Centre for
Information
Resilience

Afghan Witness

Part II: Quantitative investigation

Technology-facilitated gender-based violence (TFGBV) targeting politically engaged Afghan women.

September 2023

1 Executive Summary

This quantitative analysis seeks to complement the qualitative research set out in Part I by providing a deeper understanding of the gender-based violence that targets politically engaged Afghan women online. This report examines the following topics and questions:

1. **The scope and scale of gendered hate speech in the Dari/Farsi and Pashto information environment that women who speak these languages are exposed to.** What is the scale of hate speech that Dari/Farsi and Pashto speaking women face on the social media platform X (formerly known as Twitter)? How do common trends from June - December 2021 compare to common trends from June - December 2022?
2. **The scope and scale of gendered hate speech and abuse targeting politically engaged Afghan women**. What common trends can be seen from the second half of 2021, when the Taliban took over Afghanistan and the second half of 2022? What event(s) caused a spike in gendered hate speech?
3. **The nature of the abuse targeting politically engaged Afghan women**. What are the most commonly used hate speech terms against politically engaged Afghan women? Has there been an increase in the use of these terms between the second half of 2021 and the second half of 2022?

The findings in this investigation are based on a quantitative analysis of over 78,000 tweets/posts targeting politically engaged Afghan women from June 1, 2021 - December 31, 2021 and from June 1, 2022 - December 31, 2022. Due to X (formerly Twitter)'s changing environment, the original scope of the data collection has been narrowed down.

1.1 Key Findings

Research Question One: Scope and scale of gendered hate speech in the Dari/Farsi and Pashto information environment

- Overall, the volume of Dari/Farsi and Pashto gendered hate speech increased significantly in 2022 as compared to 2021. This finding remains true even when corrected by filtering out event-triggered hate speech, such as after the protests broke out in Iran in September 2022 following the death of Mahsa Amini, a young woman who collapsed in custody and later died after being detained by Iran's morality police. Although a lot of the hate speech and abuse might still not target Afghan women specifically per se, it negatively impacts perceived safety and freedom of Dari and/or Pashto speaking women online.

Research Question Two: Scope and scale of targeted hate speech and abuse

- AW recorded an increase of **217%** in posts containing gendered hate speech and abuse terms and the names of prominent Afghan women from the period June - December 2021 to June - December 2022. This substantial increase could be related to developments in the region; X (formerly Twitter)'s changing environment; and a generally more hostile environment towards women.
- **Before and during the Taliban takeover,** spikes in gendered hate speech could be connected to major advancements the Taliban were making in the country.

- Spikes in gendered hate speech during the **second half of 2021** and the **second half of 2022** were usually connected to policies and bans that the Taliban imposed on women, ultimately restricting their rights and freedoms. The hate speech was mainly directed at the Afghan women who protested against these policies and bans.

Research Question Three: Nature of hate speech and abuse

- Hate speech and abuse directed at Afghan women was overwhelmingly sexualised. Over **60%** of the posts in 2022 contained sexualised terms used to target Afghan women. Overall, an **11.09%** increase in the proportion of sexualised terms targeting politically active Afghan women occurred from 2021 to 2022.
- It was not possible to conduct a quantitative analysis of the perpetrators of gendered hate speech. Qualitative analysis showed that this was an issue that spanned the Afghan political spectrum, with hate speech and abuse originating from users from a range of political affiliations and ethnic backgrounds.

The limitations of the two datasets, especially the focus on mentions of female Afghan influencers, underestimate the scale and severity of gendered hate speech and abuse.
The qualitative investigation carried out prior to this report combined social media analysis with six key informant interviews with politically engaged Afghan women. The main findings of the qualitative investigation are as follows:

1. **Afghan women's online presence and social media usage**. Post-Taliban takeover, there has been a rise in women's online advocacy accompanied by a rise in online abuse and harassment. AW noticed that the main platform used by Afghan women is X (formerly Twitter).
2. **The nature of TFGBV targeting politically engaged Afghan women**. Politically engaged Afghan women experience a wide range of abuse, including (but not limited to) sexual, gendered, religious, political and ethnic abuse. AW also noticed that gendered disinformation was spread against Afghan women to discredit and undermine them.
3. **Attribution**. Perpetrators of TFGBV against politically engaged Afghan women came from a range of political affiliations, ethnic groups and backgrounds, with low-ranking Taliban and pro-Taliban social media users being responsible for the majority of the abuse. AW also found that supporters of the National Resistance Forces (NRF) would also engage in online abuse.
4. **Impact**. TFGBV impacted the daily lives of the interviewees on a personal, societal and professional level, with interviewees stressing that the online and offline worlds are intertwined. TFGBV had a chilling effect on women's participation in online activities.

The findings of the quantitative investigation aim to complement the findings from the qualitative report, especially in relation to the quantity and nature of the abuse politically engaged Afghan women receive.

Table of Contents

2 Introduction

As explored in Part I of this report, social media platforms such as X (formerly Twitter) and Facebook have emerged as increasingly important platforms for Afghan women to create communities, campaign for women's rights and establish a degree of political involvement. In doing so, politically engaged Afghan women face various risks online, including gendered disinformation[1], abuse, harassment and hate speech.

In Part I of the report, Afghan Witness (AW) used the term technology-facilitated gender-based violence (TFGBV) to describe the range of acts conducted online which cause harm to women (be it physical, psychological, social, economic or political) and infringes on their fundamental rights and dignity. In this quantitative investigation, AW focused on a subset of wider TFGBV by looking at gendered hate speech and abuse.

