Gab, etc. [30], [33]–[35]. Scholars have extensively studied this migration and how it affects content moderation, and whether it increases or decreases hate speech, etc. Jhaver et al. [166] studied the effectiveness of permanent bans on Twitter of three influencers. They found that banning significantly reduced the number of conversations about all three individuals on Twitter and the toxicity levels of supporters declined. This finding has been corroborated by studies carried out in [155],

[300]. Some scholarships have examined the effects on users rhetoric and whether their followers discuss the influencers who were deplatformed by social media [62], [90], [244], [251]. They found a common result, that deplatforming significantly decreased the reach of the deplatformed users, however the hateful and toxic rhetoric increased. Kumarswamy [188] studied the changes to Parler moderation strategies after it was taken offline by Apple, Google and Amazon Web Services. He found that the overall toxicity of the users decreased. **Research Gaps:** Scholarships should investigate if labeling every news article can decrease the *implied truth* effect. Further, the impact of changes in content moderation policies can be studied on different platforms. One possible avenue for future scholars, could be to understand if there is an *echo chamber* effect on users, when they are given the option to toggle off content with soft moderated label.

Support: Works have studied if users tend to support or disapprove the interventions (i.e., moderation) taken by social media. Through a user study, one research found that 45% of Americans think that social media companies should have a role in moderating user-created content, however that same research also found that 41% of Americans think that social media sites suppress free speech [25]. Geeng et al. [130] found that users had a positive outlook on moderation interventions. Gonçalves et al. [138] found that algorithmic moderation is perceived more favorable then human moderation, while in contrast Lyons et al. [201] found that human moderation was perceived more favorable than algorithmic moderation. Riedl et al. [248] found that users grouped by their age, their education level, and their opposition to censorship, supported social media content moderation intervention. Xiao et al. [323] found that users want social media platforms to improve design and moderate content more proactively. Works such as [78], [111], [164], [199], [260], [267] have found opposition to content moderation interventions, finding that users tend to echo that they are biased and are arbitrarily applied. Participants in [225] echoed that there should be governmental regulation as they felt that it infringes their First Amendment right to free speech. **Research Gaps:** Most user studies focused on participants that were using mainstream social media websites. Future scholarships should have participants from fringe websites, to understand what kind of regulations would be most interesting to them. Most of the user studies did not account for ethnicity as well as cultural differences.

Removal Comprehension: A key aspect of understanding trust & support in social media interventions is to understand if end-users can comprehend why their content was removed. Jhaver et al. [164] found that over a third of the participants did not understand why their content was removed and 29% expressed frustration. Haimson et al. [149] found that conservative users echoed the claim that their content was removed because they perceive social media platforms as heavily controlled by liberals, whereas black and transgender participants echoed that their content was removed because they were expressing their marginalized identities. Schoenebeck et al. [267] found that 41% of the youth participants do not trust social media platforms, this can directly point to youths not comprehending why their content was removed. Works such as [111], [138], [172], [225], [304] found that users complained about social media companies not disclosing the specifics as to why their content was removed. Scholars have also reflected upon the importance of transparency in content moderation systems [117], [167], [172], [270], [296]. **Research Gaps:** Further research should investigate novel and effective designs for a redressal system, where users whose content is removed are given specific details as to why their content was taken down.

Fairness and Bias: Social media platforms play a decisive role in promoting or constraining civil liberties [105]. How platforms make these decisions has important consequences for the communication rights of citizens and the shaping of our public discourse [134]. Shen and Rose [277] studied how the discourse on content moderation is polarized by users' ideological viewpoints. They found that right-leaning users invoked censorship while left-leaning users highlighted inconsistencies in how content policies are applied. Works have studied users' opinions about bias and fairness of content moderation on various social media platforms. Lyons et al. [201] found that users perceive human moderation as more fair and less biased than algorithmic moderation. However, Gonçalves et al. [138] found that algorithmic moderation is perceived to be more transparent and less biased than human moderation. Conservatives have often described the moderation decisions by Twitter, Facebook, etc., as biased and they claim that these companies censor their point of view [25], [162], [164], [190], [225], [250], [260], [266], [296]. Jhaver et al. [167] found that users who are on the Twitter blocklist feel they are blocked unnecessarily and unfairly. Roberts et al. [250] found that moderation is sometimes unfair for people in the marginalized communities. Scholarships have also looked into the inconsistencies and unfairness of moderation decisions and the harm of moderation on marginalized communities [78], [111], [134], [138], [149], [172], [224], [267], [270], [304]. **Research Gaps:** Scholarship should further investigate approaches for improving the precision and recall, as well as the algorithmic fairness of abuse detection systems. In addition, it is necessary to provide solutions for increasing the transparency of such algorithms, by for example, providing *accuracy labels* on the warnings, or adding sources to factchecked claims. Moreover, since most findings on this topic are based on user studies, more data driven studies are required to investigate the validity of these findings. A more thorough discussion on increasing fairness is presented in Section 5.

4. Content Moderation in Practice

To answer our second research question, we studied how various social media platforms employ content moderation, what content do they moderate, and how the content moderation policies are defined, implemented, and enforced.

Choice of Platforms: We focused on the social media platforms that were investigated by prior research studies [93], [131], [275] and from the recent events (i.e., the January 6 insurrection) [24], [30], [188]. We chose fourteen diverse and popular platforms. These platforms include both mainstream (Facebook, YouTube, Instagram, TikTok, Snapchat, Twitter, Reddit, Twitch, and Tumblr)

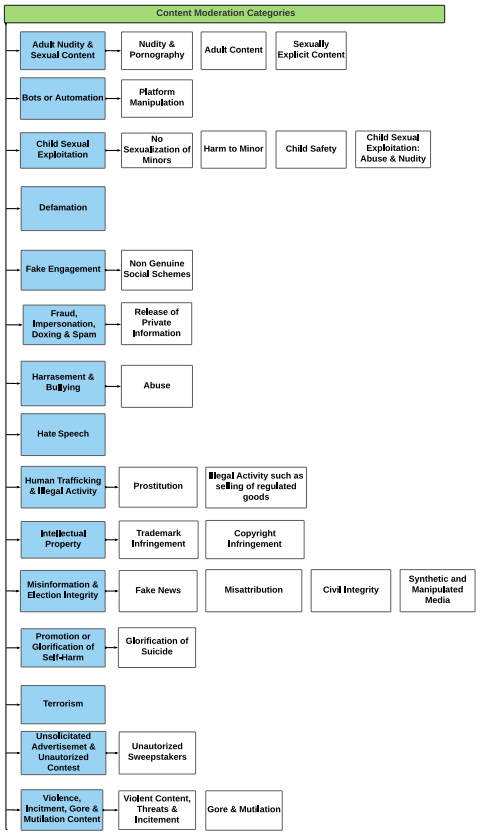

Figure 2: Content moderation categories based on analysis of guidelines

and fringe (4chan, Rumble, MeWe, Gab, and Parler) social media platforms, with different priorities, goals, and audiences.

4.1. Content Categories

To better understand the prevalent content categories that are used by social media companies to regulate and moderate content, we studied each of the platforms' *Community Guidelines/Standards (CG/CS)* that were published on their *US website* during June–August 2021. We investigated and compared the community guidelines of social media platforms for five countries, i.e., India, Germany, Australia, Brazil, and South Africa. Interestingly, we found no change in the guidelines, hence our analysis is not specific to the US. Table 3 in Section A shows the results of this comparison. We also studied the CG of every platform in September 2022 to account for any substantial changes. We found a substantial change in the CG of Parler.[1] Since each platform has a different nomenclature to how they describe some of the categories, we applied the open coding process [135] to categorize them. One of the authors manually looked at all the social media platforms CG/CS to create categories until no new categories emerged. Then applying an iterative process new categories were added, or existing ones were reorganized.

1. Example of changes in Parler CG: https://tinyurl.com/yda6pfmj

To create these categories, we followed certain guidelines: (1) Read through the CG/CS, and identify themes and subthemes; and (2) While creating the categories, identify the meaning. Some of the categories were added based on majority score of where the category was placed by the social media platforms in their community guidelines. For example, *Doxing* can also be part of *Harrasement & Bullying*, however, we found that a majority (12) platforms were placing it in *Fraud*, hence we placed it in that category. Figure 2 shows the content moderation categories. In total, we have fifteen main classes and twenty five subcategories in total. In the following, we define and describe each of the categories. The definitions of these categories were created based on differences in how various platforms were defining the categories, for example, Parler defines *terrorism* as those groups officially recognized as such by the United States Government, whereas Twitter defines it's violent organization policy in a more detailed matter [6], [37].[2]

Adult Nudity & Sexual Content: Any consensually produced and distributed media that is pornographic or intended to cause sexual arousal, full or partial nudity, and simulated sexual acts. Exceptions include content related to artistic, medical, health, breastfeeding or education.

Bots or Automation: Post content automatically, systematically, or programmatically, overutilize the service via automated tooling.

Child Sexual Exploitation: Any type of abuse or sexual exploitation, i.e., nudity toward a child, any form of images, videos, text, or links that promote child sexual exploitation, sending sexually explicit media, trying to engage a child in a sexually explicit conversation.

Defamation: Attacking in the form of oral or written communication of a false statement about another that unjustly harms their reputation.

Fake Engagement: Artificially increasing the number of views, likes, comments, or other metrics, selling or purchasing engagements, using or promoting third-party services or apps that claim to add engagement, trading, or coordinating to exchange engagement.

Fraud, Impersonation, Doxing & Spam: Impersonation of individuals, groups, or organizations with intent or effect of misleading, confusing, or deceiving others, any fraudulent schemes, such as fake lotteries, phishing links, spamming users and comments, deceptive means to generate revenue or traffic, publishing of private information about an individual for malicious intent.

Harassment & Bullying: Engage in the targeted harassment of someone, or incite other people to do so, sending threatening messages, and establishing malicious unsolicited contacts and threats.

Hate Speech: Direct attacks against people based on their protected characteristics, i.e., race, ethnicity, origin nationality, disability, religious belief, race, sexual orientation, gender, gender identity, and serious illness.

Human Trafficking & Illegal Activities: Facilitate sex trafficking, other forms of human trafficking, or illegal prostitution, unlawful purpose or in furtherance of illegal activities including selling, buying, or facilitating

2. Time-specific snapshot of platforms policies can be found at: https://tinyurl.com/4wu9vzkc

transactions in illegal goods or services, as well as certain types of regulated goods or services.

Intellectual Property: Content that has the unauthorized use of a copyrighted image as a profile photo, allegations concerning the unauthorized use of a copyrighted video or image uploaded through our media hosting services.

Misinformation & Election Integrity: Manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process, as well as sharing synthetic or manipulated media that is likely to cause harm.

Promotion or Glorification of Self-Harm: Promotes suicide, self-harm, or is intended to shock or disgust users, content that urges or encourages others to: cut or injure themselves; embrace anorexia, bulimia, or other eating disorders; or commit suicide.

