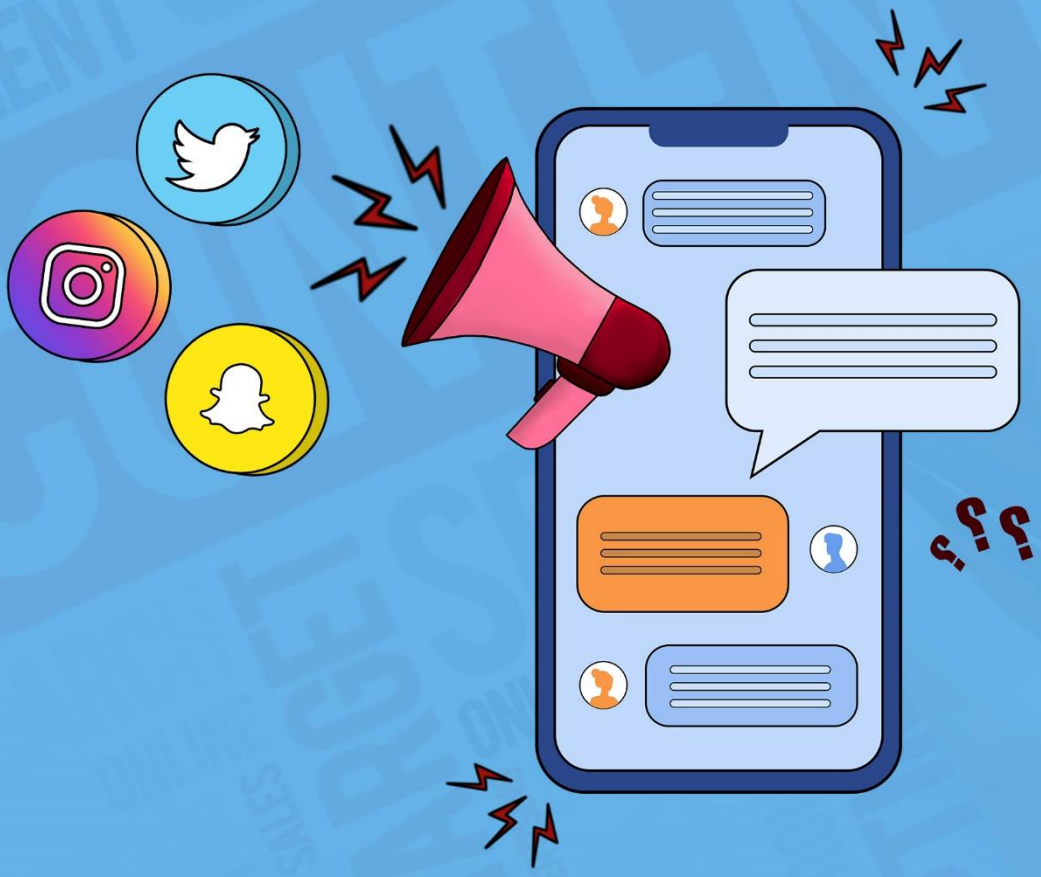
POLICY PAPER: UNPACKING CONTENT MODERATION IN PASHTO AND DARI



Meedan

DRF
DigitalRightsFoundation
"KNOW YOUR RIGHTS"

December 2022

Research: Shafeeq Gigyani and Muhammad Tufail
Supplementary Research: Asia Nabi Hassan

Author: Shmyla Khan
Editor: Noor Waheed
Designer: Ahsan Zahid

Table of Contents

Background

Content moderation and its complexities are emerging as a worldwide problem, with social media platforms failing to develop adequate policies and mechanisms to deal with content around child sexual abuse; digital violence against women and sexual and gender minorities; hate speech based on race, nationality, religion, ethnicity gender, age, sexuality, and disability; misinformation; and disinformation. Content moderation questions, particularly in light of harmful content, are deeply complex but require participatory and inclusive answers.

Content moderation is often centered around the body of the Western subject. This subject is likely to be male, white, able-bodied, and is both the subject and principal architect of content moderation mechanisms. Tech companies – that built globally used platforms and subsequently the policies governing them – are usually situated in the Global North and have become sensitive to the context of these countries. The "rest of the world", loosely referred to as the Global South, is often dealt with in one fell swoop and approached as an afterthought. Activists and some state actors have advocated extensively to put more perspectives from the Global South on the agenda of content moderation.

This "global" approach has meant that social media companies have repeatedly failed to address the specific harms that occur in countries that constitute the Global South. This has led to massive neglect of content moderation here In the recent 'Facebook revelations', it was found that Facebook's content moderation took place along tiered lines, where Brazil, India, and the United States were placed in the top tier. This meant having exclusive resources such as 'war rooms' dedicated to content moderation there. Other priority tiers, i.e. tier two, included Germany, Indonesia, Iran, Israel, and Italy, along with 22 other countries. For the "rest of the world", these special resources were not extended. As a consequence of this, the company lacked misinformation classifiers in three high-risk countries: Myanmar, Pakistan, and Ethiopia.[1]

[1] Casey Newton, "THE TIER LIST: HOW FACEBOOK DECIDES WHICH COUNTRIES NEED PROTECTION," The Verve, October 25, 2021, https://www.theverge.com/22743753/facebook-tier-list-countries-leaked-documents-content-moderation.

