**Evaluating the Impact of Internet Restrictions and Societal Unrest on Content Moderation Abuse on Social Media**

|  |  |  |
| --- | --- | --- |
| Potru Sai Bharath | Koti Naveen Chakravarthy | Elham Pourabbas Vafa |
| The University of Texas at Arlington | The University of Texas at Arlington | The University of Texas at Arlington |
| Arlington, TX, United States | Arlington, TX, United States | Arlington,TX, USA |

ABSTRACT

n recent years, the intersection of digital governance, internet freedom, and content moderation has emerged as a critical area of study, particularly in regions experiencing political upheaval and internet censorship. This research explores the complex dynamics of content moderation abuse on social media platforms during periods of societal unrest and government-imposed internet restrictions, with a primary focus on the 2022–2023 protests in Iran following the death of Mahsa Amini. Our study investigates how both users and platforms may exploit moderation systems to suppress dissenting voices or promote certain narratives, often under the guise of safety and regulatory compliance. We employ a mixed-methods approach, integrating large-scale data collection from Twitter in both English and Persian using protest-related hashtags and security-focused keywords. A multilingual dataset of over 50 million tweets was compiled and refined using advanced natural language processing techniques, including keyword filtering, lemmatization, and stem-based normalization. Manual annotation of representative samples enabled the creation of a robust categorization framework, which was subsequently used to train and evaluate machine learning models such as ParsBERT for Persian tweet classification. Our findings reveal distinct patterns of content moderation misuse, including mass-reporting campaigns, politically motivated takedowns, and misinformation-driven flagging. We identified key behaviors such as doxxing, coordinated harassment, and attempts to game moderation algorithms, particularly during internet outages when traditional oversight is weakened. Additionally, we observe that users often share personally identifiable information (PII) to incite targeted moderation actions and bypass restrictions using coded language or VPN-related hashtags. Through qualitative analysis and automated classification of over 300,000 tweets, we examine the ethical and societal implications of such abuse, including its impact on free speech, online safety, and the credibility of digital platforms. The study proposes policy recommendations to promote equitable moderation practices and emphasizes the need for transparent and culturally sensitive content governance models. Ultimately, this research contributes to the broader discourse on digital rights, state influence over online discourse, and the responsibilities of tech companies in safeguarding democratic participation in the digital age.

INTRODUCTION

In the era of digitalization, social media platforms’ moderation poli-cies have become essential for controlling and regulating online material and creating inclusive and safe spaces. Critical discussions on the abuse of content moderation techniques have resulted from the blurring of the thin line between protecting digital spaces and restricting the right to free speech. The goal of this research project is to investigate the complex dynamics of content moderation mis-use, with a special emphasis on countries like Iran where political upheaval and internet limitations pose specific obstacles to the advancement of digital freedoms.

This research is important because it has the ability to clarify the intricate relationship that exists between digital rights, content moderation, and societal effects. This study attempts to shed light on the patterns and practices of content moderation abuse across social media platform by utilizing a multi-modal research approach that combines qualitative user studies with meticulous quantitative data analysis. We want to use this lens to look at how these behaviors directly affect people’s rights to free speech and information access, particularly in environments where there is social and political unrest.

Furthermore, the project aims to balance online safety, free speech, and information accessibility without compromising core

user rights in order to contribute to the conversation about the ethical obligations of digital platforms. We intend to provide light on the realities of content moderation systems, their abuse, and their consequences for human rights and democratic participation by combining insights from user studies, and data analysis. In the end, this project aims to shape the future of online debate and digital democracy within a framework of moral moderation and hu-man rights by educating policymakers, promoting fair moderation practices, and increasing understanding of the issues surrounding digital governance.

It becomes essential to define specific research areas in order to progress our understanding of the dynamics involved in the misuse of content moderation systems, especially in light of recent political and social upheaval. As a result, we frame our study around two crucial research questions that try to break down the complexity of this problem with digital governance:

1. RQ1: How can internet shutdowns and societal unrest such as those that occurred in Iran impact the misuse of content moderation systems on social media?
2. RQ2: During internet outages, what patterns of abuse of con-tent moderation appear, and how may these be recognized with data analysis and machine learning?

Our investigation into how digital platforms can strike a balance between protecting users’ liberties and guaranteeing their safety is guided by these questions, with the goal of advancing the creation of appropriate and robust content moderation procedures.