Terrorism: Promotes, encourages, or incites acts of terrorism, content that supports or celebrates terrorist organizations, their leaders, or associated violent activities.

Unsolicited Advertisements & Unauthorized Contests: Unsolicited, unrelated advertisements, unauthorized contests, sweepstakes, and giveaways.

Violence, Incitement, Gore & Mutilation Content: Any threats of violence towards an individual or a group of people, content inciting people to commit violence, any media that depicts excessively graphic or gruesome content related to death, violence or severe physical harm, or violent content that is shared for sadistic purposes, severely injured or mutilated animals. There are however exceptions such as content may be made for documentary or educational content, religious sacrifice, food processing, and hunting.

Discussion: We found that interestingly some topics, such as *misinformation*, have different definitions on each platform, and certain platforms give no definition of it. Our analysis also shows that in some categories, such as Adult Nudity & Sexual Content & Fake Engagement, there is no consensus among platforms on what content to moderate, which makes moderation arbitrary. Even platforms that are similar in nature (i.e., attract like-minded people), such as Parler and MeWe, do not prohibit hate speech content the same way. Therefore, there is a need for not only the computer scientists but also scholars from other disciplines such as psychology, social sciences, and law to come together and come to a consensus at least at the definition level.

4.2. Content Moderation Guidelines and Polices

Table 1 shows the content moderation policies mentioned in platforms' CGs (soft or hard) as well as approaches used by them for enforcing these policies (human or ML). We also checked if any of these platforms claim to be doing factchecking. Table 1 shows *No* if some policy is not defined, *Yes* if it is defined, *Partially* if it is only defined for some moderation categories mentioned in the previous section. Below we describe the findings based on Table 1:

Factchecking: Parler, Gab, MeWe, Twitch, Rumble & 4chan do not perform any factchecking on the content, also these social media platforms were created with a promise of less content moderation, as they call themselves *champions of free speech* [18], [24], [30], [40]. Snapchat performs factchecking only on advertisements, including political advertising [26]. Meanwhile, platforms like Tumblr, TikTok, Reddit, YouTube, Facebook, Instagram, & Twitter factcheck claims that are posted on its platform. However, they do not specify if they fact check all the news. Based on the previous works though it seems these platforms only factcheck some topics, such as COVID-19 and elections.

Hard Moderation: Parler enforces some hard moderation but is kept to a minimum, as they mention "We [Parler] prefer that removing users or user-provided content be kept to the absolute minimum" [5]. After Parler was banned from Apple App Store & Amazon following the January 6th Capitol roit [41], Parler changed its community guidelines and is now not allowing *hate speech* on iOS, but allowing on Android or the web [42]. However, Parler does not remove so-called *fighting words*, which are not protected as an exercise of the right to free speech [37]. On Gab, users who are posting content from the ten categories described in Table 2, would be subjected to a ban or their content will be taken down. MeWe also has a very similar approach to that of Gab, however, MeWe does not allow hate speech on its platform. 4chan, Tumblr, Snapchat, TikTok, Reddit, YouTube, Facebook, Instagram, Twitter, Rumble & Twitch have more stringent hard moderation rules. Users can expect to have their content taken down, shadow banned, temporarily banned, or permanently banned from the platform.

Discussion: Our analysis shows that there is no universal method for hard moderation across all social media platforms, and each platform follows a different approach, which can be due to the limitations that each of these methods has. Therefore, even more, rigid analysis is needed to measure and compare the performance of these methods and also propose new models that can improve their performance.

ML or Human-based moderation: Parler, Tumblr, Snapchat, TikTok, Reddit, YouTube, Facebook, Instagram, Twitter, & Twitch have both humans and AI as moderators. We were not able to find information about what kind of moderators are employed by Gab. 4chan has human moderators and *janitors*, who are volunteers that can remove threads or replies on the imageboard they are assigned to, they can submit a request for a ban or warning of a user to the moderators (they themselves cannot ban the users). MeWe and Rumble also have human moderators [24], [50].

Soft Moderation: We found that Parler, Gab, MeWe, 4chan, Tumblr, Rumble, Snapchat & Twitch do not perform soft moderation on the content that is posted on their respective platforms. TikTok has started putting warning labels on the content where the facts are inconclusive or content is not able to be confirmed, especially during unfolding events. However, they have not specified what is considered as "unfolding" events. When a viewer tries to share a video that is flagged, they would be shown that the content is flagged for unverified content, and they can either cancel or share the post. Reddit's soft moderation technique is *quarantine*. The main purpose of quarantine is to prevent users from accidentally viewing the content.

TABLE 1: Content moderation policies defined and enforced by different social media platforms

Moderation Strategy	Platform													
	Parler	Gab	MeWe	4chan	Rumble	Tumblr	Snapchat	TikTok	Reddit	YouTube	Facebook	Instagram	Twitter	Twitch
Factchecking	No	No	No	No	No	Yes	Partially	Yes	Yes	Yes	Yes	Yes	Yes	No
Soft Moderation	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No
Hard Moderation	Partially	Partially	Partially	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Human Moderators	Yes	X	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ML Moderation	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

TABLE 2: The content categories that are moderated on different social media platforms

Abusive Content Categories	Platform													
	Parler	Gab	MeWe	4chan	Rumble	Tumblr	Snapchat	TikTok	Reddit	YouTube	Facebook	Instagram	Twitter	Twitch
Adult Nudity & Sexual Content	Partially	Partially	X	Partially	X	Partially	Partially	X	Partially	Partially	Partially	Partially	Partially	Partially
Bots or Automation	X	X	X	X	N/A	X	X	X	X	X	X	X	Allowed*	X
Child Sexual Exploitation	X	X	X	X	X	X	X	X	X	X	Partially	Partially	Partially	X
Defamation	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Fake Engagement	N/A	X	N/A	N/A	N/A	X	N/A	X	X	X	X	X	X	X
Fraud, Impersonation, Doxing & Spam	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Harassment & Bullying	N/A	N/A	X	N/A	N/A	X	X	X	X	X	X	X	X	X
Hate Speech	Allowed*	Allowed	X	Allowed	X	X	X	X	X	X	X	X	X	X
Human Trafficking & Illegal Activities	X	X	X	X	N/A	X	X	X	X	X	X	X	X	X
Intellectual Property	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Misinformation & Election Integrity	Allowed	Allowed	Allowed	Allowed	Allowed*	X	X	X	X	X	X	X	X	X
Promotion or Glorification of Self-Harm	N/A	N/A	N/A	N/A	X	X	X	X	X	X	X	X	X	X
Terrorism	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Unsolicited Advertisements & Unauthorized Contests	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Violence, Incitement, Gore & Mutilation Content	Partially	X	X	X	X	X	X	X	X	Partially	Partially	Partially	Partially	X

Reddit also gives an explanation of why the community is in quarantine. However, a user can still view the content by clicking on the continue button. Twitter also has the same mechanism both for posts and also for accounts, a user can however still view the content [4]. YouTube, Facebook, Instagram & Twitter also have a very similar soft moderation policy.

Gaps in Practice: Future scholarships should investigate ways where mainstream social media can also safeguard the First Amendment rights. Also, in practice, soft moderation has only been used for labeling misinformation. However, soft moderation might also be useful in labeling other types of abuse, e.g., malicious content, phishing urls, bot generated content, hate speech, etc.

4.3. What Content Category Gets Moderation?

We extracted the topics that social media platforms claim to be moderating in their community guidelines. Table 2 shows all different types of abusive content that each of the platforms have mentioned in their CGs. Two coders then labelled the categories for each platform. Coders followed the given nomenclature to label the categories: *X* if the content is not allowed, *Partially* if it is only allowed for some sub-topics, e.g., only for specific types of nudity and not for porn, *N/A* if the guidelines do not mention that topic, *Allowed* if the topic is allowed in the platform, and *Allowed** if the topic is allowed, however, it has restrictions. Using this methodology, we calculated the inter-coder reliability score. We found a substantial agreement of 0.76.

Misinformation & Election Integrity: We found that any content that is posting misinformation about COVID-19 or elections are not allowed on Tumblr, TikTok, Reddit, YouTube, Facebook, Instagram, Twitter, & Twitch. However, Parler, Gab, MeWe, Rumble, & 4chan allows misinformation to be present on its platform and they do not moderate them.

Hate Speech: Hate speech is not allowed on MeWe, Tumblr, Snapchat, TikTok, Reddit, YouTube, Facebook, Instagram, Rumble Twitter, & Twitch. Parler does not allow any content that is hateful on iOS devices, but it still allows it on other devices [42]. Gab and 4chan do not remove hate speech on their platform [45], [332].

Adult Nudity & Sexual Content: We found that adult nudity & sexual content of any kind are not allowed on MeWe, and TikTok. In Parler, users are allowed to post images, videos, depictions, or descriptions of adult nudity or sex as long as they are designated as *sensitive* (NSFW) [37]. However, if a user posts any content that is containing nudity or sexual content, and the user has not designated it as NSFW, then that content is removed. Exceptions are made for spiritual artwork or posts by a verified art gallery. In 4chan, users are allowed to post Anthropomorphic pornography, Grotesque images, and Loli/shota pornography only in /b/ board [1]. Platforms such as Gab, Tumblr, Snapchat, Reddit, YouTube, Facebook, Instagram, Twitter, & Twitch allow nudity content e.g. as a form of protest or for educational/medical reasons, with Twitch allowing individuals actively breastfeeding a child on stream. Twitter, Instagram, & Facebook would apply a label to content involving breastfeeding, and images/videos shared in medical or health context.

Bullying & Harassment: We found that any content that is bullying & harassing a user, or a group is not allowed on MeWe, Tumblr, Snapchat, TikTok, Reddit, YouTube, Facebook, Instagram, Twitter, & Twitch. Parler, Rumble, Gab, & 4chan do not have a policy outlining this category.

Promotion or Glorification of Self-Harm: We found that promotion or glorification of self harm is not allowed on Tumblr, Snapchat, TikTok, Reddit, YouTube, Facebook, Instagram, Rumble, Twitter, & Twitch. Parler, Gab, MeWe, & 4chan do not have a policy outlining this category.

Fake Engagement: We found that fake engagement is not allowed on Gab, Tumblr, TikTok, Reddit, YouTube, Facebook, Instagram, Twitter, & Twitch. Parler, MeWe, Rumble, Snapchat, & 4chan do not have a policy on this category.

Defamation: Defamation is not allowed across all the platforms.

Violence, Incitement, Gore & Mutilation Content: We found that some forms of Gore & Mutilation content is allowed on Parler, YouTube, Facebook, Instagram, & Twitter. On Parler, users should designate content as NSFW, if they fail, then the content will be removed. YouTube, Facebook, Instagram, & Twitter have exceptions

for religious sacrifice, food preparation or processing, and hunting.