This investigation seeks to complement Part I of the report by quantifying (to the extent possible) and deepening the understanding of the risks politically engaged Afghan women face on social media platforms. It is based on two datasets created from X (formerly Twitter). Dataset 1 contains posts with specific hate speech and abuse terms that are used to target Afghan women and the names of politically engaged female Afghan influencers. Dataset 2 contains posts with hate speech and abuse terms that explicitly mention Afghan female influencers[2]. With these two datasets of over 78,000 tweets from June to December 2021 and from June to December 2022, AW was able to present insights on:

- Trends and common patterns in the volume of gendered hate speech and abuse in the Dari/Farsi and Pashto information environment to which Afghan women are likely to be exposed: the online information environment is delimited by language rather than geographic boundaries and potential impact should thus be assessed on that level.
- Trends and common patterns in the volume of gendered hate speech and abuse targeting Afghan women comparing the second half of 2021 and the second half of 2022.
- The context behind these trends, focusing on spikes in the volume of gendered hate speech and abuse across both datasets and how these compare to other events on the ground and online.
- The nature of gendered hate speech and abuse, focusing on the language and terms most used to target politically engaged Afghan women.

[1] In Part I of this report, AW noticed that gendered disinformation is used to discredit and undermine politically engaged Afghan women, with perpetrators spreading false or inflammatory information about the women, discrediting the sources of their content, and creating fake accounts and pages using women's names and faces.
[2] In this report, the term "mention/s" refers to the explicit act of citing the handle of a social media user rather than just the name.

3 Methodology

Afghan Witness (AW) conducted a quantitative investigation based on the collection of over 78,000 tweets/posts targeting politically engaged Afghan women between June and December 2021 (pre- and immediately post Taliban takeover) and June and December 2022.

Tweets were collected in a large database through the use of Boolean strings and a relevancy algorithm. This large dataset generated two datasets with two queries within the same timeframe. Dataset 1, also referred to as the “General Dataset” (with over 73,000 tweets/posts), was created using an extensive range of Boolean strings determined and validated by Afghan analysts on the team. The Boolean strings consisted of both specific terms used to abuse Afghan women in both Dari/Farsi and Pashto and the names of politically engaged female Afghan influencers. Specific filters were added to the datasets to minimise content from other regions and/or countries. Dataset 2 has over 4000 posts (tweets). It contains the same abuse terms, but is limited only to those posts which explicitly mention politically engaged Afghan female influencers online. For this reason, it is also referred to as the “Influencer Dataset”. The image below shows a representation of how the data was collected.

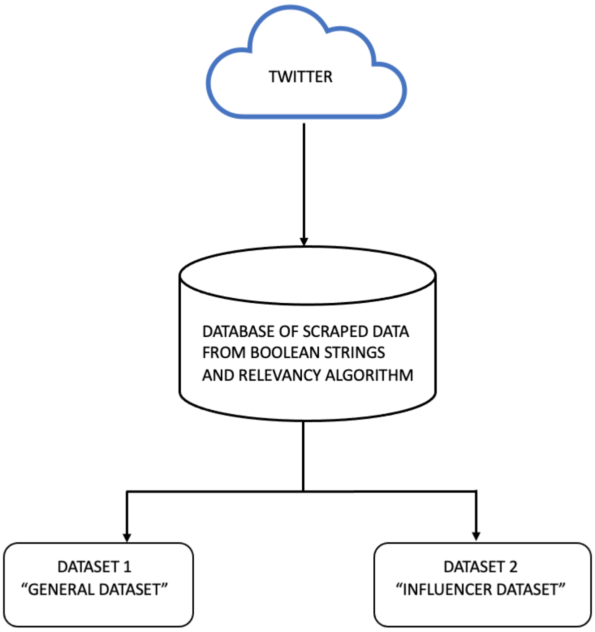
TWITTER
DATABASE OF SCRAPED DATA FROM BOOLEAN STRINGS AND RELEVANCY ALGORITHM
DATASET 1 “GENERAL DATASET”
DATASET 2 “INFLUENCER DATASET”

Figure: Data was scraped from X (formerly Twitter) using a number of Boolean strings and checked with the relevancy algorithm. AW then queried the data to create two datasets. Dataset 1 contains a more generalised dataset of tweets/posts that contain terms targeting women in Dari/Farsi and Pashto, as well as the names of politically engaged female Afghan influencers. Dataset 2 contains all tweets/posts mentioning female Afghan influencers only.

3.1 Research Scope and Definitions

Consistent with the definitions in Part 1:

- ‘Afghan women’ refers to women living in Afghanistan and Afghan women living in the diaspora.

- ‘Politically engaged’ encompasses women politicians, activists and prominent women who, for simply excelling in their field and receiving public recognition for it - may be viewed as engaging in women’s rights activism and/or politics[3].

- Throughout this quantitative investigation, AW has analysed “spikes” related to gendered hate speech and abuse targeting politically engaged Afghan women. In this investigation, “spikes” refer to a 50%+ increase in data compared to the general average/trend.

- Technology-facilitated gender-based violence (TFGBV) is violence committed and amplified through information and communications, technologies or digital spaces against a person based on gender (UNFPA). The analysis considers a wide range of online violence experienced by Afghan women - from sexualised language to disinformation, doxxing, and death threats – and, where possible, includes issues of intersectionality, such as ethnicity. As mentioned in the introduction, this quantitative investigation focuses on a subset of TFGBV, namely gendered hate speech and abuse.

3.2 Online gendered hate speech and abuse: scope and scale

Data collected from the second half of each year is demonstrated through graphs showing relative spikes. Dataset 1 shows an increase in the volume of general gendered hate speech and abuse, as measured through Boolean strings, in the wider Dari/Farsi and Pashto information environment. Dataset 2 shows an increase in the volume of hate speech keywords targeted directly at selected female Afghan influencers.

The largest spikes within the targeted dataset were then singled out, and AW carried out further research to determine what might be driving the spark in gendered hate speech relating to events in Afghanistan and internationally. Examples of the tweets/posts shared during that particular spike were also added. Data and spikes from after September 2022 were heavily focused on developments in Iran following the death of Mahsa Amini. Although these developments are relevant for evaluation of the respective information environment, due to the scope of this respective report, AW decided to apply filters to try to correct for these event-triggered spikes.

[3] This definition of politically engaged seeks to include the female Afghan influencers mentioned in the datasets. The Afghan influencers included in the datasets are all public facing women who engage in women’s rights activism in some form or another.