Additionally, the "Global South" is not a monolith. The term, when taken as a placeholder for a multitude of countries, cultures and peoples, becomes vacuous in the absence of context and lived experiences. Even within individual countries, there is so much diversity in terms of experiences, languages and marginality. This report seeks to get at the granularity of content moderation by focusing on a region typically excluded from conversations related to content moderation.

Focusing on Pakistan, cases coming from peripheral regions and in regional languages within the country might not be dealt with uniformly. It is important to deconstruct the "Global South" through focused data collection and research that can look at these issues in depth. This policy paper thus seeks to focus on the Pak-Afghan region to understand the nature of misinformation and hate speech being circulated there that is often not centered in conversations around content moderation.

There has been a gap when it comes to content moderation, where the focus has been "*driven by high-profile incidents and people: the 2016 election in the United States put worries about misinformation front and centre; the Christchurch shooting pushed hate and domestic terrorism as the highest priority; the Covid-19 pandemic put misinformation and conspiracy back in front*."[2] The data collection for this policy paper seeks to address this structural gap in our understanding of content moderation. It focuses on content emerging in Pashto and Dari in the Pak-Afghan region, particularly the north-west region of the countries in the aftermath of the US withdrawal and Taliban take-over.

Since the late summer of 2021, content has been moderated through a narrow securitized lens, emphasising the "Dangerous Individuals and Organizations (DIO)" policy at Facebook and its equivalent in other tech companies. This heavy-handed content moderation approach led to the indiscriminate removal of Taliban-related content and resulted in

[2] Tarleton Gillespie and Patricia Aufderheide. "Expanding the debate about content moderation: scholarly research agendas for the coming policy debates," Volume 9, Issue 4, Internet Policy Review, https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy.

moderation overreach where even posts and pages countering the Taliban's disinformation and presenting alternative narratives were removed.[3]

On the other hand, the pervasiveness of misinformation and hate speech on social media platforms surged during the period, adding to the uncertainty and lack of security felt in the region. Fake news regarding Taliban attacks and policies surged as anxiety concerning the takeover raged.[4]

The shift towards more automated decision-making, as opposed to human moderation, particularly during the COVID-19 pandemic, has opened up a new set of problems. It is unlikely that these algorithms are trained in languages such as Pashto and Dari in the same way as English and other Western languages. This increases the likelihood of errors, both in the form of false positives and negatives.

[3] Sheera Frenkel and Ben Decker, "Taliban Ramp Up on Social Media, Defying Bans by the Platforms," The New York Times, August 20, 2021, https://www.nytimes.com/2021/08/18/technology/taliban-social-media-bans.html.

[4] Michael Kugelman, "Avalanche of Misinformation Follows Taliban Takeover," Foreign Policy, September 9, 2021, https://foreignpolicy.com/2021/09/09/afghanistan-misinformation-taliban-takeover-social-media/.

Research Findings

Content moderation is described as "*the detection of, assessment of, and interventions taken on content or behaviour deemed unacceptable by platforms or other information intermediaries, including the rules they impose, the human labour and technologies required, and the institutional mechanisms of adjudication, enforcement, and appeal that support it*."[5] For this research, data was collected primarily from Facebook through the tool Crowdtangle to analyze the nature of misinformation/disinformation and hate speech in Pashto and Dari languages spoken in Khyber Pakhtunkhwa and Afghanistan. The researchers scraped through hundreds of content on the platform for classification purposes. The researchers monitored 8 pages of traditional media and 21 pages from non-traditional media/influencers posting in Pashto and Dari to identify cases of misinformation and hate speech over a period of three months. Furthermore, a list of 35 keywords was also developed to help monitor any viral content that was reviewed by the researchers.

From the scraped data, the researchers shortlisted 14 pieces of content from traditional media pages, 25 from pages categorized as non-traditional media/influencers, 10 pieces of content that went viral, and 7 tweets for the purposes of this research. The data has been categorized in the following tables:

[5] Tarleton Gillespie and Patricia Aufderheide. "Expanding the debate about content moderation: scholarly research agendas for the coming policy debates," Volume 9, Issue 4, Internet Policy Review, https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy.