Child Sexual Exploitation: We found that Facebook, Instagram, & Twitter, allow only educational content for child sexual exploitation i.e., documentaries, news media reportage. These types of content are normally accompanied by a label in Facebook and Instagram. While platforms explicitly do not mention how they identify images or videos depicting child sexual exploitation, previous works have reported that major social media companies use PhotoDNA [143].

Fraud, Impersonation, Doxing & Spam: We found that across all the platforms, any type of fraud, impersonation, doxing & spam are not allowed.

Terrorism: We found that across all the platforms, the promotion of terrorist propaganda, or violent extremism are not allowed. This includes recruiting for a violent organization, or using the insignia or symbol of violent organizations to promote them. Previous work has reported that various social media companies use a *Shared Industry Hashed Database*.

Bots or Automation: All the platforms except Twitter & Rumble prohibit the use of bots or automation to post content or over-utilize the service by sending excessive queries. Twitter allows users to send automated tweets, sending replies and mentions etc., however, they have to authorize an app or service via OAuth [14]. We were not able to find any such policy outlined in Rumble's CGs.

Unsolicited Advertisements & Unauthorized Contests: None of the platforms allow users to post unsolicited advertisements or hold any unauthorized contests. Tumblr has a separate policy outlying the rules to hold contest, sweepstakes, and giveaways [10]. While platforms have regulations for product placements and influencers adverts, it is currently out of the scope of this paper.

Intellectual Property: All the platforms ban any content that infringes the copyright. Any content that infringes the copyright will be removed. Content ID is a state of the art system that is used to detect copyright infringement content [289]. However, content that is satire, parody, and news reporting among others are not in violation of intellectual property, and countries such as the US and some other countries, follow the *fair use* doctrine, whereas the EU has some exceptions [147], [303].

Human Trafficking & Illegal Activities: All the platforms except Rumble do not allow any content about Human Trafficking & Illegal Activities. We were not able to find any such policy outlined in Rumble's community guidelines.

Discussion: We found that content that is legally required to be removed from the social media platforms, such as *Child Sexual Exploitation*, *Violence*, *Intellectual Property*, and *Terrorism*, has a more uniform moderation across all the fourteen platforms. We also found a very consistent pattern in fringe social media platforms with respect to *Misinformation & Election Integrity*, where all these platforms allowed them. We also see that there are spectrum of differences in the definitions of *Adult Nudity & Sexual Content* on Parler and 4chan, both allowing them, but they have various differences to what is allowed, however, mainstream social medias such as Twitter, Facebook have uniformity in the content that may be allowed.

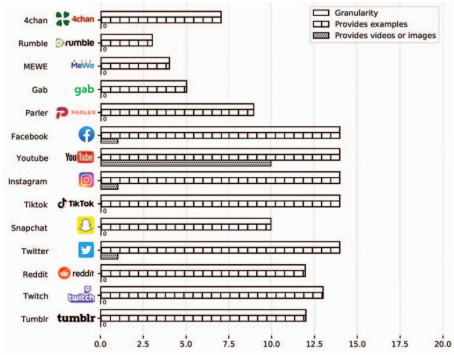

Figure 3: Community Guidelines Comprehensibility

4.4. Community Guidelines Comprehensibility

Comprehensibility of community guidelines means whether the end-user can understand completely what type of content is allowed on the platforms and what would be the repercussions if there is a violation of these terms. The goal of this analysis is to identify places where guidelines are not comprehensible enough. For that, we first check the granularity of community guidelines in terms of covering various content categories, and then we check if they have provided examples, images, and videos for each of the categories. We used the same mechanism to label the categories as described in Section 4.2. Figure 3 shows the results. It shows three broad categories: (1) granularity, (2) provides an example, & (3) provides videos or images. Two authors independently coded all the categories. For inconsistent results, coders discussed how to resolve disagreements. To assess the inter-coder reliability, we performed a Cohen-Kappa test [268]. The Kappa score was 0.75, which shows substantial agreement.

Granularity: Granularity can help users to understand in more detail the content that is allowed and what content is not allowed. To measure granularity, we used the data provided in Table 2 and coded *Yes* and *Partially* as 1 and *No* as 0, and then computed the sum. In our analysis, we found that YouTube, Facebook, TikTok, Instagram, & Twitter were the top five platforms to have very granular CGs. However, we found that MeWe and Rumble were the only two platforms, whose community guidelines were not granular, except for four and three sub-categories, respectively. The average number of categories for which the platforms' community guidelines were granular was ten out of fifteen categories, the minimum was four, and the maximum was fourteen.

Provides Examples: In order to facilitate the rules, social media companies also provide examples to help users understand what type of content will not be permissible and what will be. In our analysis, we found that YouTube, Facebook, TikTok, Instagram, & Twitter were the top five platforms that were providing examples in their community guidelines. Rumble and MeWe were the two platforms with the least number of examples in the subcategories studied (i.e., three & four respectively). The average number of categories for which the platforms were providing examples was ten out of fifteen, with the minimum as four, and the maximum as fourteen. Interestingly, among the fringe social media that we studied, Parler

provided the most number of examples, i.e., nine. For the subcategory, *Violence, Incitement, Gore & Mutilation Content* we found that 4chan, & Snapchat were only providing examples for violence and incitement, however, these platforms failed to provide examples for gore and mutilation content. TikTok provides examples for automation, however, the platform does not provide examples about the use of bots. Twitch provides an ample number of examples for misinformation, however, the community guidelines do not mention posts/videos challenging the integrity of an election.

Provides Videos or Images: Audio and visual context aid users to grasp the content that will not be moderated. We found that only YouTube provides videos for the ten sub-categories. However, we found that for misinformation and election integrity, Facebook, Instagram, & Twitter with YouTube provided visual material for the user. The average number of categories for which the platforms' community guidelines were providing videos or images was one out of fifteen, with the minimum as zero, and the maximum as ten.

Summary: Figure 3 shows that only Youtube community guidelines provide videos or images. We also found that a lot of mainstream social media, such as Tumblr, Twitch, Reddit, etc., do not provide such. Future scholarships can investigate if by providing such examples, users would pay more attention to the guidelines and whether it helps users to understand them. This can have implications regarding users' complaints that they do not understand the reason behind the removal of their content.

5. Discussion

Content Moderation from the Legal Perspective. Freedom of speech is equivocally a basic human right, and prior works have found that users complain that their voices are being suppressed by social media companies [164], [225], [260], [266], [296]. Gillespie [133] argued that platforms are innately political entities that have attempted to maintain an image of neutrality. It is this image of neutrality in the domain of speech that has come under question most in recent years, and the inability of platforms to be truly neutral is core [269]. People expect to have the same protections on social media platforms as they would have in the real world, while prior works have found that users want government oversight [96], [225] and a possible solution to this is to have government actors determine which platforms' content moderation practices could be subject to government oversight, with the platforms following legally defined set of rules. Platforms have to follow laws in the jurisdictions in which they are operating. For example, in the US, social media companies are not considered as a *state* entity under *Marsh v. Alabama*, and hence do not have to guarantee First Amendment protection to the user for protected speech [7]. Social media platforms are also shielded by Section 230 of the Communication Decency Act (CDA) which gives online intermediaries broad immunity from liability for user-generated content posted on their sites [8]. In the landmark case of *Zeran v. America Online, Inc.*, the purpose of §230 is to encourage platforms to act as a *Good Samaritans* and to take an active role in removing offensive content [9]. However, two currently pending cases in the Supreme Court, i.e., Gonzalez v. Google LLC [38] & Twitter, Inc. v. Taamneh [55] could upend §230 and consequentially pose serious risks to Internet speech [136], [184]. In India, pursuant to Article 4(d) of India's Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, social media platform companies must publish a monthly report regarding their handling of complaints from users in India, including actions taken on them [214], [215]. In Germany, The Network Enforcement Act (NetzDG) obliges social media platforms with over 2 million users to remove *clearly illegal* content within 24 hours and all illegal content within 7 days of it being posted or face a maximum fine of 50 million Euros [118]. While the law is one of the most stringent in the world, scholarships, and human rights organizations have criticized it for incentivizing social media platforms to preemptively censor valid and lawful expression [158]. In Australia, users can submit a complaint to the Australian eSafety Commissioner and get a remedy from the commissioner by using the available safety tools and resources [69].

One way to moderate content in the grey areas and still uphold the statutory and legal doctrines is to apply soft moderation tags to posts that are inaccurate or explicit. This would be a welcoming step for users who echo that platforms are deliberately censoring their point of view [149], [225].

Transparency in Moderation. Over the past few years, researchers have reflected on the importance of transparency in content moderation [164], [225], [267], [270], [295], [296]. Legal scholars have argued that platforms should disclose the content moderation policies and procedures and in order to increase users' trust [190], [316]. Prior works have also found that users often do not understand why their content was removed, as they get a very generic response from the social media companies [164], [225], [267]. We concur with the suggestion from Díaz and Hecht [107], that social media platforms should provide transparency reports which will be an essential step in allowing users and watchdog organizations alike to identify issues ranging from lack of language expertise to biased algorithms. The transparency report should report the identities of organizational trusted flaggers and the distinct content policies for which they have special flagging privileges, the types of automated tools deployed to identify and remove the offending content, such as the use of hashing systems and natural language processing systems and whether the moderators are employed by the company itself or are they outsourced to a third party. We also echo that platforms should make public log reports for the moderated content and more social media companies should provide APIs for independent researchers to review the mechanisms and also conduct a fairness analysis on the system. Platforms should also publish the number of orders received from government agencies to remove content or suspend accounts, and whether the platform took action, and if so was it based on actual infringements of the law or was it based on the violations of community guidelines?

One Size does not Fit All. The amorphous nature of social media platforms' content moderation policies gives companies enormous discretion in their enforcement, however, they should exercise this discretion in a way that

truly balances free expression with equity and prioritizes user safety and due process [107]. Following the death of George Floyd, users have tried to make their voices heard both in offline and online spaces. However, due to companies enforcement for speech talking about terrorism or racism, users especially those of color have found that their voices are getting suppressed [21], [29]. Prior researches have also corroborated that, voices of marginalized communities are often suppressed [149], [250], [270]. On the other hand, hate speech policies attempt to take a more measured approach with narrower restrictions, which can create a convoluted order of operations that ends up protecting powerful groups or allowing all but the most explicit attacks based on protected characteristics [107]. Hence, platforms should take steps to redress their moderation systems, so that all users are treated equally. We recommend that social media platforms should take into account the current political and cultural context of the country when they are moderating content. This will help in users from marginalized communities the freedom of expression and at the same time punish the powerful groups that are often involved in incitement of violence or predominately engaging in hateful rhetoric.