3.3 Nature of abuse

AW analysts used word cloud and graphical analysis to identify and understand the most common keywords used in gendered hate speech against politically engaged Afghan women. This analytical process is described in more detail below. For this chapter, the graphs have been created from Dataset 2 as it provides a more accurate representation of the hate speech directed at politically engaged Afghan women.

3.4 Data Collection

Data was collected from X (formerly Twitter) using an open-source scraping tool called *snscrape* that can be used to scrape social networking services (SNS). AW used the tool to automate data collection from X (formerly Twitter) following the terms of service of X (formerly Twitter) scraping. Multiple Boolean search terms were used to query X (formerly Twitter) to create the datasets from June to December 2021 and June to December 2022. AW's initial objective was to compile a comprehensive dataset covering the entirety of 2021, 2022, and the first half of 2023. However, due to restrictions and the unpredictable nature of access to X (formerly Twitter), AW was forced to refine the report's focus to the period from June 1 to December 31 for both 2021 and 2022. (see data limitations for further detail).

To ensure that the data AW collected was relevant to the Afghan information space, AW further refined the scraped data using a relevancy algorithm and refined the search terms once the data was collected for Dataset 1. This adjustment was necessary to address the issue of data overlap with Iran. By excluding the names of Iranian influencers and cities and making the language specific to Afghanistan, we aimed to eliminate this crossover. AW's relevancy algorithm is designed to assess whether the body of a tweet/post contains a sufficient number of predetermined terms, thereby ensuring its relevance to the study. This was implemented to decrease the amount of noise in the dataset and the requirement for data storage. The algorithm checks if the body of the tweet/post matches with the keywords defined by AW, then calculates an average based on the number of keywords to create a relevancy score.

No Match

The relevance system uses a mathematical equation to assess a post's relevance for collection. This equation is based on a comprehensive list of keywords. The calculation calculates the average usage of these keywords to determine the post's relevance to analysts. In other words, the higher the number of matching keywords, the higher the average relevance score.

When the term "No Match" appears in the collected data, it signifies that the script has identified the post as relevant based on the keywords. However, it couldn't identify a specific keyword with the strongest match from the lists. This might occur due to translation issues when matching English keywords against Arabic-style lettering and right-to-left sequencing. This limitation arises from the coding language regex's inability to fully recognise certain Arabic languages as text.

The Python scripting mandates that a post must achieve a relevance score above 5. This threshold was established after weeks of testing to ensure optimal accuracy. Posts with a score below 5 were not collected. It's important to note that any post collected and featuring "No

Match" still remained relevant to analysts and the project. These posts needed a score of 5 for collection, but their keywords couldn't be precisely matched due to the Arabic style lettering.

Data Analysis: Word Clouds

Word clouds were generated for the influencer datasets using the Wordcloud and NLTK Python packages. Due to the lack of support for Dari/Farsi and Pashto in the packages being used, AW used the translated versions of the tweets/posts and not the original text. The text was cleaned using a standard set of procedures for analysing X (formerly Twitter) data; this includes:

- Removing contractions through contraction expansion (“won’t” becomes “will not” etc);
- Removing stopwords, commonly used words that don’t convey meaning but are required in a sentence, for example, “the”, “this”, or “and”;
- Removing emoticons and symbols;
- Removing URLs;
- Removing X (formerly Twitter) mentions;
- Removing strings that are less than two characters to remove trailing letters;
- Finally, setting all of the text to lowercase so that the same word is not counted twice because it was in a different case (“That” and “that” would be counted separately).

3.5 Research Limitations

The findings in this report should be regarded as a sample of the wider patterns of abuse facing politically engaged women in Afghanistan. The findings presented here are likely to underestimate the scale and severity of the problem. This is a result of:

- **Scope.** The initial scope of the project was to collect a dataset for the years 2021, 2022 and early 2023, consisting of tweets/posts containing gendered hate speech towards Afghan women. However, due to several limitations, discussed further below, AW was unable to deliver a full dataset for those years. Using the data we were able to collect, AW chose to examine two comparable periods of the latter halves of 2021 and 2022.

- **Data collection on other platforms**. Part I of this report included an analysis of posts collected from Facebook. AW could not gather data from Facebook for this quantitative investigation due to the platform's restrictions on collecting large amounts of data. For this reason, data for the quantitative investigation was only taken from X (formerly Twitter).

- **Data limitations and X (formerly Twitter) takedowns**. Only information publicly available without authentication was used for this investigation. Since Elon Musk’s takeover, X (formerly Twitter) as a platform has become increasingly hostile towards data scraping, regardless of whether the official API or a third-party scraping tool is being used. X (formerly Twitter) completely shut off access to data to researchers and academics unable to pay for full API access from 30 June 2023. Academic access has been removed as an option, meaning that only the paid version of the API is available, which was unaffordable for this project. AW was able to collect part of the data but was forced to reduce the scope of the dataset due to the rules that X (formerly Twitter) imposed. Issues affecting the open-source scraper were reported in March, April and June 2023. As of June 2023, access to X (formerly Twitter) to view posts requires authenticated access which is outside the scope of the project for snscraper and will not be added. For this investigation, AW could not access or collect private messages, which were instead analysed in Part I of this report.

- **Content takedowns**. While Afghan Witness found many examples of abuse targeting politically engaged women remaining live months after they were posted, some publicly available abuse, particularly the most egregious, will likely have been removed by platform moderators before researchers could archive and analyse it. In addition, some of the posts AW analysed were comments containing gendered hate speech in response to tweets/posts which had been taken down. In these cases, it was not possible to understand the full nature and impact of the tweet/post. In addition, with Musk's takeover of X (formerly Twitter), the platform might not be taking down tweets/posts that violate the company's policies but only making them harder to find. Therefore, it could be possible that more content was taken down in 2021 (before Musk took over) than in 2022.

- **Language.** Some of the data analysis was completed using Natural Language Processing (NLP); however, neither Dari/Farsi nor Pashto are supported in the widely available NLP packages such as NLTK or spaCy. Due to this, word clouds were created using the translated version of the tweets and not the original text.