Traditional Media Pages	
Misinformation	5
Disinformation	3
Hate speech	2
Other Harmful Content	4

Non-legacy/Influencers	
Misinformation	13
Disinformation	1
Hate speech	1
Other Harmful Content	11

Viral Content	
Misinformation	2
Other Harmful Content	8

Tweets	
Misinformation	2
Disinformation	1
Other Harmful Content	4

In addition to the data collected through Crowdtangle, information was collected through a focus group discussion (FGD) with 7 local digital media experts and stakeholders to validate the initial findings and garner additional context. The FGD included: Riaz Ghafoor - Assistant Director, Information Department, Khyber Pakhtunkhwa, Danish Yousafzai - Deputy Director, Information Department, Dr. Ameer Hamza - Asst Professor, Journalism Department, University of Peshawar, Sajid Takar - Editor, Pakhtun Magazine, Murtaza Hussain - Director Cyber Emergency Response Khyber Pakhtunkhwa (KPITB), as well as Tayyab Afridi and Iftikhar Khan of Tribal News Network. This FGD took place in Peshawar towards the end of the study.

Limitations of Research

The research's focus on Facebook – to the exclusion of other widely used platforms in the country as well as private-messaging platforms such as WhatsApp – is a limitation of its scope. It is recognised that private-messaging platforms widely used across the region, such as WhatsApp, host large volumes of false information and harmful content, but are under-studied by researchers. While tweets were mined for the purpose of this study, given Twitter's restrictive APIs they could only be accessed for the previous 7 days. Facebook was chosen as the focus of this research for two primary reasons: 1) it is the largest social media platform in Pakistan, with 50 million users[6] and the most widely used platform in Afghanistan had 3.7 million users;[7] and 2) there was greater access to content on Facebook through the CrowdTangle tool which focused primarily on Facebook-related content.

Furthermore, content moderation happens in many ways: removal of content, suspension of accounts, de-platforming users, placing labels on content, determining content reach, change in platform content distribution infrastructure, and demonetization. However, not all of these are captured in the methodology used in this study. The study focuses on the nature of harmful content in the chosen regional languages.

Perhaps most importantly, language – the focus of the research – was perhaps the biggest limitation. Pashto and Dari, while widely spoken in the AfPak region, are often not written in a fixed structure. There is a lot of reliance on the oral tradition, which means that it lends itself to videos and pictures more often than to text posts. Extracting visual-based data from social media platforms is a lot more difficult as the OCR for each picture was not public. Furthermore, when it comes to these languages being written in roman form, as opposed to their original script, there is a lack of standardization which makes it difficult to filter through keywords because one word can have multiple variations. Moreover, understanding Pashto in written form is rare. Even the courses in universities, colleges and schools in Khyber Pakhtunkhwa in Pakistan are taught in English or Urdu. Pashto is often

[6] "Annual Report 2021," Pakistan Telecommunications Authority, https://www.pta.gov.pk/en/data-&-research/publications/annual-reports.
[7] Simon Kemp, "Digital 2022: Afghanistan," February 15, 2022, https://datareportal.com/reports/digital-2022-afghanistan.

confined to the oral medium – unless you are a student of Pashto literature; otherwise, it is difficult for the people of AfPak region to read and write Pashto. Additionally, research tools such as Crowdtangle are better suited for English language content since the AI relies on and is trained on English-based content.

The focus on written content was a further limitation since a lot of hate and misinformation material is shared in the form of videos, audio clips and, now, in Twitter space. These mediums are in some aspects more inclusive given that they do not require literacy as an entry point. Other emerging mediums are more ephemeral, for instance, stories which last from around 24 hours to 7 days depending on the platform. The data collected for this research does not capture this type of content.

Lack of digital literacy also means that using reactions to posts as data points, such as likes and reactions, might not be an accurate barometer for researchers. During the research, it was pointed out those reacting to a post might not understand what the reaction means and whether the emotion they attach to the reaction is the same as the researchers. Given that these platforms are increasingly becoming more complex and there is frequent tweaking, 'digitally illiterate' users can behave in ways online that might not be reflective of their intent.

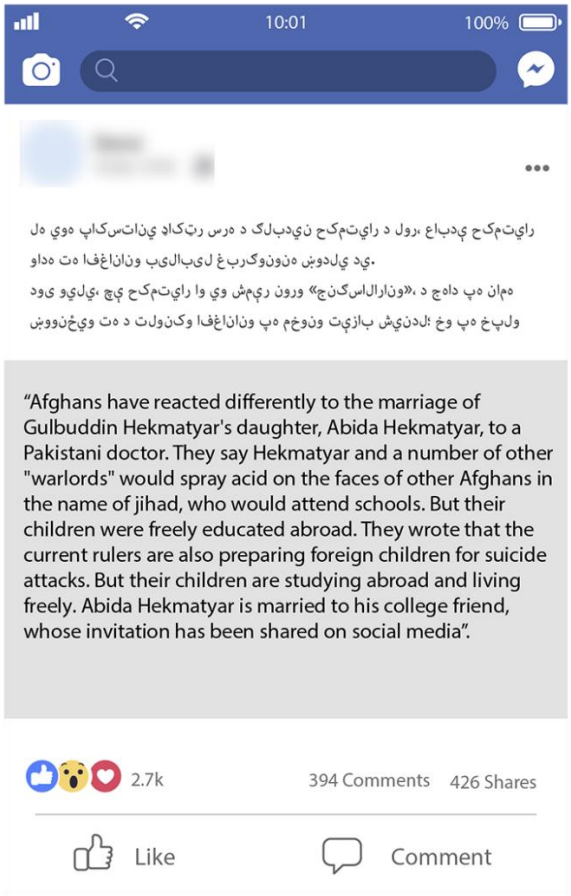
Lastly, the research was a time-consuming process as the Crowdtangle tool was good for filtering content, however, it required manual sorting of data and searches.