Demonetization of videos is one of the broader suites of content moderation governance mechanisms that is available to YouTube [142]. However, creators have echoed that inconsistencies present in the demonetization process lead to beliefs of being algorithmically controlled or censored [86]. Interestingly, it has been found that YouTube was treating established media personalities differently than they were treating users with a few thousand subscribers [28], [113], [187]. Hence, we echo that content moderation policy must be overhauled to account for power flux among different groups such as users with large follower bases, i.e., influencers, and structural inequalities that may curtain opportunity for equal access to platforms. Platforms should acknowledge the fact that the ability to protect the speech of ordinary people when challenging the elite or influential people of the society requires a different matrix than the protections necessary for influential figures who often have multiple avenues for disseminating their message. Platforms should publish policies that govern public figures, heads of state, and other influencers in community guidelines. Social media companies should make sure that both the automated moderation tool and human moderators are able to accurately assess different languages, dialects, slang, and related variations of context. Social media companies should make sure that their AI systems are not potentially discriminating because of the train cases given to them.

Collaborative Human-AI Decision Making. With the growing scale of content, platforms employ human moderators, ML-based moderators, or both to regulate the content. While human moderation comes at a cost of time and scalability, ML-based moderation tools also suffer from the lack of training examples specific to regions or local dialects. It is arguably impossible to make perfect automated moderation systems because their judgments need to account for the context, the complexity of language, and emerging forms of obscenity and harassment, and they exist in adversarial settings where they are vulnerable to exploitation by bad actors [165], [210]. Even though hard coding the criteria helps in scalability and consistency, they still suffer from being insensitive to the individual differences of content, for example, when distinguishing hate speech from newsworthiness [85]. Hence, there is a need for more proactive human interventions to offset the errors done by ML based moderation. However, more studies should be conducted to examine the effectiveness of having interconnected moderation systems, where human moderators proactively provide feedback on ML-based moderation.

Fairness is another challenge of such systems, especially when they mainly rely on user-labeled data. Therefore, it is also crucial to develop fully or partially automated methods and algorithms that minimize the impact of bias on moderation decisions. One more human-based method for increasing fairness of content moderation is allowing users to appeal decisions, where an impartial, independent authority can verify the decisions. An example of this type of impartial jury is *Oversight Board* [39]. The Oversight Board appeals process gives people a way to challenge content decisions on Facebook or Instagram. While this is a good step forward, the oversight board does not evaluate all the submitted cases but selects eligible cases that are difficult, significant, and globally relevant that can inform future policy [36]. Pan et al. [230] work also found strong evidence of support for independent expert juries, similar to The Oversight Board.

Open sourcing vs. restricting access to content. To provide a sense of fairness and accountability, open sourcing public data and public moderation logs are necessary. As we found in Section 3, previous works have largely studied moderation on platforms such as Twitter and Reddit, as they provide mechanisms for researchers to archive data. However, there are platforms from mainstream to fringe that are still restricting researchers from obtaining the vast amount of public data.

Uniformity vs. favorability. With the Russian invasion of Ukraine, big tech companies such as Twitter, Facebook and Instagram, loosened their hate speech, for Ukranians to post about calling for general violence against Russians [54]. While, this was ultimately changed three days later, such was not the case for protesters in Iran, where posts that include *death to the dictator* - a key protest slogan, are being proactively flagged and taken down [48], [49], [52], [53]. One has to ask, if platforms should remain neutral and uniform in implementing content moderation policies? In addition, more studies should be conducted to research the implications of using a threshold to detect hate speech more or less rigorously depending on the offline events.

6. Limitations & Future Work

In this work, we focus on the platforms from the consumer point of view and not from business point of view. We plan to study that in the future to look at the platforms from a business point of view.

7. Conclusion

In this work, we have presented three taxonomies based on an extensive review of the social media community guidelines and previous works. Using the taxonomies

we answer the two research questions. We concluded that the most popular and mainstream social media platforms moderate for all categories studied as well as using both hard moderation and soft moderation categories. Out of these six platforms, only YouTube provides image or video examples. On the other hand, fringe platforms do not moderate for all the studied categories and prefer minimal interventions.

8. Acknowledgment

This material is based upon work supported by the National Science Foundation under Award NSF III Medium 2107296 and NSF RAPID: SaTC 2309318 at the University of Texas at Arlington, and Awards CNS-1942610 and CNS-2114407 at Boston University. Chen Ling was supported by the Meta Research PhD Fellowship. We also would like to thank professor Andrew Sellars for his helpful feedback and fruitful discussion on the legal aspect of content moderation. We would also like to thank our shepherd Dr. Jason (Minhui) Xui and anonymous reviewers for their helpful comments.

References

[1] 4chan Rules. https://www.4chan.org/rules.

[2] Civic integrity policy. https://help.twitter.com/en/rules-and-policies/election-integrity-policy.

[3] COVID-19 misleading information policy. https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.

[4] Notices on Twitter and what they mean. https://help.twitter.com/en/rules-and-policies/notices-on-twitter.

[5] Parler Community Guidelines. https://parler.com/documents/guidelines.pdf.

[6] Violent organizations policy. https://help.twitter.com/en/rules-and-policies/violent-groups.

[7] Marsh v. alabama, 1946.

[8] 47 U.S. Code §230 - Protection for private blocking and screening of offensive material. https://www.law.cornell.edu/uscode/text/47/230, 1996.

[9] Zeran v. america online, inc., 1997.

[10] Contest, Sweepstakes, and Giveaway Guidelines. https://www.tumblr.com/policy/en/contest-guidelines, 2012.

[11] The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. https://www.wired.com/2014/10/content-moderation/, 2014.

[12] General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-22-gdpr/, 2016.

[13] Recital 71 - Profiling. https://gdpr-info.eu/recitals/no-71/, 2016.

[14] Automation rules. https://help.twitter.com/en/rules-and-policies/twitter-automation, 2017.

[15] Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism. https://perma.cc/6QYS-ML76, 2017.

[16] 2018 was the year we (sort of) cleaned up the internet. https://mashable.com/article/deplatforming-alex-jones-2018, 2018.

[17] Deplatforming Works. https://www.vice.com/en/article/bjbp9d/do-social-media-bans-work, 2018.

[18] Gab, the Social Media Site for the Alt-Right, Gets Deplatformed. https://nymag.com/intelligencer/2018/10/gab-the-alt-right-social-media-site-gets-banned.html, 2018.

[19] Inside Facebook's Fast-Growing Content-Moderation Effort. https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/, 2018.

[20] Facebook Bans White Nationalism and White Separatism. https://www.vice.com/en/article/nexpbx/facebook-bans-white-nationalism-and-white-separatism, 2019.

[21] Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech. https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/, 2019.

[22] GIFCT. https://perma.cc/44V5-554U, 2019.

[23] Here's What Happens When News Comes With a Nutrition Label. https://www.wired.com/story/gallup-poll-fake-news-ratings/, 2019.

[24] Inside MeWe, Where Anti-Vaxxers and Conspiracy Theorists Thrive. https://www.rollingstone.com/culture/culture-features/mewe-anti-vaxxers-conspiracy-theorists-822746/, 2019.

[25] Most conservatives believe removing content and comments on social media is suppressing free speech. https://today.yougov.com/topics/technology/articles-reports/2019/04/29/content-moderation-social-media-free-speech-poll, 2019.

[26] Snapchat fact-checks political ads, unlike Facebook, says CEO Evan Spiegel. https://www.cnbc.com/2019/11/18/snapchat-fact-checks-political-ads-unlike-facebook-ceo-evan-spiegel.html, 2019.

[27] The Downfall of Alex Jones Shows How the Internet Can Be Saved. https://www.vanityfair.com/news/2019/04/the-downfall-of-alex-jones-shows-how-the-internet-can-be-saved, 2019.

[28] YouTube's arbitrary standards: Stars keep making money even after breaking the rules. https://www.washingtonpost.com/technology/2019/08/09/youtubes-arbitrary-standards-stars-keep-making-money-even-after-breaking-rules/, 2019.

[29] Black Lives Matter Activists Say They're Being Silenced By Facebook. https://www.buzzfeednews.com/article/craigsilverman/facebook-silencing-black-lives-matter-activists, 2020.

[30] Fact-Checked on Facebook and Twitter, Conservatives Switch Their Apps. https://www.nytimes.com/2020/11/11/technology/parler-rumble-newsmax.html, 2020.

[31] Here's how we're using AI to help detect misinformation. https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/, 2020.

[32] How Facebook uses super-efficient AI models to detect hate speech. https://ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/, 2020.

[33] Parler, mewe, gab gain momentum as conservative social media alternatives in post-trump age. https://www.usatoday.com/story/tech/2020/11/11/parler-mewe-gab-social-media-trump-election-facebook-twitter/6232351002/, 2020.

[34] Right-wing users flock to Parler as social media giants rein in misinformation. https://www.pbs.org/newshour/nation/right-wing-users-flock-to-parler-as-social-media-giants-rein-in-misinformation, 2020.

[35] Why some Americans are trading in mainstream social networks for ones that tout "freedom of speech". https://www.cbsnews.com/news/americans-trade-in-their-mainstream-social-networks-for-ones-that-tout-freedom-of-speech/, 2020.

[36] Appealing Content Decisions on Facebook or Instagram. https://oversightboard.com/appeals-process/, 2021.

[37] Elaboration-on-Guidelines. https://legal.parler.com/documents/Elaboration-on-Guidelines.pdf, 2021.

[38] Gonzalez v. google llc, 2021.

[39] Oversight Board. https://oversightboard.com/, 2021.

[40] Parler, a Platform Favored by Trump Fans, Struggles for Survival. https://www.wsj.com/articles/parler-struggles-survival-amazon-lawsuit-trump-fans-11610414745, 2021.

[41] Parler Ban: Employees At Apple, Google, And Amazon Back Booting Far-Right Platform. https://www.forbes.com/sites/ajdellinger/2021/01/16/parler-ban-employees-at-apple-google-and-amazon-back-booting-far-right-platform/?sh=64d1955c4f5a, 2021.

[42] Parler will be hate speech–free — on iOS only. https://www.niemanlab.org/2021/05/parler-will-be-hate-speech-free-on-ios-only/, 2021.

[43] Reddit Quarantines AntiMask Antivax Subreddit. https://www.vice.com/en/article/akgqmg/reddit-quarantine-nonewnormal, 2021.

[44] Restricting accounts. https://transparency.fb.com/enforcement/taking-action/restricting-accounts/, 2021.

[45] What is Gab? The far-right social media site that Google and Apple banned and that is still gaining thousands of new users after Twitter and Facebook deplatformed Trump. https://bit.ly/2Vr8jeV, 2021.

[46] YouTube Community Guidelines enforcement. https://transparencyreport.google.com/youtube-policy/removals?hl=en, 2021.

[47] About suspended accounts. https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts, 2022.

[48] Facebook's Ukraine-Russia Moderation Rules Prompt Cries of Double Standard. https://theintercept.com/2022/04/13/facebook-ukraine-russia-moderation-double-standard/, 2022.

[49] How Instagram Is Failing Protesters in Iran. https://slate.com/technology/2022/06/instagram-meta-iran-protests-exceptions.html, 2022.