- **Hate speech targeting women from different ethnic backgrounds.** AW was not able to quantify the amount of ethnically-targeted hate speech due to the size of the datasets. However, in Part I of the report, the qualitative investigation was able to shed some light on the type of abuse that ethnic minorities, particularly Hazara women, receive.

4 Online gendered hate speech in the Afghan information environment: scope and scale

4.1 Overview

Analysis of the datasets have shown the following:

- **The total volume of identified gendered hate speech in Dari/Farsi and Pashto increased significantly between June - December 2021 and June - December 2022.** Although this hate speech might not have been targeted at Afghan women per se, the data illustrates the increasingly hostile information environment that Dari/Farsi and Pashto speaking women are forced to navigate. Unfortunately, it is difficult to pinpoint the catalyzer for the increase, although it is evident that it cannot be dedicated to developments in Afghanistan alone. Nonetheless, the findings remain true even when corrected for event-triggered developments, such as the Iranian protests following the death of Mahsa Amini in September 2022.
- **In the three months prior to and during the Taliban takeover** (June - August 2021), spikes in gendered hate speech were correlated with major advancements the Taliban were making in the country.
- **After the Taliban took over Afghanistan** (August 2021) and the United States left, the volume of hate speech appears to have increased, and spikes appear to have become slightly more frequent.
- Spikes in the volume of gendered hate speech and abuse were usually in relation to the **indoor and outdoor protests** that Afghan women carried out against policies and bans that the Taliban imposed on women, ultimately restricting their rights and freedom.
- In the **second half of 2022**, AW noticed larger and more frequent spikes in gendered hate speech and abuse against politically engaged Afghan women. Overall, AW reported an increase of 217% in posts containing gendered hate speech and abuse from June - December 2021 to June - December 2022.

4.2 Data from June to December 2021

The graph below demonstrates the volume of gendered hate speech and abuse from both datasets between June and December 2021. While Dataset 1 (dark blue) shows a relatively consistent baseline of gendered hate speech and abuse all throughout 2021, Dataset 2 (light blue) presents significant increases in hate speech against female Afghan influencers during certain months of the year, namely August and September 2021.

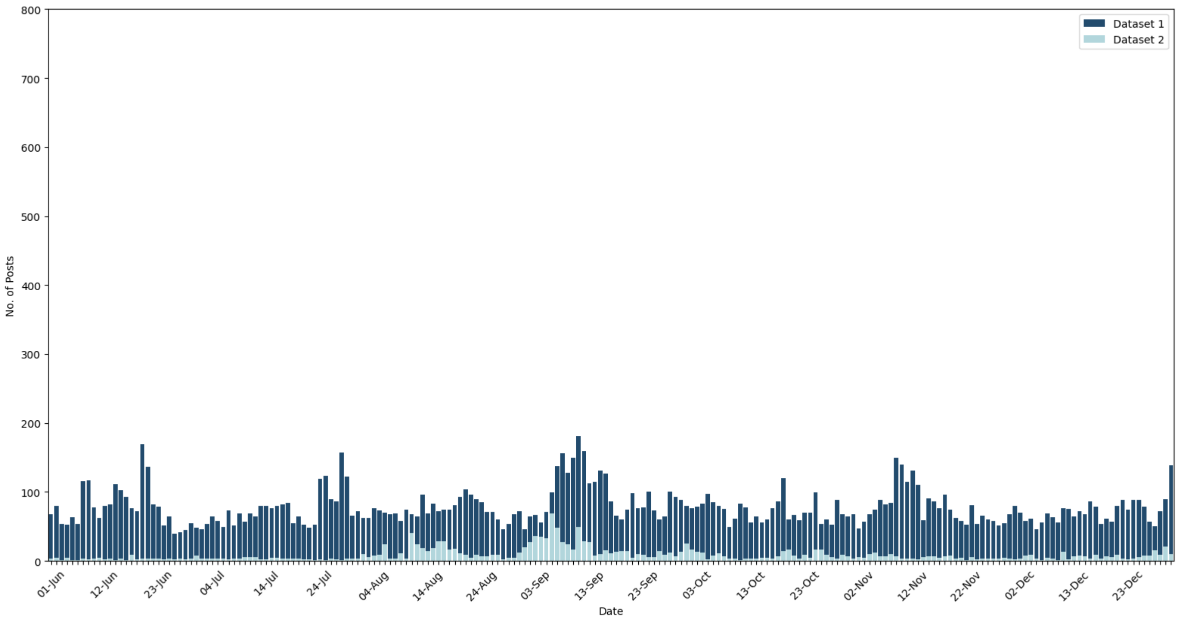

Figure: graph showing the overlapping datasets for data collected from June to December 2021. Bars in the graph are overlaid, not stacked.

From the graph, AW analysts noticed the following:

- Spikes **before and during the Taliban takeover** (June - August 2021) could be related to Taliban advancements in the country. Spikes around mid-June are correlated with Taliban advancement in several provinces, including Faryab and Zabul. Around July 25, the spike could be connected to an announcement from the US, vowing to support Afghan troops against Taliban forces. Lastly, spikes in August 2021 correlate with the Taliban officially taking over Afghanistan around August 15 and 16, 2021.
- Spikes between **September and December 2021** are correlated with a series of protests and marches by women in support of women's rights that occurred in several provinces. As documented in Part 1 of the report, Afghan women took to the streets in cities such as Kabul, Balkh and Herat to protest the Taliban and their restrictive measures. Afghan women protested for their right to work, their right to an education and their right to social welfare.

A qualitative deep-dive into the largest spike on September 6 and 7 provided further evidence of the link between women's political activism in the form of the protest movement and gendered hate speech and abuse.

4.2.1 Spike: September 6-7, 2021

With 69 posts in Dataset 2 and 300 posts from Dataset 1, the most prominent spike after the Taliban takeover was on September 6, 2021. The graph below, taken from Dataset 2, captures the spike more accurately, showing the number of posts during the month of September 2021.