RELATED WORK

The Internet’s tremendous expansion has fundamentally shifted the dynamics of global information sharing and communication. While it has revolutionized the way we interact, it has also transformed into a battleground for both individuals and governments striving to regulate online content and actions. This has prompted the adop-tion of internet filtering and restriction policies in several nations, including [Iran.[5][6]](#page4) Particularly during times of social unrest these internet limitations and filtering techniques have had a substantial impact on social media platform content [moderation.[15]](#page4) Iran’s in-ternet censorship and limitations have gotten more severe specially during these periods and social media companies have been pushed by this to enact stronger content moderation guidelines to abide with government [regulations.[4] [11]](#page4)

As a result, both individuals and platforms are abusing content moderation more frequently. During times of social chaos, users may report or flag posts in violation of the content moderation

criteria to silence opposing viewpoints or [voices.[10][7]](#page4) Conversely, platforms could abuse content moderation to manipulate narratives or appease government [authorities.[13][9][3] [12]In](#page4) the context of Iranian censorship, it is crucial to investigate how social unrest and internet restrictions affect the misuse of content moderation. The primary objective of this research is to evaluate the impact of social turmoil and internet constraints on the improper utilization of content moderation within social media platforms. Specifically, the study will concentrate on the censorship practices in Iran and how they hinder access to vital information and prevent the exercise of free speech.

A mixed-methods approach will be used for this project, integrat-ing qualitative user studies and quantitative data analysis. Patterns of content moderation violations on social media platforms will be examined as part of the quantitative study. In addition, the qualita-tive case studies will shed light on the viewpoints and experiences of those impacted by abuse of content moderation throughout these times. This study intends to determine the the volume of content moderation abuse, the tactics used by users and platforms, and the implications for freedom of expression and access to information in the context of Iranian censorship. This study aims to clarify the intricate relationship between internet restrictions, societal unrest and abuses of content moderation on social media platforms.

DATA COLLECTION

Our initial step in data collection involved curating a set of hashtags relevant to the Iran Protests, encompassing both the languages, Eng-lish and Farsi. Beginning with widely recognized hashtag "#Mah-saAmini", we employed snowball sampling technique [[8]](#page4) to identify and record related hashtags. Considering the dynamic nature of so-cial media discourse, we observed that new hashtags emerge every month. To capture the tweets accordingly, we constructed monthly specific lists of hashtags. Table provides the comprehensive list of hastags that were obtained during our study, at the end of January 2023.



Figure 1: Hashtag list for data collection

Twitter Data Collection

To collect tweets from Twitter, we utilized Twitter V2 Archive Search endpoint [[1],](#page4) which allows us to access complete historical archive of public Tweets dating back to the first Tweet in March 2006. Our data collection spanned from September 18th, 2022 to Jan-uary 31st, 2023, where discussions were heavily centered on issues of security, privacy and censorship imposed by the government. We mainly focused on English and Persian posts to capture the discussions on the usage of VPNs, proxies and related privacy tools. This approach was critical due to internet shutdown in Iran, during which such tools were utilized to bypass censorship restrictions. To collect data for each month, we utilized month specific hash-tags, running separately for both English and Persian. We excluded retweets to prioritize the conversational content consisting of main tweets and direct replies.

Upon collecting data for each month, we filtered the data to obtain the tweets related to security and privacy. To achieve this filtration process, initially we prepared a set of keywords includ-ing Internet shutdown, Internet restrictions, Internet speed, Cen-sor, Bypassing, VPN, Starlink, Proxy server, Internet filtering, Tor, Server, Internet, Net, Network, Censorship, Bypass. Later tweets were preprocessed using the NLTK WordNet Lemmatizer [[2]](#page4) to de-rive lemmas, thereby standarizing the words for effective keyword matching. Similar filtration process was applied to persian tweets as well. Given the bilingual nature of Persian posts, where tweets often contain a mixture of english and persian words, we prepared a comprehensive list of keywords in both languages. For persian tweets, we utilized PersianStemmer [[14]](#page4) to stem words, matching with our bilingual keyword list.

While filtering content with keywords, we observed that the hashtags such as vpn or server were commonly used but not al-ways associated with the Iran protests. These hashtags sometimes included general tweets about commercial VPN services such as NordVPN. Furthermore, the VPN hashtag was prevalent in tweets pertaining to issues in China at that time. This resulted in subset of data that included chinese tweets, which contained a hashtag VPN, but were unrelated to the events in Iran. To remove such irrelevant tweets, the dataset underwent another filtering process where the posts were preserved in the dataset if they contain protest related hashtags as well.