[50] How Rumble, a Toronto-based YouTube alternative, became a refuge for the MAGA crowd (with a US$2-billion valuation). https://www.theglobeandmail.com/business/article-rumble-toronto-video-platform-youtube-alternative-valuation/, 2022.

[51] I don't think Facebook should have taken down my post. https://www.facebook.com/help/2090856331203011, 2022.

[52] Iran's protests pose a challenge for Washington and Silicon Valley. https://www.washingtonpost.com/world/2022/10/12/iran-tech-sanctions-censorship-protest/, 2022.

[53] Meta changes Ukraine content moderation policy to ban death threats against heads of state. https://freespeechproject.georgetown.edu/tracker-entries/meta-changes-ukraine-content-moderation-policy-to-ban-death-threats-against-heads-of-state/, 2022.

[54] Meta withdraws Ukraine war content policy guidance request. https://www.reuters.com/technology/meta-withdraws-ukraine-war-content-policy-guidance-2022-05-11/, 2022.

[55] Twitter, inc. v. taamneh, 2022.

[56] Why India banned TikTok — and what the US can learn from it, as pressure mounts for Biden to follow suit. https://www.businessinsider.com/why-did-india-ban-tiktok-us-states-fcc-government-2023-1, 2023.

[57] Sara Abdali. Multi-modal misinformation detection: Approaches, challenges and opportunities. *arXiv preprint arXiv:2203.13883*, 2022.

[58] Shivang Agarwal and C Ravindranath Chowdary. Combating hate speech using an adaptive ensemble learning model with a case study on covid-19. *Expert Systems with Applications*, 185:115632, 2021.

[59] Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer, 2018.

[60] Usman Ahmed and Jerry Chun-Wei Lin. Deep explainable hate speech active learning on social-media data. *IEEE Transactions on Computational Social Systems*, 2022.

[61] Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374, 2021.

[62] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, pages 187–195, 2021.

[63] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference 2020*, pages 3033–3040, 2020.

[64] Nujud Aloshban. Act: Automatic fake news classification through self-attention. In *12th ACM Conference on Web Science*, pages 115–124, 2020.

[65] Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Daehun Nyang, and David Mohaisen. Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in youtube. In *Companion Proceedings of the Web Conference 2021*, pages 508–515, 2021.

[66] Ashley A Anderson, Sara K Yeo, Dominique Brossard, Dietram A Scheufele, and Michael A Xenos. Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, 30(1):156–168, 2016.

[67] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54, 2019.

[68] Zahra Ashktorab and Jessica Vitak. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3895–3905, 2016.

[69] Australian Government. Australia esafety commissioner. https://www.esafety.gov.au/, 2015.

[70] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59, 2019.

[71] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

[72] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics, 2019.

[73] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[74] Samuel Benson. Twitter to label all state-affiliated Russia media. https://www.politico.com/news/2022/02/28/twitter-label-state-affiliated-russia-media-00012351, 2022.

[75] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. Nudgecred: Supporting news credibility assessment on social media through nudges. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30, 2021.

[76] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556, 2020.

[77] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19, 2017.

[78] Danielle Blunt, Ariel Wolf, Emily Coombes, and Shanelle Mullin. Posting into the void: studying the impact of shadowbanning on sex workers and activists. *Hacking/Hustling*, pages 1–87, 2020.

[79] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[80] Lia Bozarth and Ceren Budak. Toward a better performance evaluation framework for fake news classification. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 60–71, 2020.

[81] Alexander Brown. What is so special about online (as compared to offline) hate speech? *Ethnicities*, 18(3):297–326, 2018.

[82] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[83] Amy Bruckman, Pavel Curtis, Cliff Figallo, and Brenda Laurel. Approaches to managing deviant behavior in virtual communities. In *Conference companion on Human factors in computing systems*, pages 183–184, 1994.

[84] Cheng Cao and James Caverlee. Detecting spam urls in social media via behavioral analysis. In *European conference on information retrieval*, pages 703–714. Springer, 2015.

[85] R Caplan. Content or context moderation? artisanal, community-reliant, and industrial approaches [white paper]. data & society, 2018.

[86] Robyn Caplan and Tarleton Gillespie. Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media+ Society*, 6(2):2056305120936636, 2020.

[87] Arthur TE Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, et al. Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS, 2019.

[88] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, August 2021. Association for Computational Linguistics.

[89] Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, 28(11):1531–1546, 2017.

[90] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4):1–26, 2022.

[91] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric 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*, 1(CSCW):1–22, 2017.

[92] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

[93] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3175–3187, 2017.

[94] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22, 2017.

[95] Hao Chen, Susan McKeever, and Sarah Jane Delany. The use of deep learning distributed representations in the identification of abusive text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 125–133, 2019.

[96] Myojung Chung and John Wihbey. Social media regulation, third-person effect, and public views: A comparative study of the united states, the united kingdom, south korea, and mexico. *New Media & Society*, page 14614448221122996, 2022.

[97] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4):1073–1095, 2020.

[98] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[99] Kate Crawford and Tarleton Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428, 2016.

[100] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502, 2020.

[101] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

[102] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[103] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95, 2017.

[104] Marco L Della Vedova, Eugenio Tacchini, Stefano Moret, Gabriele Ballarin, Massimo DiPierro, and Luca de Alfaro. Automatic online fake news detection combining content and social signals. In *2018 22nd conference of open innovations association (FRUCT)*, pages 272–279. IEEE, 2018.

[105] Laura DeNardis and Andrea M Hackl. Internet governance by social media platforms. *Telecommunications Policy*, 39(9):761–770, 2015.

[106] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[107] Ángel Díaz and Laura Hecht. Double standards in social media content moderation. 2021.

[108] Julian Dibbell. A rape in cyberspace; or, how an evil clown, a haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. In *Flame wars*, pages 237–262. Duke University Press, 1994.

[109] Periwinkle Doerfler, Andrea Forte, Emiliano De Cristofaro, Gianluca Stringhini, Jeremy Blackburn, and Damon McCoy. " i'm a professor, which isn't usually a dangerous job": Internet-facilitated harassment and its impact on researchers. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–32, 2021.

[110] Ehsan Doostmohammadi, Hossein Sameti, and Ali Saffar. Ghmerti at SemEval-2019 task 6: A deep word- and character-based approach to offensive language identification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 617–621, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[111] Brooke Erin Duffy and Colten Meisner. Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility. *Media, Culture & Society*, page 01634437221111923, 2022.

[112] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 81–89, 2021.

[113] Arun Dunna, Katherine Keith, Ethan Zuckerman, Narseo Vallina-Rodriguez, Brendan O'Connor, Rishab Nithyanand, et al. Paying attention to the algorithm behind the curtain. bringing transparency to youtube's demonetization algorithms. In *ACM Conference on Computer Supported Cooperative Work*, 2022.

[114] Ahmed Elnaggar, Bernhard Waltl, Ingo Glaser, Jörg Landthaler, Elena Scepankova, and Florian Matthes. Stop illegal comments: A multi-task deep learning approach. In *Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference*, pages 41–47, 2018.

[115] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[116] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 183–193, 2022.

[117] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[118] Federal Ministry of Justice. Network enforcement act, netzdg. https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html, 2017.

[119] Thomas Felber. Constraint 2021: Machine learning models for covid-19 fake news detection shared task. *arXiv preprint arXiv:2101.03717*, 2021.

[120] Pnina Fichman and Elizabeth Peters. The impacts of territorial communication norms and composition on online trolling. *International Journal of Communication*, 13:20, 2019.

[121] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. Falling for fake news: investigating the consumption of news via social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2018.

[122] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[123] Paula Fortuna, Juan Soler, and Leo Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794, 2020.

[124] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114, 2019.

[125] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[126] Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752, 2019.

[127] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90, 2017.

[128] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, 2017.

[129] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–16, 2018.

[130] Christine Geeng, Tiona Francisco, Jevin West, and Franziska Roesner. Social media covid-19 misinformation interventions viewed positively, but have limited impact. *arXiv preprint arXiv:2012.11055*, 2020.

[131] Christine Geeng, Savanna Yee, and Franziska Roesner. Fake news on facebook and twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[132] Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence*, pages 1–6, 2018.

[133] Tarleton Gillespie. The politics of 'platforms'. *New media & society*, 12(3):347–364, 2010.

[134] Tarleton Gillespie. Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. 2018.

[135] Barney G Glaser and Anselm L Strauss. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, 2017.

[136] Eric Goldman. Why section 230 is better than the first amendment. *Notre Dame L. Rev. Online*, 95:33, 2019.

[137] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478, 2020.

[138] João Gonçalves, Ina Weber, Gina M Masullo, Marisa Torres da Silva, and Joep Hofhuis. Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *new media & society*, page 14614448211032310, 2021.

[139] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[140] Leo A Goodman. Snowball sampling. *The annals of mathematical statistics*, pages 148–170, 1961.

[141] Google Perspective API. https://www.perspectiveapi.com/, 2022.

[142] Robert Gorwa. What is platform governance? *Information, communication & society*, 22(6):854–871, 2019.

[143] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, 2020.

[144] Sukeshini Grandhi, Linda Plotnick, and Starr Roxanne Hiltz. By the crowd and for the crowd: Perceived utility and willingness to contribute to trustworthiness indicators on social media. *Proceedings of the ACM on Human-Computer Interaction*, 5(GROUP):1–24, 2021.

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

[146] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12, 2018.

[147] Lucie Guibault, Guido Westkamp, and Thomas Rieber-Mohn. Study on the implementation and effect in member states' laws of directive 2001/29/ec on the harmonisation of certain aspects of copyright and related rights in the information society. *Report to the European Commission, DG Internal Market, February*, pages 2012–28, 2007.

[148] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of misinformation detection: new perspectives and trends. *arXiv preprint arXiv:1909.03654*, 2019.

[149] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, 2021.

[150] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 117:47–58, 2021.

[151] Md Mahfuzul Haque, Mohammad Yousuf, Ahmed Shatil Alam, Pratyasha Saha, Syed Ishtiaque Ahmed, and Naeemul Hassan. Combating misinformation in bangladesh: roles and responsibilities as perceived by journalists, fact-checkers, and users. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–32, 2020.

[152] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812, 2017.

[153] Stefan Helmstetter and Heiko Paulheim. Weakly supervised learning for fake news detection on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277. IEEE, 2018.

[154] Benjamin D Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh international AAAI conference on web and social media*, 2017.

[155] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24, 2021.

[156] Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*, 2018.

[157] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763, 2021.

[158] Human Rights Watch. Germany: Flawed social media law. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law, 2018.

[159] Hyunseo Hwang, Porismita Borah, Kang Namkoong, and A Veenstra. Does civility matter in the blogosphere? examining the interaction effects of incivility and disagreement on citizen attitudes. In *58th Annual Conference of the International Communication Association, Montreal, QC, Canada*, 2008.

[160] Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. Fermi at semeval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 70–74, 2019.

[161] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):1–20, 2020.