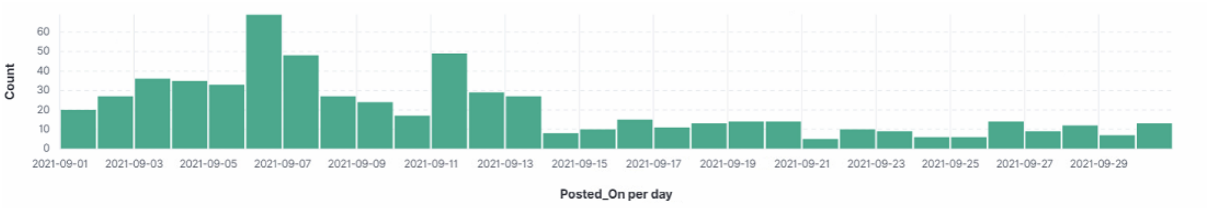

Figure: graph showing the spikes in Dataset 2 during the month of September 2021.

The first week of September 2021 saw a wave of female-led protests. On September 2 and 3, 2021, Afghan women protested for their rights in Herat and Kabul. On September 6, 2021, a group of Afghan women took to the streets in Mazar-i Sharif, Balkh, to protest the Taliban and their violation of women's rights and advocate for female political participation. Several videos surfaced on X (formerly Twitter) showing the Taliban trying to disrupt the protests with vehicles. AW wrote a report on the protests and the measures used by the Taliban to suppress them.

The protests triggered widespread reactions on social media. While many users commended the women for their courage in protesting for their rights, the protests and the women participating also faced online hate. Qualitative analysis of these posts shows that the majority came from what appeared to be low-ranking Taliban and pro-Taliban attributed accounts. AW identified and investigated several disinformation campaigns targeting women protesters in its 2022 report.

AW analysts also found the spike to be related to clashes that occurred around September 7, 2021, between the Taliban and the National Resistance Forces (NRF) in Panjshir as well as a march by men and women on the same day in Kabul, following a call from the NRF leader Ahmad Massoud for "a national uprising" against the Taliban. Among the dataset, AW researchers identified hateful comments and tweets targeting the women who took part in the protests and those who supported or denounced the protests online. AW analysts noticed that users who were pro-Taliban and anti-resistance would usually target women who were anti-Taliban and pro-resistance. The opposite would also occur, with anti-Taliban and pro-resistance users targeting women who were pro-Taliban and anti-resistance. This finding was also investigated in Part I of the report.

The majority of tweets/posts shared on September 7, 2021, in Dataset 2 contained explicit and highly sexualised language, often accompanied by gendered slurs. Many of these posts mentioned the influencers directly, labelling them as "whores" and "prostitutes." Tweets/posts from Dataset 1 denounced the women for disrupting society and would call for them to be punished by the Taliban.

4.3 Data from June to December 2022

Data collected from June to December 2022 shows a substantial increase in the volume of gendered hate speech and abuse from 2021 to 2022. The data collected also shows a gradual and steady increase in volume from June to December 2022. When comparing all the tweets/posts from both datasets in 2021 with all the tweets/posts from both datasets in 2022, AW analysts recorded a substantial increase in the volume of hate speech, with a 217% increase in posts from 2021 to 2022[4].

The graph below demonstrates the volume of gendered hate speech and abuse from both datasets between June and December 2022. AW analysts reported a substantial increase in posts containing gendered hate speech and abuse from Dataset 1, while Dataset 2 shows a steady increase all throughout the year.

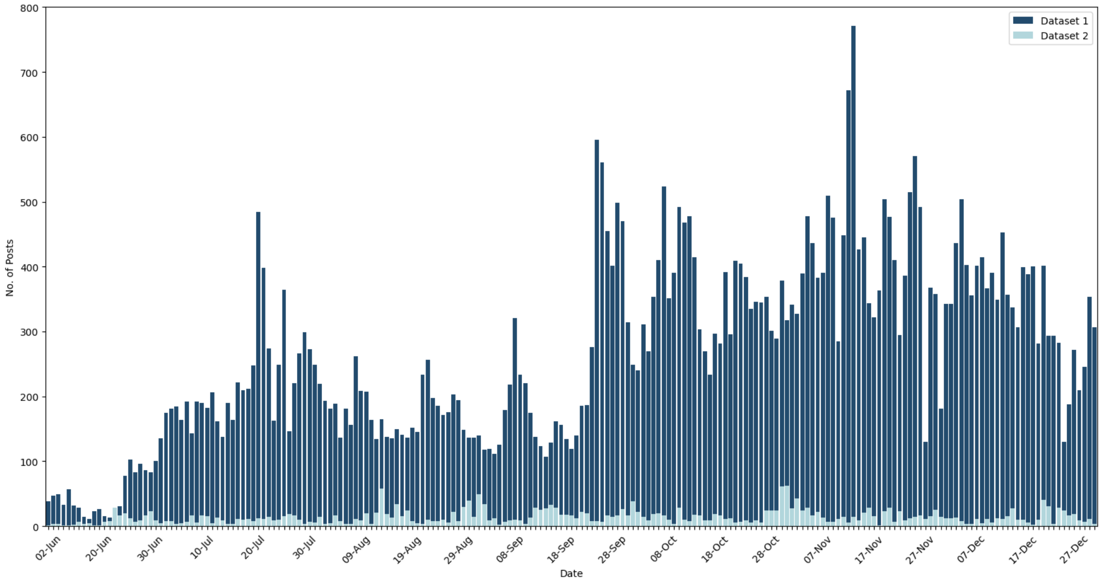

Figure: graph showing the overlapping datasets for data collected from June to December 2022. *Bars in the graph are overlaid, not stacked.*

From the graph, AW analysts noticed the following:

- Spikes between **July and August 2022** correlate with a series of bans the Taliban imposed on women. During this period, women were banned from entering health centres; female government employees were told to stay at home or asked to introduce a male relative to work in their place; women were ordered to cover their faces and bodies in public and forced to wear the hijab; the female moral police was established. All of these restrictions pushed Afghan women to protest against the Taliban.
- Spikes between **mid-September and December 2022** were characterised by bans imposed by the Taliban on both the educational and social spheres and women's protests against the restrictions. The Taliban banned women from universities, banned women and girls from parks and gyms, and banned women from working in NGOs. AW wrote a report on the protests happening during these months and how the Taliban disrupted them.
- **Between June - December 2022**, significant increases in Dari/Farsi and Pashto gendered hate speech dwarf Afghan targeted gendered hate speech trends, despite the latter also facing substantial growth as compared to the same period the year prior. Although it is difficult to identify the exact trigger behind the increase in Dari/Farsi and

[4] The increase was calculated by summing the number of total tweets/posts from both datasets in 2021 and in 2022. In 2021, the total number of tweets/posts was 18703, while in 2022 the total number of tweets/posts was 59358.

Pashto gendered hate in general, it is evident that the changes were not triggered by developments in Afghanistan alone: comparing spikes in Dataset 1 (dark blue) and Dataset 2 (light blue) indicates a negative correlation between the two. Some of the Dari/Farsi and Pashto gendered hate will have been triggered by the protests in Iran following the death of Mahsa Amini. However, filters applied to minimise results related to these events, as well as the fact that increase commenced prior to the start of the protests (in July 2022 instead of September 2022), suggest that there must be more factors at play.

- Other possible examples include X (formerly Twitter)'s changing environment. Although Musk became X (formerly Twitter)'s CEO in October 2022, talks about acquiring the social media company commenced from April 2022. With Musk promising to turn X (formerly Twitter) into a safe space where people can talk freely, X (formerly Twitter) users might have become more inclined to spread gendered hate speech and abuse against women. Researchers looking into other forms of hate speech, such as antisemitism, came out with similar findings. Hence, while it is likely that there are multiple factors that have contributed to the increase of gendered hate speech in the Dari/Farsi and Pashto information environment, AW has unfortunately not yet been able to find conclusive evidence for the exact cause.
- In addition, with Musk's takeover of X (formerly Twitter), the platform might not be taking down tweets/posts that violate the company's policies but only making them harder to find. Therefore, it could be possible that more content was reported or taken down in 2021 (before Musk took over) than in 2022.
- Although less exponential than compared to gendered hate speech in Dari/Farsi and Pashto in general, gendered hate speech targeting Afghan women also faced significant increases in June - December 2022 as compared to June - December 2021. Zooming in specifically on the targeted hate speech in Dataset 2 (light blue), AW identified a correlation between the spikes and a sequence of restrictions imposed by the Taliban on the rights and freedoms of women and girls. These restrictions prompted further protests by Afghan women, which in turn were violently repressed by the Taliban. Online activism in response to these developments is likely to have triggered increased gendered hate speech, conceivably by individuals who experience a greater sense of impunity since the Taliban takeover.

A prominent spike can be seen around October 31 and November 1, 2022. AW analysts carried out further research to determine the connection between the event and the spark in gendered hate speech (section below).

4.3.1 Spike: October 31 and November 1, 2022

The most significant spike of targeted hate speech, comprising over 62 posts in Dataset 2, occurred on November 1, 2022. The graph below, extracted from Dataset 2, provides a more precise visualisation of this spike, depicting the post count throughout the month of November 2022.

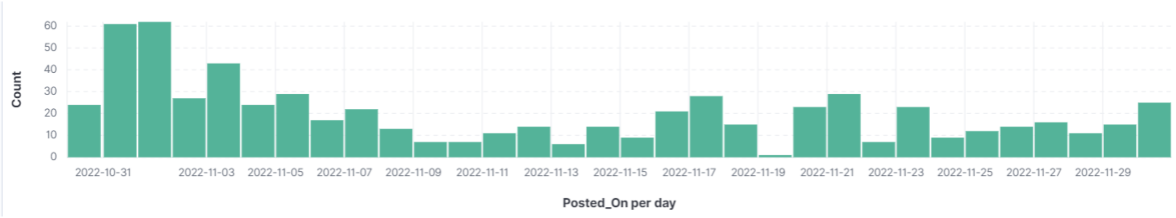

Figure: graph showing the spikes from Dataset 2 during the end of October and November 2022.

From the end of October and the start of November 2022, Afghan women staged several protests against the Taliban. The protests centred around the restrictions on women's education, access to public places and amenities, and freedom of movement and clothing.

On October 31, a group of women protesters gathered in a park in Kabul, showing their academic documents as a form of protest against the Taliban's constraints on women's rights to education and employment. However, the Taliban reportedly intervened and disrupted the gathering. AW analysts determined that the spike in posts during this period is also linked to the protest and the subsequent online reactions it sparked.

Most of the tweets/posts shared on October 31 included mentions of the influencers alongside highly explicit sexualised and gendered language, with most of the perpetrators calling the influencers "whores" and "prostitutes". Similarly, from Database 1, AW analysts noticed that in some tweets/posts, perpetrators called out women for not wearing a hijab, for betraying their religion, spying for the West, and attempting to get asylum cases approved in Western countries. AW covered these four trends extensively in Part I of the report.

The spike in posts observed on November 1, 2022, was primarily driven by online reactions to an incident involving Tamana Zaryab Paryani, a women's rights activist exiled in Germany. On October 30, 2022, Paryani posted a video on her X (formerly Twitter) account in which she publicly removed and burned her burqa. In the video, she discussed her past protests against mandatory hijab and chanted the slogan "*No to compulsory hijab.*" This video triggered a substantial response across social media platforms. Upon qualitative analysis of the dataset, it became evident that women who supported Paryani by sharing her video or retweeting/reposting supportive comments were targeted with hate speech.