Case Studies in Pashto

#1: Wedding of the Daughter of Gulbadin Hikmatyar (Leader and founder of Hezb-e-Islami)

This fake news spread in no time and was viral in Afghanistan and the North-West region of Pakistan when some media outlets shared news claiming that the daughter of the current leader and founder of Hezb-e-Islami, a Political Party (Gulbadin Hikmatyar), was getting married to a Pakistani doctor.

The news surfaced on March 5, 2022 around 5:33 AM in the morning and was posted by one of the prominent news providers in the Afghan region called “ZAWIA News”. The daughter of Gulbadin Hikmatyar, an ex-commander of the Afghan Taliban, was believed to be studying MBBS in the Northwest Medical College (One of the Private Medical Colleges in Peshawar), the original post (in Pashto) stated the following:



بیلابیل له یوه پاکستاني ډاکټر سره د ګلبدین حکمتیار د لور، عابدې حکمتیار واده ته افغانانو غبرګونونه ښودلي دي.

دوی ویلي، چې حکمتیار او یو شمېر نورو «جنګسالارانو»، د جهاد په نامه ښوونځیو ته د تلونکو افغانانو په مخونو تېزاب شیندل؛ خو په خپلو اولادونو یې بهرنیو هېوادونو کې ازادانه زده کړې کولې.

دوی لیکلي، چې اوسني واکمنان هم پردي اولادونه ځانمرګو بریدونو ته چمتو کوي؛ خو خپل اولادونو یې په بهرنیو هېوادونو کې زده کړې او ازاد ژوند کوي.

عابدې حکمتیار د خپل کالج له ملګري سره واده کړی، چې بلنلیک یې په ټولنیزو رسنیو خپور شوی دی.

Translation:

> “Afghans have reacted differently to the marriage of Gulbuddin Hekmatyar's daughter, Abida Hekmatyar, to a Pakistani doctor.
> They say Hekmatyar and a number of other "warlords" would spray acid on the faces of other Afghans in the name of jihad, who would attend schools. But their children were freely educated abroad.
> They wrote that the current rulers are also preparing foreign children for suicide attacks. But their children are studying abroad and living freely.
> Abida Hekmatyar is married to his college friend, whose invitation has been shared on social media”.

The post started to spread as soon as it was posted and has been shared 426 times with 394 comments and 2.7k overall interactions. The post was even shared with a fake wedding card with the wedding date as March 10, 2022. The intended impact was to defame the Hezb-e-Islami and spread hatred against his political party. The target audience for the news ostensibly was the people of Afghanistan.

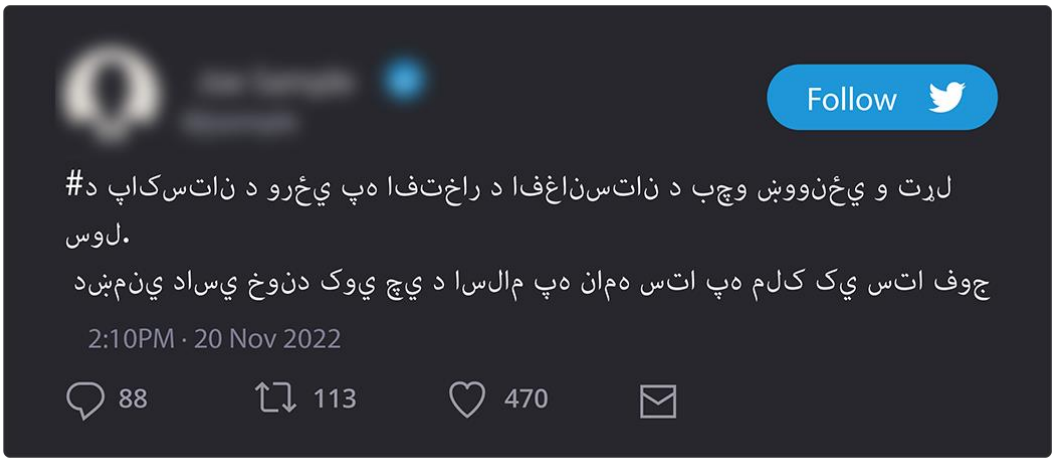
#2: The News of Pakistani Flag in Afghanistan

Posts claiming that the Pakistani Ambassador to Afghanistan Mansoor Ahmad Khan held a flag hoisting ceremony on March 23, 2022 in Kabul went viral on social media. The

ceremony was allegedly attended by members of the Pakistani embassy and Pakistanis working in Afghanistan. Mansoor Ahmad Khan shared photos of the flag-raising ceremony on his Twitter account. In the photo, Mansoor Ahmad Khan is seen unfurling the flag, and it is claimed that he is accompanied by two Afghan security personnel.