[162] Farnaz Jahanbakhsh, Amy X Zhang, Adam J Berinsky, Gordon Pennycook, David G Rand, and David R Karger. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–42, 2021.

[163] Catherine Jennifer, Fatemeh Tahmasbi, Jeremy Blackburn, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. Feels bad man: Dissecting automated hateful meme detection through the lens of facebook's challenge. *arXiv preprint arXiv:2202.08492*, 2022.

[164] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. " did you suspect the post would be removed?" understanding user reactions to content removals on reddit. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–33, 2019.

[165] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019.

[166] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30, 2021.

[167] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33, 2018.

[168] Chenyan Jia, Alexander Boltz, Angie Zhang, Anqing Chen, and Min Kyung Lee. Understanding effects of algorithmic vs. community label on perceived accuracy of hyper-partisan misinformation. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 2022.

[169] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017.

[170] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816, 2017.

[171] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[172] Prerna Juneja, Deepika Rama Subramanian, and Tanushree 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.

[173] Ben Kaiser, Jerry Wei, Elena Lucherini, Kevin Lee, J Nathan Matias, and Jonathan Mayer. Adapting security warnings to counter online disinformation. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021.

[174] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.

[175] Satyajit Kamble and Aditya Joshi. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *arXiv preprint arXiv:1811.05145*, 2018.

[176] Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1546–1557, 2018.

[177] Muhammad US Khan, Assad Abbas, Attiqa Rehman, and Raheel Nawaz. Hateclassify: A service framework for hate speech identification on social media. *IEEE Internet Computing*, 25(1):40–49, 2020.

[178] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.

[179] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020.

[180] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design*, pages 125–178, 2012.

[181] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 324–332, 2018.

[182] Sangyeon Kim, Omer F Yalcin, Samuel E Bestvater, Kevin Munger, Burt L Monroe, and Bruce A Desmarais. The effects of an informational intervention on attention to anti-vaccination content on youtube. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 949–953, 2020.

[183] Jan Kirchner and Christian Reuter. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–27, 2020.

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

[185] Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–25, 2022.

[186] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3), 2020.

[187] Sangeet Kumar. The algorithmic dance: Youtube's adpocalypse and the gatekeeping of cultural content on digital platforms. *Internet Policy Review*, 8(2):1–21, 2019.

[188] Nihal Kumarswamy et al. *"Strict Moderation?" The Impact of Increased Moderation on Parler Content and User Behavior*. PhD thesis, 2022.

[189] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao 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*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[190] Kyle Langvardt. Regulating online content moderation. *Geo. LJ*, 106:1353, 2017.

[191] Erwan Le Merrer, Benoît Morgan, and Gilles Trédan. Setting the record straighter on shadow banning. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

[192] Jiaxuan Li and Yue Ning. Anti-asian hate speech detection via data augmented semantic relation inference. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 607–617, 2022.

[193] Chen Ling, Krishna P Gummadi, and Savvas Zannettou. " learn the facts about covid-19": Analyzing the use of warning labels on tiktok videos. *arXiv preprint arXiv:2201.07726*, 2022.

[194] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020.

[195] Yang Liu and Yi-Fang Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[196] Yang Liu and Yi-Fang Brook Wu. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–33, 2020.

[197] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[198] Yi-Ju Lu and Cheng-Te Li. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online, July 2020. Association for Computational Linguistics.

[199] Zhicong Lu, Yue Jiang, Cheng Lu, Mor Naaman, and Daniel Wigdor. The government's dividend: complex perceptions of social media misinformation in china. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[200] Zhicong Lu, Yue Jiang, Chenxinran Shen, Margaret C Jack, Daniel Wigdor, and Mor Naaman. " positive energy" perceptions and attitudes towards covid-19 information on social media in china. *Proceedings of the ACM on human-computer interaction*, 5(CSCW1):1–25, 2021.

[201] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. What's the appeal? perceptions of review processes for algorithmic decisions. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.

[202] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics, 2018.

[203] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. Discovering the sweet spot of human-computer configurations: A case study in information extraction. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

[204] Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, 2020.

[205] Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria, September 2017. INCOMA Ltd.

[206] Shervin Malmasi and Marcos Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202, 2018.

[207] Daniel Martens and Walid Maalej. Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering*, 24(6):3316–3355, 2019.

[208] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380, 2019.

[209] J Nathan Matias. The civic labor of online moderators. In *Internet Politics and Policy conference. Oxford, United Kingdom*, 2016.

[210] Patrick McDaniel, Nicolas Papernot, and Z Berkay Celik. Machine learning in adversarial settings. *IEEE Security & Privacy*, 14(3):68–72, 2016.

[211] Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380, 2022.

[212] Shahan Ali Memon and Kathleen M Carley. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*, 2020.

[213] Paul Mena. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & internet*, 12(2):165–183, 2020.

[214] Ministry of Electronics and Information Technology. Information technology(intermediary guidelines and digital media ethics code) rules, 2021. https://mib.gov.in/sites/default/files/IT%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20English.pdf, 2021.

[215] Ministry of Electronics and Information Technology. Information technology(intermediary guidelines and digital media ethics code) rules, 2021. https://www.meity.gov.in/writereaddata/files/Intermediary_Guidelines_and_Digital_Media_Ethics_Code_Rules-2021.pdf, 2021.

[216] Rahul Mishra. Fake news detection using higher-order user to user mutual-attention progression in propagation paths. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 652–653, 2020.

[217] Sandip Modha, Prasenjit Majumder, and Daksh Patel. Da-ld-hildesheim at semeval-2019 task 6: tracking offensive content with deep learning using shallow representation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 577–581, 2019.

[218] Ayme Arango Monnar, Jorge Pérez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, 2022.

[219] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. 2019.

[220] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer, 2019.

[221] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.

[222] Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200, 2012.

[223] Kevin Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, 2017.

[224] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. Experiences of harm, healing, and joy among black women and femmes on social media. In *CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

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

[226] Xia-mu Niu and Yu-hua Jiao. An overview of perceptual hashing. *ACTA ELECTONICA SINICA*, 36(7):1405, 2008.

[227] Debora Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, 2021.

[228] Brendan Nyhan, Ethan Porter, Jason Reifler, and Thomas J Wood. Taking fact-checks literally but not seriously? the effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, 42(3):939–960, 2020.

[229] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China, November 2019. Association for Computational Linguistics.

[230] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022.

[231] Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. Content based fake news detection using knowledge graphs. In *International semantic web conference*, pages 669–683. Springer, 2018.

[232] Orestis Papakyriakopoulos and Ellen Goodman. The impact of twitter labels on misinformation spread and user engagement: Lessons from trump's election tweets. In *Proceedings of the ACM Web Conference 2022*, pages 2541–2551, 2022.

[233] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.

[234] Sungkyu Park, Jamie Yejean Park, Hyojin Chin, Jeong-han Kang, and Meeyoung Cha. An experimental study to understand user experience and perception bias occurred by fact-checking messages. In *Proceedings of the Web Conference 2021*, pages 2769–2780, 2021.

[235] Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D Dikaiakos, and Evangelos Markatos. Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–21, 2020.

[236] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11):4944–4957, 2020.

[237] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*, 2018.

[238] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, 2018.

[239] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[240] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 518–527. IEEE, 2019.

[241] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI*, volume 18, pages 3834–3840, 2018.

[242] Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3550–3559, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[243] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.

[244] Adrian Rauchfleisch and Jonas Kaiser. Deplatforming the far-right: An analysis of youtube and bitchute. *Available at SSRN*, 2021.

[245] Sahil Raut, Nikhil Mhatre, Sanskar Jha, and Aditi Chhabria. Hate classifier for social media platform using tree lstm. In *ITM Web of Conferences*, volume 44, page 03034. EDP Sciences, 2022.

[246] Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. Adversarial active learning based heterogeneous graph neural network for fake news detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 452–461. IEEE, 2020.

[247] Alison Ribeiro and Nádia Silva. Inf-hateval at semeval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 420–425, 2019.

[248] Martin J Riedl, Kelsey N Whipple, and Ryan Wallace. Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others. *Information, Communication & Society*, pages 1–18, 2021.

[249] Sarah T Roberts. Commercial content moderation: Digital laborers' dirty work. 2016.

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

[251] Richard Rogers. Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229, 2020.

[252] Björn Ross, Anna Jung, Jennifer Heisel, and Stefan Stieglitz. Fake news on social media: The (in) effectiveness of warning messages. 2018.

[253] Alon Rozental and Dadi Biton. Amobee at SemEval-2019 tasks 5 and 6: Multiple choice CNN over contextual embedding. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 377–381, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[254] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

[255] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. Hateminers: detecting hate speech against women. *arXiv preprint arXiv:1812.06700*, 2018.

[256] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. Hatemonitors: Language agnostic abuse detection in social media. *arXiv preprint arXiv:1909.12642*, 2019.

[257] Haji Mohammad Saleem and Derek Ruths. The aftermath of disbanding an online hateful community. *arXiv preprint arXiv:1804.07354*, 2018.

[258] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[259] Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1–34, 2020.

[260] Emily Saltz, Claire R Leibowicz, and Claire Wardle. Encounters with visual misinformation and labels across platforms: An interview and diary study to inform ecosystem approaches to misinformation interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery.

[261] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. " they don't leave us alone anywhere we go" gender and digital abuse in south asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[262] Justin Sampson, Fred Morstatter, Liang Wu, and Huan Liu. Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 2377–2382, 2016.

[263] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[264] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics.

[265] Anna Schmidt and Michael 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*, pages 1–10, 2017.

[266] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. Drawing from justice theories to support targets of online harassment. *new media & society*, 23(5):1278–1300, 2021.

[267] Sarita Schoenebeck, Carol F Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. Youth trust in social media companies and expectations of justice: Accountability and repair after online harassment. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–18, 2021.

[268] Christof Schuster. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, 64(2):243–253, 2004.

[269] Joseph Seering. Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact.*, 3, 2020.

[270] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.

[271] Haeseung Seo, Aiping Xiong, and Dongwon Lee. Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. In *Proceedings of the 10th ACM Conference on Web Science*, pages 265–274, 2019.

[272] Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. If you have a reliable source, say something: Effects of correction comments on covid-19 misinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 896–907, 2022.

[273] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media*, 22:100104, 2021.

[274] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908. IEEE, 2021.

[275] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. Misinformation warnings: Twitter's soft moderation effects on covid-19 vaccine belief echoes. *Computers & security*, 114:102577, 2022.

[276] Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. Identifying coordinated accounts on social media through hidden influence and group behaviours. KDD '21, New York, NY, USA, 2021. Association for Computing Machinery.

[277] Qinlan Shen and Carolyn Rose. The discourse of online content moderation: Investigating polarized user responses to changes in reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 58–69, 2019.

[278] Qinlan Shen and Carolyn P Rosé. A tale of two subreddits: Measuring the impacts of quarantines on political engagement on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 932–943, 2022.