5 Nature of the abuse

5.1 Overview

The common themes from the qualitative report were: sexual, gendered, religious, political and ethnic abuse, which were used to structure the keywords for the quantitative analysis. Data collected from the second half of 2021 and the second half of 2022 has provided an understanding of which themes are most common in posts targeting Afghan women. The main findings from this section are the following:

- From the datasets, AW analysed that the most common theme is sexualised abuse. From the second half of 2021 to the second half of 2022, there was an 11.09% increase in the proportion of sexualised terms targeting politically active Afghan women.
- The most commonly used words in hate speech targeting politically engaged Afghan women are: “dirty”, “prostitute”, and “whore”. In comparison to 2021, the use of the terms "whore" and "prostitute" increased by 3.56% and 6.56%, respectively, in 2022. The term "dirty" instead experienced a decline of 8.05% between the years.

5.2 Sexualised abuse

Based on AW’s qualitative report, sexualised abuse against Afghan women has been a common form of online abuse. Interviewees in the qualitative report emphasised the pervasive sexual and gendered abuse they received in comments to their posts on social media and in direct messages that sometimes contained pornographic content.

Qualitative analysis of the influencer dataset demonstrates that online abusers used sexual slurs against the Afghan influencers regardless of ethnicity, background, or political affiliations[5]. The slurs were also used against a number of male influencers, targeting them and their female family members. For instance, slurs such as “son of a whore” were frequently used in posts. The slurs were also combined with other words; for instance, “political prostitute” and “fugitive prostitute” were commonly used.

In both 2021 and 2022, the words used more frequently to target Afghan influencers online were “prostitute”, “whore”, and “dirty”, as seen from the word clouds below. The image on the left shows the most used words in 2021, while the image on the right shows the ones most used in 2022. The word clouds are based on the words used in the body of the tweets/posts.

[5] AW was not able to quantify the amount of ethnically-targeted hate speech due to the size of the datasets. However, in Part I of the report, the qualitative investigation was able to shed some light on the type of abuse that ethnic minorities, particularly Hazara women, receive.

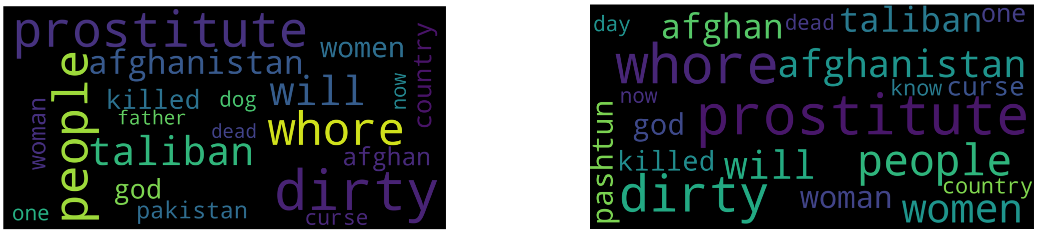

Figure: the images represent the most used words within Dataset 2. On the left, the word cloud was generated for 2021; on the right, the word cloud was generated for 2022.

5.2.1 Data from June to December 2021

Data collected from June to December 2021 from tweets targeting Afghan influencers identifies "dirty" (31.25%), "whore" (17.44%), and "prostitute" (12.92%) as the most commonly used words in posts, as seen in the graphs below.

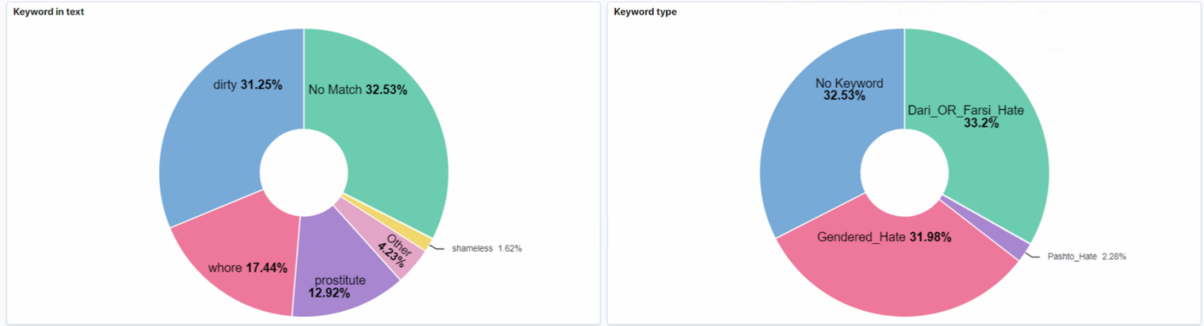

Figure: graph showing the percentages of specific keywords targeting Afghan influencers for 2021. In this graph, the percentage for gendered hate is equal to the percentage of the proportion of sexualised terms[6].

The data also shows the distribution of hate speech across the Dari/Farsi and Pashto languages. Notably, the prevalence of hate speech in Dari/Farsi (33.2%) exceeds that in Pashto (2.28%). As demonstrated by the graphs above, hate speech in Dari/Farsi exceeds that in Pashto by approximately 30.92%. Considering that over 77% of the Afghan population speaks Dari/Farsi and a significant number of female influencers in the dataset communicate in this language, it becomes evident that hate speech against them would be predominantly expressed in Dari/Farsi.

[6] The relevance system uses a mathematical equation to assess a post's relevance for collection. This equation is based on a comprehensive list of keywords. The calculation calculates the average usage of these keywords to determine the post's relevance to analysts. In other words, the higher the number of matching keywords, the higher the average relevance score. When the term "No Match" appears in the collected data, it signifies that the script has identified the post as relevant based on the keywords. However, it couldn't identify a specific keyword with the strongest match from the lists. This might occur due to translation issues when matching English keywords against Arabic-style lettering and right-to-left sequencing. This limitation arises from the coding language regex's inability to fully recognise certain Arabic languages as text. The Python scripting mandates that a post must achieve a relevance score above 5. This threshold was established after weeks of testing to ensure optimal accuracy. Posts with a score below 5 were not collected. It's important to note that any post collected and featuring "No Match" still remained relevant to analysts and the project. These posts needed a score of 5 for collection, but their keywords couldn't be precisely matched due to the Arabic style lettering.