Afghans worldwide commented on what they believed was Afghanistan celebrated Pakistan Day on March 23. Someone said that Afghanistan was the “fifth province of Pakistan”. Another condemned the Taliban in harsh words that because of them the government the flag of Afghanistan was destroyed and the white flag which belongs to the Taliban was hoisted in the government offices of Afghanistan. The video became symbolic of the resentment many had against the Taliban government in Afghanistan, which is widely believed to be supported by Pakistan.

The truth of the matter was that the flag-raising ceremony took place in India and the video was shared from the Twitter Account of the Pakistani High Commission to India. It was posted from an account named “Mohammad Marhoon'' which stated that “On the day of Pakistan's Pride, the educational institutions of the children of Afghanistan were closed. The real enmity should be like this that in the name of Islam, on the day of Pride, your army is barred and then forced to do their official parade, your flag is put down and the flag of Prophet SAW (The Flag of Taliban) is moved up. Oh Lord! We are aidless but not this much”.



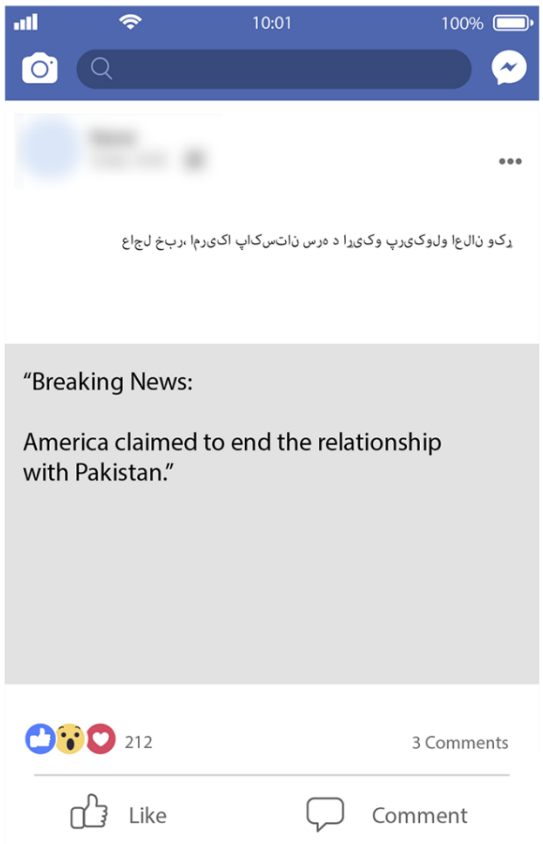
The original post is as follows:

#د پاکستان د ورځي په افتخار د افغانستان د بچو ښوونځي و ټرل سول.
دښمني داسي خوند کوي چې د اسلام په نامه ستا په ملک کي ستا فوج چپه کي خپل نظامي پرېښټ درته وکړي.
ستا ملي بيرغ کښته کي د پيغمر ص بيرغ پورته کړي او د خپل ملک ملي بيرغ هم درته پورته کي.
خدايه بې وسي دي و خو دونه هم نه

This particular post on Twitter got 88 Comments, 113 Retweets and 470 Likes.

#3: America to finish Diplomatic Relationship with Pakistan

Pakistan's historically complex relationship with the US has also been the target of propaganda and sensationalism. In this post, one of the media channels - "Meena TV" - posted news about Pakistan's relationship with the US saying the following:



عاجل خبر، امریکا پاکستان سره د اړیکو پریکولو اعلان وکړ

Translation:

"Breaking News: America claimed to end the relationship with Pakistan."

The news was posted on October 11, 2021 that got 3 comments, 212 overall interactions and 5.8k views posted through a Digital Media Page run by a local girl named "Meena"' from Afghanistan. The supposed motive for the post was to gain more audience for the page's content and views, perhaps to increase monetization. This news was intended to be shown to the people of both Pakistan and Afghanistan.

#4: Pakistani Military Facilitation Centers in Afghanistan

This fake news broke out on October 3, 2021 by a Social Media News page titled "Meena TV" when the Taliban had just taken control of Afghanistan after the American troops withdrew from the country.