[279] Imani N Sherman, Jack W Stokes, and Elissa M Redmiles. Designing media provenance indicators to combat fake media. In *24th International Symposium on Research in Attacks, Intrusions and Defenses*, pages 324–339, 2021.

[280] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[281] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 626–637, 2020.

[282] Kai Shu, Suhang Wang, and Huan Liu. Understanding user profiles on social media for fake news detection. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 430–435. IEEE, 2018.

[283] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320, 2019.

[284] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439, 2019.

[285] Vivek K Singh, Souvick Ghosh, and Christin Jose. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2090–2099, 2017.

[286] Mohit Singhal. *Analysis and Categorization of Drive-By Download Malware Using Sandboxing and Yara Ruleset*. PhD thesis, The University of Texas at Arlington, 2019.

[287] Mohit Singhal, Nihal Kumarswamy, Shreyasi Kinhekar, and Shirin Nilizadeh. Cybersecurity misinformation detection on social media: Case studies on phishing reports and zoom's threats. *arXiv preprint arXiv:2110.12296*, 2021.

[288] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022*, pages 726–734, 2022.

[289] Michael Soha and Zachary J McDowell. Monetizing a meme: Youtube, content id, and the harlem shake. *Social Media+ Society*, 2(1):2056305115623801, 2016.

[290] Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437, 2021.

[291] Chenguang Song, Kai Shu, and Bin Wu. Temporally evolving graph neural network for fake news detection. *Information Processing & Management*, 58(6):102712, 2021.

[292] Saurabh Srivastava, Prerna Khurana, and Vartika Tewari. Identifying aggression and toxicity in comments using capsule network. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 98–105, 2018.

[293] Nili Steinfeld. "i agree to the terms and conditions":(how) do users read privacy policies online? an eye-tracking experiment. *Computers in human behavior*, 55:992–1000, 2016.

[294] John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

[295] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette*, 80(4):385–400, 2018.

[296] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. What do we mean when we talk about transparency? toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13:18, 2019.

[297] Eric S Tellez, Daniela Moctezuma, Sabino Miranda-Jiménez, and Mario Graff. An automated text categorization framework based on hyperparameter optimization. *Knowledge-Based Systems*, 149:110–123, 2018.

[298] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. "it's common and a part of being a content creator": Understanding how creators experience and cope with hate and harassment online. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.

[299] Pamela Bilo Thomas, Daniel Riehm, Maria Glenski, and Tim Weninger. Behavior change in response to subreddit bans and external events. *IEEE Transactions on Computational Social Systems*, 8(4):809–818, 2021.

[300] Amaury Trujillo and Stefano Cresci. Make reddit great again: Assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 2022.

[301] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018*, pages 517–524, 2018.

[302] Jacqueline Urakami, Yeongdae Kim, Hiroki Oura, and Katie Seaborn. Finding strategies against misinformation in social media: A qualitative study. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[303] US Copyright Office, Washington, DC, USA. *Copyright Law of the United States (Title 17)*, 2021 [Online].

[304] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. " at the end of the day facebook does what itwants" how users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–22, 2020.

[305] Ameya Vaidya, Feng Mai, and Yue Ning. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693, 2020.

[306] Rama Adithya Varanasi, Joyojeet Pal, and Aditya Vashistha. Accost, accede, or amplify: Attitudes towards covid-19 misinformation on whatsapp in india. In *CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[307] Francielle Alves Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Alexandre Salgueiro Pardo. Contextual lexicon-based approach for hate speech and offensive language detection. *arXiv preprint arXiv:2104.12265*, 2021.

[308] Neeraj Vashistha and Arkaitz Zubiaga. Online multilingual hate speech detection: experimenting with hindi and english social media. *Information*, 12(1):5, 2020.

[309] Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*, 2020.

[310] Bertie Vidgen and Taha Yasseri. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78, 2020.

[311] Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*, 2021.

[312] Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 275–284, 2018.

[313] Svitlana Volkova and Jin Yea Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, pages 575–583, 2018.

[314] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[315] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[316] Richard Ashby Wilson and Molly K Land. Hate speech on social media: Content moderation in context. *Conn. L. Rev.*, 52:1029, 2020.

[317] Kevin Winter and Roman Kern. Know-center at semeval-2019 task 5: multilingual hate speech detection on twitter using cnns. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 431–435, 2019.

[318] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, pages 637–645, 2018.

[319] Liang Wu, Fred Morstatter, Xia Hu, and Huan Liu. Mining misinformation in social media. *Big data in complex and social networks*, pages 123–152, 2016.

[320] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569, 2021.

[321] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[322] Tomer Wullach, Amir Adler, and Einat Minkov. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[323] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents' needs for addressing online harm. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.

[324] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 2501–2510, 2022.

[325] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18, 2019.

[326] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–14, 2020.

[327] Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 2021.

[328] Kanwal Yousaf and Tabassam Nawaz. A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access*, 10:16283–16298, 2022.

[329] Hua Yuan, Jie Zheng, Qiongwei Ye, Yu Qian, and Yan Zhang. Improving fake news detection with domain-adversarial and graph-attention neural network. *Decision Support Systems*, 151:113633, 2021.

[330] Savvas Zannettou. " i won the election!": An empirical analysis of soft moderation interventions on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 865–876, 2021.

[331] Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. Measuring and characterizing hate speech on news websites. In *12TH ACM WEB SCIENCE CONFERENCE*. ACM, 2020.

[332] Asta Zelenkauskaite, Pihla Toivanen, Jukka Huhtamäki, and Katja Valaskivi. Shades of hatred online: 4chan duplicate circulation surge during hybrid media events. *First Monday*, 2021.

[333] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California, June 2016. Association for Computational Linguistics.

[334] Xianchao Zhang, Shaoping Zhu, and Wenxin Liang. Detecting spam and promoting campaigns in the twitter social network. In *2012 IEEE 12th international conference on data mining*, pages 1194–1199. IEEE, 2012.

[335] Yixuan Zhang, Nurul Suhaimi, Nutchanon Yongsatianchot, Joseph D Gaggiano, Miso Kim, Shivani A Patel, Yifan Sun, Stacy Marsella, Jacqueline Griffin, and Andrea G Parker. Shifting trust: Examining how trust and distrust emerge, transform, and collapse in covid-19 information seeking. In *CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2022.

[336] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer, 2018.

[337] Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*, pages 500–507, 2021.

[338] Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25, 2020.

[339] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[340] C Ziems, B He, S Soni, and S Kumar. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. arxiv 2020. *arXiv preprint arXiv:2005.12423*.

[341] Steven Zimmerman, Udo Kruschwitz, and Chris Fox. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

A. Analysis of Community Guidelines of other countries:

We investigated and analyzed community guidelines of five countries, with the highest populations in each continent i.e., India, Germany, Australia, Brazil, and South Africa and used a VPN location, to access the community guidelines for the platforms. We excluded China because most of the platforms investigated in our paper are prohibited in China, and Chinese users rely on VPNs to access these websites, causing IP addresses to be inaccurate. Interestingly, we found no change in the guidelines. Table 3 shows them in detail. TikTok was banned in India in 2020, due to a geopolitical dispute with China [56].

B. Related work conference list

Table 4 shows the distribution of papers from various conferences and journals that were analyzed by the authors.

TABLE 3: Community guidelines analysis of different countries. Note that ✓ denotes that the guidelines were similar to USA and × that the platform is banned in the country studied.

Platforms	Germany	India	Australia	Brazil	South Africa
Parler	✓	✓	✓	✓	✓
4chan	✓	✓	✓	✓	✓
Gab	✓	✓	✓	✓	✓
MeWe	✓	✓	✓	✓	✓
Rumble	✓	✓	✓	✓	✓
TikTok	✓	×*	✓	✓	✓
Twitch	✓	✓	✓	✓	✓
Tumblr	✓	✓	✓	✓	✓
Reddit	✓	✓	✓	✓	✓
Snapchat	✓	✓	✓	✓	✓
YouTube	✓	✓	✓	✓	✓
Twitter	✓	✓	✓	✓	✓
Facebook	✓	✓	✓	✓	✓
Instagram	✓	✓	✓	✓	✓

TABLE 4: Distribution of paper analyzed by the authors

Conference	Number of papers
IEEE S&P	2
ACM CCS	1
NDSS	0
USENIX Security	1
IEEE Euro S&P	0
Proc. of ACM CSCW	26
ACM CHI	25
WWW	14
ICWSM	16
AAAI AI	6
ACL Anthology	37
ACM WebSci	6
NeurIPS	3
IEEE/CVF	2
ACM KDD	5
ACM WSDM	4
ACM CIKM	2
Other Conf. & Workshops	35
New Media & Society	6
Political Behaviour	5
Other Journals	52
Law Reviews	7
Thesis	1
Books	4
Transactions	4
Others	9

C. Related work paper list

Table 5 shows the paper that conducted user studies to understand moderation. In the table, we show the category of paper that was created in Section 3, we also show which platform the authors conducted their study on, the sample size, the demographics of the study and they type of the study conducted. Note that in sample size, if there is a number in red color brackets, that means that the authors of the study conducted multiple studies. We however, present the demographics of the study with highest number of participants.

Table 6 shows the papers that are about hate speech and misinformation detection. Table 7 shows the papers that are data driven study to unpack the intricate details about content moderation. Similar to Table 5, we characterize the papers into the category, the platforms they studied. In this table, we also give the size of the dataset, if the dataset is public, whether the authors created the dataset or they used some other papers dataset, the type of data used and the intervention studied. The paper that are surrounded by red brackets, signifies that the authors studied both hate speech and misinformation moderation, whereas, the paper that is given in blue color signifies that that paper also conducted a user study.