After this, the dataset was de-duplicated to ensure that each tweet was unique. The criteria for de-duplicating data were based on tweet text and its associated metadata. The final dataset comprised of tweets that were both related to identified hashtags and contained the required security and privacy keywords, post the removal of irrelevant content and duplicates.

METHODOLOGY

Farsi Tweets

Data Acquisition and Preprocessing: We began our research with an advanced data gathering strategy that utilized regular ex-pressions (regex) to filter a large dataset of 39375975 Farsi tweets using a keyword-based approach. The key feature of this regex approach was its ability to cleverly extract tweets that referenced entities like "user," "account," and "tweet," as well as specific key-words linked to content moderation actions like "block," "report,"

Evaluating the Impact of Internet Restrictions and Societal Unrest on Content Moderation Abuse on SocialThe WebMdiaConference (WWW) ’23, April 30–May 4, 2023, Austin, Texas

and "delete," thus guaranteeing relevance to our research focus on content moderation practices. The regex pattern was cleverly cre-ated to detect tweets that had solo keywords suggestive of content moderation difficulties, or that directly addressed terms linked to moderation alongside terms connected to users or content. The efficient extraction of a highly relevant subset of tweets from the larger dataset made possible by this method laid the groundwork for the in-depth research that came next. This keyword filtration approach provided us with 120836 tweets.

Sample Selection and Manual Labeling: After the initial data filtration, we took the filtered dataset and used random selection to choose 1,500 tweets for manual labeling. Developing a catego-rization schema that could guide additional automated analysis and comprehending the subtleties of content moderation concerns reflected in the dataset were made possible by this crucial step. In addition to making it easier to find common themes in the data, the manual labeling procedure guaranteed the accuracy and relevance of the categories that would be utilized in later studies. Two coders went through the random sample and labeled them independently to conduct the Cohen’s kappa analysis.

Implementation of ParsBERT for Classification: Our study used a Persian language model called ParsBERT to classify the larger dataset, utilizing the knowledge gathered from manual labeling. We were able to improve our classification technique by taking into account the subtleties of the language and the particular context of content moderation that was mentioned in the tweets, thanks to ParsBERT’s suitability for Persian text. We preprocessed our data thoroughly in order to improve the model’s performance before classifying it. The input data to ParsBERT was ideally prepared for analysis by this preprocessing, which included normalizing letters, eliminating unnecessary symbols and URLs, and making modifications for linguistic consistency. A stratified K-fold cross-validation approach was used to further enhance the classification process, enabling us to evaluate the model’s performance across various data subsets and guaranteeing the validity and consistency of our conclusions.

Results Interpretation and Model Evaluation:After classi-fication, our analysis included all of the data from September 18, 2022, to January 31, 2023, and we were able to identify 308,992 tweets that were relevant to our study’s content moderation em-phasis. This comprehensive analysis shed light on the intricacies of digital discourse regulation by highlighting the common themes of user interaction with moderation systems and provide a substantial foundation for comprehending the dynamics of content moderation abuse. All things considered, our methodological approach inte-grated thorough data preparation, perceptive manual labeling, and cutting-edge machine learning algorithms to effectively decipher the nuances of content moderation misuse.

We now focus on important questions that have surfaced in relation to user behavior and the effects of content moderation systems in an effort to better comprehend their wider ramifications.

Q1) Are users misuing the exsiting content moderation tools to supress other users?

Q2) Are users providing PII to provoke other users to misuse content moderation tools?

Q3) Are there coordinated instances of "SWATing" attacks?

Q4) What are the ill effects of misuse of content moderation on daily users life?

Q5) How are users’ circumventing content moderation practices of social media

English Tweets

Data Acquisition and Preprocessing: Our data collection pro-cess involved sifting through an extensive dataset of 11,264,108 English tweets. By employing a combination of nouns (including "user," "account," "tweet," "post," and "content"), verbs (such as "re-move," "delete," "suspend," "report," "moderate," "block," "raid," "dox," and "harass"), and meticulously chosen standalone keywords (like "content moderation," "shadowbanned," "ban," and variations of "ha-rassment"), alongside a finely crafted regular expression pattern, we pinpointed 117,256 pertinent tweets. These tweets depicted various actions related to content moderation and user interactions, en-suring their direct relevance to our investigation. This methodical approach ensured that the gathered tweets were highly pertinent to our analytical pursuits.