The news stated that the Pakistan Army had built its facilitation centers in the Helmand Province of Afghanistan. This was classified as misinformation as no electronic or digital media outlet reported on it. The post appears to be disseminating hearsay to draw links between the Pakistan Army and the situation in Afghanistan. The original post says the following:

د پاکستان پوځ هلمند کی تاسیسات جوړ کړی تالیبانو څه ویلی

Translation:

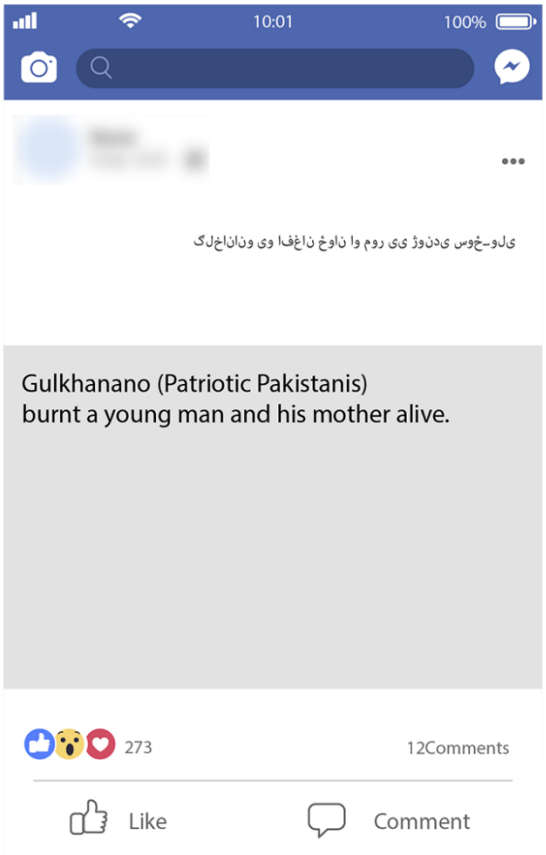
> “The Pakistan Army has set up facilitation Centers in Helmend, what Taliban has to say.”

This news went viral on Facebook and has been viewed 124k times with 146 Comments and 4.7k overall interactions.

#5: Mother Along with her Son Burnt Alive

A news item claiming that a mother was burnt alive along with her son propagated in Pakistan's Khyber Pakhtunkhwa went viral quite quickly. The alleged incident supposedly occurred in the Charsadda region of the Khyber Pakhtunkhwa province.

Furthermore, this news spread by "Meena TV" (a digital media platform) was altered even more to make it more contentious. Meanwhile, in reality, the news was that one of the residents of the Charsadda (believed to be Afghani) killed a young boy in a fight about pigeons. The family of the deceased gathered around the accuser's home (who was disabled) and set his home on fire where he was hiding along with his mother. He died in the fire and his mother was shifted immediately to the hospital to treat her injuries. Moreover, the video shown in this particular post is fake and had nothing to do with the news itself.



The original post has to say the following:
گلخانانو یو افغان ځوان او مور یې ژوندی سوځـولی

Translation:

Gulkhanano (Patriotic Pakistanis) burnt a young man and his mother alive.

This fake news got 8.3k views, 12 comments and 273 overall interactions, the rest of the information about this news can be gotten from the Traditional Media Section of the Master sheet.

Analysis

The data collected tells us that, in many instances, fake news is not always easily identifiable. As seen from some of the selected case studies, identification of misinformation often requires a nuanced understanding of the context, since misinformation is often built upon exaggerations rather than outright "lies". One focus group participant put in the following terms:

> "Fake news/misinformation usually [stems] from misunderstanding through rumors and hearsay in which the people mold the exact words of the speculation in their own way and in turn give them their own taste."

The researchers engaged in this study were able to pick up on these contextual cues because they were native language speakers and residing in the areas where the misinformation emerged from, thus embedded in the socio-political context.

Most of the themes identified in the collected content showed that a lot of the content in Pashto related to misinformation regarding the Taliban and Afghanistan-Pakistan relations. For Dari, a majority of the content related to hate speech along sectarian lines. Furthermore, it was noted that there was a substantial amount of misinformation regarding less political matters, such as rumors about the death of famous actors or singers. This content was not included in the dataset though it was more likely to go viral. It was also noted that there was an uptick in misinformation regarding drone attacks immediately after the US-backed regime in Afghanistan fell to the Taliban. There were also several false reports regarding the murder of journalists by the Taliban.[8] While journalists, activists, minorities and women are uniquely vulnerable at the hands of the Taliban, misinformation exploiting these fears was also fairly common.

[8] Reuters Fact Check, "Fact Check-Fabricated story about a journalist killed in Kabul," Reuters, August 18, 2021, https://www.reuters.com/article/factcheck-journalist-kabul/fact-check-fabricated-story-about-a-journalist-killed-in-kabul-idUSL1N2PP1W3.
Ciara O'Rourke, "Fake news accounts are spreading false information about journalists executed in Afghanistan," Poynter, August 18, 2021, https://www.poynter.org/fact-checking/2021/fake-news-accounts-are-spreading-false-information-about-journalists-executed-in-afghanistan/.

Out of the content shortlisted for the research, 26 posts were removed by the platform. Interestingly, most of the viral posts (all but one) were removed, whereas 8 out of 26 posts from non-traditional digital media pages were removed, and 9 out of 14 pieces of content were removed from traditional media outlets, during the length of the research period and at the time of editing this report.[9] It appears from the findings that viral content classified as hate speech or misinformation was most likely to be removed by the platform; perhaps its virality attracted greater scrutiny by automated systems or reporting. Furthermore, it appears that verified or traditional media outlets also faced more content moderation as opposed to non-traditional digital media pages or influencers. The inference drawn here is that violating content by non-verified outlets was more likely to stay on the platform. It seems automated systems are not adept at identifying and removing content in Dari and Pashto that violated policies regarding hate speech and false information.