TABLE 5: User Study Paper Analysis

Paper	Categories	Platforms	Size	Demographics	Study
Hate Speech Moderation					
[267]	Support, Removal Comprehension, Fairness and Bias	MyVoice	832	Only youth participants (14-24), 54.4% females	Survey
[266]	Effectiveness, Fairness and Bias	MTurk	573	Diverse demographics with transgender and non-binary participants, left-leaning (57%)	Survey
[167]	Effectiveness, Removal Comprehension, Fairness and Bias	Twitter	14	Mix of gender, mostly from US	Semi-structured interviews
[77]	Effectiveness	HeartMob	18	10 participants females, median age of participants was 40, high number of non-heterosexual participant	Semi-structured interviews
[224]	Effectiveness, Fairness and Bias	Social media platforms and flyers	49	Black female participants, 30 participates from ages 18-30 years	Survey
[298]	Effectiveness	Residency program	135	56% female, 99% created content for YouTube, 30% from the ages 35-44	Survey
[261]	Effectiveness	NGOs	199	Women participants from south-east Asia, 103 from India	Survey
[109]	Effectiveness	Personal contacts	17	10 females, 4 LGBTQ+ participants, predominately white participants	Survey
[138]	Support, Removal Comprehension, Fairness and Bias	Dynata	2,870	Equal representations from US, Portugal and The Netherlands	Survey
[323]	Effectiveness, Support	College students	28	All participants from 18-20 years old, 18 participants were female, 23 students were of asian decent	Interviews
Misinformation Moderation					
[131]	Consumption of (fake/misinformation) news, Engagement of user	Facebook and Twitter	25	One-third are college students, left leaning	Semi-structured interviews and Simulation
[130]	Consumption of (fake/misinformation) news, Support	Facebook and Twitter	311	Majority 18-24 year old (130), 38% from UK, majority left leaning	Mixed-method survey
[121]	Consumption of (fake/misinformation) news	Facebook	309 (9)	Majority 18-25 year old (70%), 55% female, All from UK	Pilot study and simulation
[173]	Effectiveness	MTurk	238 (40)	Two-thirds male, over half of participants from 30-49 year old (92), left-leaning (159)	Pilot study and simulation
[162]	Consumption of (fake/misinformation) news,Fairness and Bias	MTurk	1,807	47% left-leaning, median age 35 year old, 42% females	Pilot Study and survey
[213]	Engagement of users, Effectiveness	Facebook	501	51% females, majority left-leaning (47.6%), avg. age 36 year	Survey
[97]	Effectiveness	Facebook	2,994	Majority female (54%), majority left-leaning (58%)	Survey
[275]	Effectiveness	Twitter	319	56.4% males,33.3% from 25-34 age group, highly left-leaning (49.2%)	Survey
[129]	Effectiveness	MTurk	122	60% females, almost equal number political leaning, 80% from 25 to 54 year old	Survey
[260]	Effectiveness, Support, Fairness and Bias	Dscout	23 (15)	Equal genders,multiple ethnicities,balanced political views	Semi-structured interview and diary study
[228]	Effectiveness	Morning Consult and MTurk	4,186 (1,546)	54% females, largely left-leaning (49%), 42% from 18 to 34 years old	Survey
[271]	Effectiveness	MTurk	800	55% females, 75% participants between the age of 20 to 40 years	Survey
[116]	Engagement of users, Effectiveness	Lucid	1,473	Mean age of participants were 47.87, 54.1% were female	Survey
[236]	Engagement of users, Effectiveness	MTurk	5,271 (1,568)	Female dominated (55%), largely left leaning	Survey
[252]	Effectiveness	Facebook	151	57% females, mean age was 25 year old	Survey
[234]	Engagement of users, Effectiveness	MTurk	11,145	52.9% female participants	Pilot study and survey
[279]	Consumption of (fake/misinformation) news	MTurk	1,456 (24)	Almost equal number of male and females,large number of participants from 31-40 year of age	Pilot study and survey

[302]	Consumption of (fake/misinformation) news	Twitter	15	66% females, non-native English speakers	Survey
[306]	Consumption of (fake/misinformation) news	WhatsApp	28	53% males, average age of 32	Survey
[151]	Consumption of (fake/misinformation) news	Facebook	519	74% of the participants were from the 21-30 age range,64% were students	Survey and interviews
[199]	Consumption of (fake/misinformation) news, Effectiveness	WeChat	44	23 were male participants, 38 partipants were residing in China	Semi-structured interviews
[168]	Engagement of users, Effectiveness	MTurk	1,677	Majority left-leaning 52%, majority female 55%	Simulation
[144]	Effectiveness	MTurk	376	57% females, 39% from the ages 18-24 years, 52% students	Survey
[183]	Effectiveness	Respondi	2,057	Mostly male participants	Survey, Semi-structured interview and simulation
[272]	Engagement of users, Effectiveness	MTurk	2,841	Mostly female (52.3%), largely between the ages of 28-37	Survey
[326]	Engagement of users, Effectiveness	MTurk	1,512	mostly male participants (51%), slightly older crowd mean age 38	Survey
Moderation Techniques					
[248]	Support	U.S. national panel survey	1,022	53.2% female, 75% white ethnicity, 38.4% from the ages 30-49	Survey
[225]	Support, Removal Comprehension, Fairness and Bias	OnlineCensorship.org	519	Participants largely from US (295)	Survey
[164]	Effectiveness, Support, Removal Comprehension, Fairness and Bias	Reddit	907	Participants from 81 countries, highest from US (61%), majority male (81%), under 25 year old (55%)	Survey
[149]	Removal Comprehension, Fairness and Bias	Prolific, Qualtrics	909 (207)	Mixed genders, balanced ethnicity's, largely young, mix of conservatives and moderates	Survey
[270]	Removal Comprehension	Twitch, Reddit, Facebook	56	Largly female and LGBTQ participants	Semi-structured interviews
[117]	Removal Comprehension	Yelp	15	40% were between ages 35-44	Survey
[75]	Consumption of (fake/misinformation) news	Twitter	430 (12)	Slightly skewed towards females, higher number of independents and republicans	Pilot Study and Simulation
[200]	Consumption of (fake/misinformation) news	WeChat	33	More number of females (18), average age of 34	Semi-structured interviews
[335]	Consumption of (fake/misinformation) news	Qualtrics	177 (21)	Largely female (66%), between the ages of 25-34	Pilot study and survey
[172]	Removal Comprehension, Fairness and Bias	Reddit	13	70% male population, 2 people from other countries other than US	Interviews
[96]	Support	Prolific, Qualtrics, Embrain	5,392	Diverse population from US, UK, South Korea and Mexico, female dominated study (52.35%), older participants (median age = 41)	Survey
[111]	Support, Removal Comprehension, Fairness and Bias	TikTok, YouTube, Twitch	30	All participants had historically marginalized identities	Semi-structured interviews
[201]	Support, Fairness and Bias	MTurk	100	62% male participants, all from USA, older participants (avg. age = 37)	Survey
[296]	Removal Comprehension, Fairness and Bias	OnlineCensorship.org	380	N/A	Survey
[78]	Support, Fairness and Bias	Twitter, Instagram	262	Females, 38.9% were sex workers and activists	Survey

TABLE 6: Papers about Detection Techniques

Categories	Broad Description	Papers
Hate Speech Detection		
Content based	What type of features (i.e., TF-IDF, BoWV, POS tags etc.) can we obtain solely from the content to detect hate speech?	[101], [102], [255], [258]
Lexicon Based	How can we utilize the syntactic and semantic orientations of the existing lexicons to identify hate speech?	[72], [87], [115], [126], [205], [263], [297], [307], [310], [321]
Deep Neural Based	How can we use the advancements in Deep Neural Networks to identify hate speech in platforms without using the handcrafted features ?	[58]–[61], [65], [70], [71], [82], [88], [95], [106], [110], [114], [123], [124], [127], [132], [137], [146], [160], [175], [177], [192], [197], [206], [217], [220], [221], [229], [233], [237], [242], [245], [247], [253], [256], [292], [305], [308], [317], [322], [328], [336], [337], [340]
Hybrid Approaches	How can we effectively combine content based, lexicon based and deep neural based methods to identify hate speech?	[63], [94], [103], [128], [208], [235], [259]
Multi-modal Based	How can we combine multiple modalities of posts (such as images, texts, memes) together to identify hate speech?	[137], [179], [194], [239], [285], [309], [311], [325]
Misinformation Detection		
Content Based	What type of features (i.e., TF-IDF, BoWV, POS tags etc.) can we obtain solely from the content to detect misinformation?	[104], [119], [150], [154], [156], [238], [243], [287], [313], [314], [338]
Propagation Structure	How can we detect posts that contain misinformation by mining their spreading patterns in the underlying social networks?	[195], [216], [262], [281], [282], [284], [287], [318]
Hybrid Approaches	How can we effectively combine both content based features and propagation structure to detect misinformation?	[153], [170], [176], [254], [283]
Crowd Intelligence	Can we use the wisdom of the crowd to identify misinformation on the platform?	[171], [181], [202], [241], [280], [301], [312]
Deep Neural Based	How can we use the advancements in Deep Neural Networks to identify misinformation in platforms without using the handcrafted features ?	[64], [76], [100], [174], [196], [198], [202], [211], [219], [246], [254], [280], [291], [315], [318], [324], [329], [341]
Knowledge Based	Can we effectively use Knowledge Graphs to identify and automatically detect posts that contains misinformation?	[100], [112], [157], [185], [211], [231], [291], [324]
Multi-modal Based	How can we combine multiple modalities of posts (such as images, texts, and audio) together to identify misinformation?	[178], [240], [274], [288], [290], [315], [320]

TABLE 7: Data Driven Study Paper Analysis

Paper	Category	Platforms	Size of Dataset	Is Dataset Public?	Created the dataset?	Type of Data used	Intervention Studied
Hate Speech Moderation							
[90]	Engagement of users, Effectiveness	Reddit	∼85M	No	Yes	Posts and metadata	Soft Moderation
[155]	Engagement of users, Effectiveness	Reddit	∼6.3M	Yes	Partially	Comments, posts and metadata	Hard Moderation
[91]	Engagement of users, Effectiveness	Reddit	∼100M	Partially	No	Comments and posts	Hard Moderation
[62]	Engagement of users, Effectiveness	Gab, Reddit and Twitter	∼30M	Partially	Partially	Posts and metadata	Hard Moderation
[166]	Engagement of users, Effectiveness	Twitter	∼49M	No	Yes	Tweets and metadata	Hard Moderation
[257]	Engagement of users, Effectiveness	Reddit	∼1.9M	Partially	No	Comments and metadata	Hard Moderation
[92]	Effectiveness	Reddit	∼2.8M	No	Yes	Comments	Hard Moderation
[223]	Effectiveness	Twitter	N/A	No	Yes	Tweets and metadata	Hard Moderation
[300]	Engagement of users, Effectiveness	Reddit	∼15M	Yes	No	Posts	Soft Moderation
[244]	Engagement of users, Effectiveness	YouTube and BitChute	∼11K	No	Partially	Videos and metadata	Hard Moderation
[251]	Effectiveness	Telegram	N/A	No	Yes	Posts and metadata	Hard Moderation
[188]	Effectiveness	Parler	∼200M	Partially	Partially	Posts and metadata	Hard Moderation
[278]	Engagement of users, Effectiveness	Reddit	∼3.7M	Partially	No	Posts, metadata	Soft Moderation
Misinformation Moderation							
[330]	Engagement of users	Twitter	∼18K	Yes	Yes	Tweets and metadata	Soft Moderation
[193]	Engagement of users	TikTok	∼41K	No	Yes	Videos and metadata	Soft Moderation
[182]	Engagement of users, Effectiveness	YouTube	105	No	Yes	Videos metadata	Soft Moderation
[232]	Engagement of users, Effectiveness	Twitter	∼2.4M	Yes	Yes	Tweets and metadata	Soft Moderation
Moderation Techniques							
[277]	Fairness and Bias	Reddit	∼9K	Yes	No	Posts	Soft Moderation
[299]	Effectiveness	Reddit	N/A	Partially	No	Posts and comments	Hard Moderation
[172]	Removal Comprehension, Fairness and Bias	Reddit	∼0.5M	No	Yes	Moderation logs	Hard Moderation