5.2.2 Data from June to December 2022

AW's collected data from June to December 2022 shows similar findings to 2021. The commonly used words are the same; however, the percentages slightly differ. As seen from the graph below, the usage of the term “whore” (24%) was more prevalent than the words “dirty” (23.2%) and “prostitute” (16.48%). Similarly, more hate speech in 2022 was in Dari/Farsi than in Pashto.

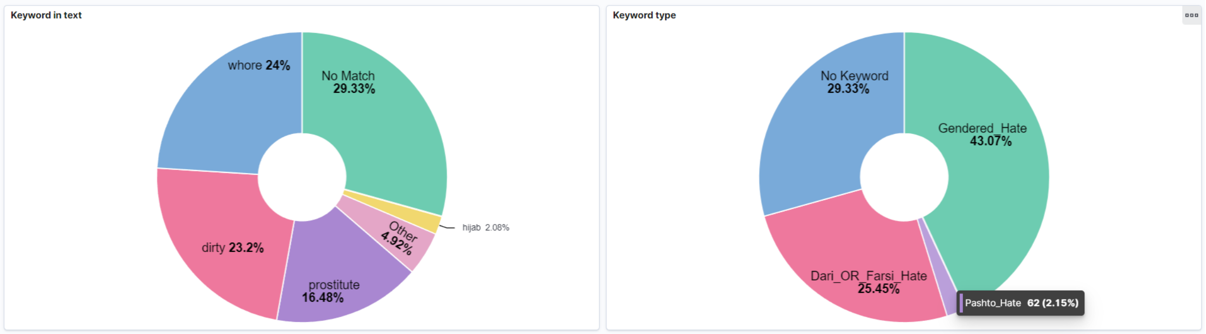

Figure: graph showing the percentages of specific keywords targeting Afghan influencers for the year 2022. In this graph, the percentage of gendered hate is equal to the percentage of the proportion of sexualised terms.

There was an overall increase of 11.09% in the proportion of sexualised terms aimed at Afghan female influencers between 2021 and 2022. In comparison to 2021, the utilisation of the terms "whore" and "prostitute" increased by 3.56% and 6.56%, respectively, in 2022. On the other hand, the term "dirty" experienced a decline of 8.05%. Similarly to 2021, the majority of hate speech targeting Afghan influencers in 2022 was predominantly in the Dari/Farsi language rather than Pashto. In the 2021 graph, a small percentage of the term "shameless" (1.62%) was utilised, while in 2022, the term "hijab"[7] (2.08%) was used on a similar, albeit smaller, scale.

[7] The term “hijab” was not used as a standalone keyword in the datasets, but it was used alongside other words that could be interpreted as hate speech.

6 Conclusion

Afghan Witness' quantitative study into online abuse targeting politically engaged Afghan women aimed to complement the previous qualitative study by quantifying the gendered hate speech Afghan women receive online. Although the scope and scale of this investigation have been narrowed down due to platform and research limitations, this report still managed to shed light on the following findings:

Research Question One: Scope and scale of gendered hate speech in the Dari/Farsi and Pashto information environment

- The general Dari/Farsi and Pashto information environment is the online space that Afghan women are most likely to move around in on a daily basis. Specific analysis of gendered hate speech within this environment shows a significant rise of hate speech and thus an increasingly hostile online space for women to navigate. Even if this hate speech might not target Afghan women specifically per se, exposure to high levels of hate speech coming from similar (cultural or geographical) contexts will likely have a negative impact on women's perceived safety and freedom online, albeit by proxy rather than directly.

Research Question Two: Scope and scale of targeted hate speech and abuse

- AW recorded an increase of **217%** in posts containing gendered hate speech and abuse terms and the names of prominent Afghan women from the period June - December 2021 to June - December 2022. This substantial increase could be related to developments in the region; X (formerly Twitter)'s changing environment; and a generally more hostile environment towards women.

- **Before and during the Taliban takeover,** spikes in gendered hate speech could be connected to major advancements the Taliban were making in the country.

- Spikes in gendered hate speech during the **second half of 2021** and the **second half of 2022** were usually connected to policies and bans that the Taliban imposed on women, ultimately restricting their rights and freedoms. The hate speech was mainly directed at the Afghan women who protested against these policies and bans.

Research Question Three: Nature of hate speech and abuse

- Hate speech and abuse directed at Afghan women was overwhelmingly sexualised. Over **60%** of the posts in 2022 contained sexualised terms used to target Afghan women. Overall, an **11.09%** increase in the proportion of sexualised terms targeting politically active Afghan women occurred from 2021 to 2022.

- It was not possible to conduct a quantitative analysis of the perpetrators of gendered hate speech. Qualitative analysis showed that this was an issue that spanned the Afghan political spectrum, with hate speech and abuse originating from users from a range of political affiliations and ethnic backgrounds.

The findings from this investigation complement those previously found in the qualitative investigation by providing an understanding of the volume of gendered hate speech targeting politically engaged Afghan women. In the previous qualitative report, AW had noticed a significant amount of hate speech targeting politically engaged Afghan women. Due to the nature of the two datasets, in particular, with one being predominantly focused on mentions of female Afghan influencers, the data collected might not fully represent the scale and severity of the problem. In addition, given the research and platform limitations found throughout the investigation, AW was not able to provide a section on attribution and was not able to further identify the perpetrators carrying out gendered hate speech. From the datasets, however, AW noticed perpetrators coming from different ethnic backgrounds and political affiliations.

Politically engaged Afghan women use X (formerly Twitter) to advocate, mobilise and attract international attention and support for women's rights in the country. For fear of being attacked, harassed or targeted, Afghan women have in some cases limited their online interactions and activity. AW has noticed that gendered hate speech targeting Afghan women is usually carried out in relation to a protest. When the Taliban imposed bans and policies restricting women's rights, Afghan women took to the streets of Kabul to protest for their rights and freedoms. This action usually sparks hate speech coming from many different perpetrators online.