In the sorting process, it was difficult to clearly delineate hate speech; it was only after analyzing the comments and overall content shared by a particular page that the intention and impact of the speech would become obvious. For instance, while criticism of Pakistani state institutions for their role and complicity in the crisis in Afghanistan was legitimate political speech, some posts occupied more gray areas. Posts using the popular hashtag #SanctionPakistan ranged from criticism and political speech to incitement to violence against the state. These are tricky calls to make for Meta, however, there seems to be no thoughtful policy by the platform on speech in this context. This is in stark contrast to the company's decision to allow violent speech on the platform in Ukraine:

[9] This report was edited in May 2022.

> *"As a result of the Russian invasion of Ukraine we have temporarily made allowances for forms of political expression that would normally violate our rules like violent speech such as 'death to the Russian invaders.' We still won't allow credible calls for violence against Russian civilians."*[10]

While there is appreciation for the fact that these are difficult judgement call, the fact that there was a complete absence of guidelines by the platform in the region, which could also be reasonably classified as a conflict zone, is concerning. This is not to say that social media companies did not put in work towards the crisis in Afghanistan, however, Meta's interventions were geared more towards digital security in the form of features that included removing the ability to search friend lists of users in Afghanistan and a one-click tool allowing users to lock their accounts.[11] It appears that security interventions by the platform were a lot more thoughtful as compared to the heavy-handed content moderation policies in place for the region.

These issues were validated by the decision by the Meta Oversight Board regarding an appeal against a content removal decision by the company on Facebook relating to the Taliban made by a user from Afghanistan. The post in Dari, written by a user identifying as a journalist, was removed under Meta's 'Dangerous Individuals and Organisations Community Standard'. The Board framed the questions and areas of focus for the case to include the following:

[10] "Facebook and Instagram let users call for death to Russian soldiers over Ukraine," The Guardian, March 11, 2022, https://www.theguardian.com/technology/2022/mar/11/facebook-and-instagram-let-users-call-for-death-to-russian-soldiers-over-ukraine.
[11] "Facebook moves to protect Afghan users' accounts amid Taliban takeover," August 20, 2021, BBC, https://www.bbc.com/news/technology-58277175.

> *"Content moderation challenges specific to Afghanistan and languages spoken in the country. How content moderation policies affect public discourse in Afghanistan, before and after the Taliban takeover. The safety of journalists and extent of media freedom in Afghanistan since the Taliban takeover, and how these factors affect reporting about the Taliban and the public's access to information on the political and security challenges facing the country. Whether Facebook's Dangerous Individuals and Organisations Community Standard unnecessarily limits discussion of designated groups that either form or take the place of governments. The relationship between US law prohibiting material support of designated terrorist organisations and Facebook's content policies, and how this may affect freedom of expression globally."*[12]

In its judgment released in September 2022, the Oversight Board overturned Meta's original decision to remove the Facebook post and found that it failed to protect users' freedom of expression to report on terrorist regimes. While the original post was in Urdu, the Board recommended that there be an increase the capacity allocated to high-impact false positive override (HIPO) system review across all languages.[13]

Content moderation consists of much more than simply community guidelines and formal systems in place at platforms. There needs to be increased recognition that content moderation is more than the sum of its parts – it consists of intangible factors such as the cultural understanding of content moderators, commercial interests of tech companies, lobbying and power of government actors, and access of civil society and activists to these platforms. Content moderation is as much a political issue as it is legal or technological. Once we recognize a more expansive understanding of content moderation, that "the policy ecosystem that affects how companies design, implement, and enforce their

[12] "Announcing the Board's next cases and changes to our bylaws," Oversight Board, November 2021, https://oversightboard.com/news/3138595203129126-announcing-the-board-s-next-cases-and-changes-to-our-bylaws/.

[13] "Oversight Board overturns Meta's original decision in "Mention of the Taliban in news reporting" (2022-005-FB-UA)," Oversight Board, September 2022, https://www.oversightboard.com/news/484790580170915-oversight-board-overturns-meta-s-original-decision-in-mention-of-the-taliban-in-news-reporting-2022-005-fb-ua/.

content standards has become increasingly complex," we will be able to have more honest conversations about the subject.[14] The geopolitical status of content, the identity of the users impacted by the content and the priorities of private tech companies and Western governments are the unwritten content moderation rules in many, if not all, of these cases.