Sample Selection and Manual Labelling: Following the initial data filtration, we proceeded by randomly choosing 2,500 tweets for manual labeling. This step was crucial as it enabled the devel-opment of a categorization framework to aid further automated analysis. Moreover, manual labeling provided valuable insights into the subtleties of content moderation issues present in the dataset. By carefully categorizing these tweets, we ensured the precision and relevance of the categories for future studies. Additionally, we conducted Cohen’s Kappa analysis on this set of 2,500 manually labeled tweets to evaluate inter-rater agreement.

EVALUATION

The assessment of our research technique takes into account multi-ple factors to guarantee the reliability and appropriateness of our findings:

Validation of Machine Learning Models: Metrics including accuracy, precision, recall, and F1 scores will be used to evaluate how well machine learning algorithms perform in spotting patterns of content moderation abuse. This quantitative assessment will guarantee that our methods for data analysis are reliable.

Effectiveness of Qualitative insights: Carefully choosing user studies, and applying content analysis techniques methodically will all help to assure the validity of our qualitative conclusions. As a result, we can be certain that our qualitative insights are representative of the intricacies of content moderation procedures and based on actual data.

Impact Assessment: The total effect of our study will be evalu-ated by looking at how much our research adds to our knowledge of the misuse of content moderation and its consequences for online freedoms and rights. This involves considering how our findings might influence discussions about policy and platform governance tactics, as well as how applicable they are to actual situations.

This study attempts to provide a nuanced knowledge of the difficulties and ramifications related to content moderation usage in the setting of internet shutdowns and social unrest by using a methodologically sound and comprehensive approach. We endeavor to provide insightful analysis and evaluation in order to add to the

conversation around digital governance and human rights in the digital age.

REFERENCES

1. 2021. Twitter V2 API. [https://developer.twitter.com/en/docs/twitter-api/tweets/](https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction) [search/introduction.](https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction)
2. 2022. Source code for nltk.stem.wordnet. [https://www.nltk.org/\_modules/nltk/](https://www.nltk.org/_modules/nltk/stem/wordnet.html) [stem/wordnet.html.](https://www.nltk.org/_modules/nltk/stem/wordnet.html)
3. Oxford Analytica. 2023. Content moderation issues on social media are evolving. Emerald Expert Briefings oxan-db (2023).
4. Simurgh Aryan, Homa Aryan, and J Alex Halderman. 2013. Internet censorship in Iran: A first look. In 3rd USENIX Workshop on Free and Open Communications on the Internet (FOCI 13).
5. Constance Bitso, Ina Fourie, and Theo JD Bothma. 2013. Trends in transition from classical censorship to Internet censorship: selected country overviews. Innovation: journal of appropriate librarianship and information work in Southern Africa 2013, 46 (2013), 166–191.
6. Marie Eneman. 2019. Internet filtering: A solution to harmful and illegal content?. In 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Ad-

vanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smart-World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, 354–359.

1. Catalina Goanta and Pietro Ortolani. 2022. Unpacking content moderation: The rise of social media platforms as online civil courts. In Frontiers in Civil Justice. Edward Elgar Publishing, 192–216.
2. Leo A Goodman. 1961. Snowball sampling. The annals of mathematical statistics (1961), 148–170.
3. Emmie Hine. 2023. Content Moderation in the Metaverse Could Be a New Frontier to Attack Freedom of Expression. Philosophy & Technology 36, 3 (2023), 43.
4. Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Cama-cho. 2023. Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage. Applied Soft Computing 145 (2023), 110552.
5. Simin Kargar and Keith McManamen. 2018. Censorship and collateral damage: Analyzing the Telegram ban in Iran. Berkman Klein Center Research Publication 2018-4 (2018).
6. Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P). IEEE, 868–895.
7. Sebastian Smart and Alberto Coddou McManus. 2022. Closing the gap between UNGPS and content regulation/moderation practices. Braz. J. Int’l L. 19 (2022), 269.
8. Hossein Taghi-Zadeh, Mohammad Hadi Sadreddini, Mohammad Hasan Diyanati, and Amir Hossein Rasekh. 2017. A new hybrid stemming method for Persian language. Digital Scholarship in the Humanities 32, 1 (2017), 209–221.
9. BURCU SELIN YILMAZ, HÜMEYRA DOĞRU, and VOLKAN BAHÇECI. 2017. WHAT IF YOU CANNOT ACCESS THE INTERNET IN THE SURVEILLANCE SOCIETY? INDIVIDUALS’PERCEPTIONS RELATED TO THE INTERNET CEN-SORSHIP AND SURVEILLANCE IN TURKEY. Journal of Media Critiques 3, 11 (2017), 73–86.