Content moderation policies, encapsulated in community guidelines and the design of automated systems, are geared towards companies largely located in the United States and carry with them a very particular sensibility around what constitutes permissible content and free speech. While companies claim conformity with "international" human rights standards, these standards have a definite "Western" tilt. This faux-neutrality means that content moderation for users situated in all parts of the world must be regulated under largely Global North sensibilities. This creates a disconnect between the regulatory mechanisms for content moderation and the actual content, which represents the stories, lives and experiences of people across the world, mediated through complex systems designed elsewhere and coded with the values of a limited population.

Lastly, calls for contextual content moderation should not be conflated with demands by nation-states to impose local laws on private companies, particularly repressive laws designed to limit free speech. Social media companies often present a false binary between flawed community standards and repressive state laws as a way to avoid accountability and avoid employing greater resources to the Global South. This policy paper calls for greater nuance that centers on users and the impact on marginalized communities as opposed to abstracted binaries constructed by big tech.

[14] Robert Gorwa, "Content Moderation as a Regulatory Politics," https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy.

Conclusion & Recommendations

Content moderation at scale is an extremely complicated issue, however by looking at specific examples such as the case studies and data highlighted in this study, the conversation can start to take into account more diverse experiences and context that is normally overlooked. Emerging from these experiences are recommendations for reform and structural change reflected in focus group discussions and demands by activists in the region, some of which are reproduced below.

1. Over-reliance on automated systems should be revised in light of issues emerging from non-English speaking markets. The failure of these systems to adequately account for context should be reason enough to fundamentally revise systems and protocols underpinning them.
2. Dedicating more resources to human-based content moderation in non-Western contexts. The disparity of material resources between countries considered "key economies" and the "rest of the world" is startling and has resulted in enormous challenges for societies and political structures elsewhere.
    a. Furthermore it should be noted that increased human moderation should not mean outsourcing practices which have resulted in abject labor practices for human moderators and detrimental impact on the mental well-being of moderators.[15]
    b. There must be greater safeguards to ensure the safety and well-being of content moderators, i.e., living wages, humane workloads and well-being resources to prevent traumatisation.
    c. There needs to recognition for the unpaid labor by activists and individual users who are often burdened with the task of flagging, reporting and escalating content to fill in the gaps created by inadequate automated systems.

[15] Paul M. Barrett, "Who Moderates the Social Media Giants? A Call to End Outsourcing," NYU Center for Business and Human Rights, June 2020, https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf.

3. Radical transparency by tech platforms regarding the ways in which content moderation policies are formulated and implemented should be high on the priority of digital platforms.
    a. The current transparency practices by tech platforms are inadequate as they provide very little specific information. Transparency must include open information regarding the extent of cooperation and collaborations with governments, beyond statistics regarding take-downs; and must include data on all reporting and the nature of action or inaction taken particularly for content related to misinformation, disinformation, hate speech and online harassment as well as reasons for action or inaction.
    b. Furthermore, transparency must also be practiced at the individual user level, providing users with complete information, including reasoning, regarding decisions and the process of decision-making, i.e. whether the decision was made by human intervention or automated systems, pertaining to their content.
    c. Platforms should be pushed towards complete algorithmic transparency by making these algorithms public and open to public scrutiny as a way to off-setting the imbalance between platforms and users.
4. Content moderation decisions are often one-sided, with little recource for users who are aggrieved by the decisions, both for false positives or inaction by platforms. Meta's Oversight Board is a positive start but the model only impacts select cases. There needs to be a robust and time-responsive system for appeals that provides users with complete information regarding content moderation decisions and responsive action on appeals.
5. Content moderation decisions by tech platforms, and inaction in equal measure, have resulted in tangible real-world harms in the past and present.
    a. In order to course-correct the structural problems that have resulted, there must be meaningful accountability of big tech platforms through voluntary sharing of information for human rights audits as well as compensatory measures to repair the harm that has been caused. True accountability means truth-telling and accountability towards those harmed.[16]

[16] At the time of writing this paper, the Digital Services Act that includes some of these systems has been enacted in the EU however does not apply to regions which are the focus of this paper.

6. The importance of greater user awareness and empowerment cannot be overstated. Complex community guidelines and reporting mechanisms serve internal institutional content moderation needs but are often not accessible to the average user.
    a. Simplifying the reporting process by orienting them towards user-friendly systems is key to ensuring victim-led content moderation.
    b. Allocation of resources for user awareness that caters to different languages and user literacies is the need of the hour.
7. Fact-checking, required for counter-speech and fact-check labels introduced by platforms, for non-English languages is extremely inadequate. In the study outlined above, there were very few resources available for fact-checking available in Pashto and Dari resulting in mis/disinformation being disseminated uncontested.
8. Lastly, measures and advocacy for moving towards pro-competition approaches to content moderation challenging the concentration of power by big tech companies should be a collective priority. Proposals such as Article 19's recommendation of unbundling hosting and content-curation services on large social-media platforms need to be seriously considered.[17] DRF's recommendations to the Islamabad High Court on content moderation models can be accessed here.

[17] "Taming Big Tech," Article 19, 2021, https://www.article19.org/wp-content/uploads/2021/12/Taming-big-tech_FINAL_8-Dec-1.pdf.