THREAT MATRIX

FIFA / FIFPRO

SOCIAL MEDIA PROTECTION SERVICE

FIFA WORLD CUP QATAR 2022™ TOURNAMENT ANALYSIS

respondology | Signify

THREAT MATRIX FIFA SOCIAL MEDIA PROTECTION SERVICE
WORLD CUP QATAR 2022 – ANALYSIS

CONTENTS

SUMMARY INTRO

SCOPE + PROCESS

MONITORING + REPORTING

INSIGHTS + LEARNINGS

MODERATION

WORLD CUP FINAL

MENTAL HEALTH

NEXT STEPS

TACKLING ONLINE ABUSE

THREAT MATRIX FIFA SOCIAL MEDIA PROTECTION SERVICE WORLD CUP QATAR 2022 – ANALYSIS

INTRODUCTION

This report represents a summary analysis of all monitoring and moderation activities carried out by FIFA's Social Media Protection Service (SMPS) across FIFA World Cup Qatar 2022 ™.

The data and insights in this document were gathered from Sun 20th Nov through to 24hours after the final on Sun 18th Dec 2022, incorporating all 64 fixtures of FIFA World Cup Qatar 2022 ™.

The report covers SMPS activities across all major platforms including Twitter, Instagram, Facebook, TikTok and YouTube.

At the FIFA World Cup Qatar 2022 ™ FIFA's Social Media Protection Service delivered the most comprehensive level of defence to players, coaches and officials from social media abuse ever activated at a global sports event.

This report will provide the following analysis:

- Monitoring data + insights
- Moderation activities
- Issue categorisations
- Timelines of abusive comments
- Targeted national squads and team accounts

20m
Posts / comments analysed

287k
Comments hidden

434k
Posts flagged by AI and reviewed by humans

19.6k
Posts / comments verified as abusive and reported to platforms

12.6k
Unique accounts detected sending abusive messages

THREAT MATRIX FIFA SOCIAL MEDIA PROTECTION SERVICE WORLD CUP QATAR 2022 – ANALYSIS

MODERATION + MONITORING SCOPE

FIFA introduced the FIFA Social Media Protection Service (SMPS) to provide participating teams, players, officials and other stakeholders during FIFA events with the following:

- **Monitoring** – FIFA monitors public accounts and mentions for abusive, discriminatory or threatening content. No action is required from individuals or representatives for this service to run. Where content is detected and verified, it is actioned by FIFA, working in partnership with platforms for swift removal.
- **Moderation** – an individual or representative can opt-in to moderation of their account which allows abusive, discriminatory or threatening content to be detected and hidden in real time.

The central goal of these services is to protect players, teams, officials and fans from abuse, keeping their social feeds free from hate and allowing them to focus on enjoying their part in FIFA events.

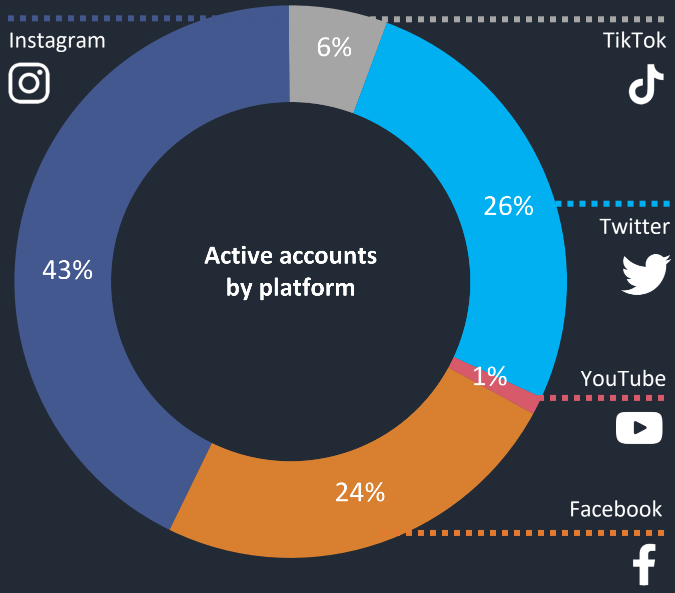
Players, teams and officials are protected against threatening, discriminatory or abusive content in FIFA's official languages and their respective nations' official language(s).

ACCOUNTS MONITORED

Across the FIFA World Cup Qatar 2022™ protective coverage was set up to incorporate the following participants:

- 864 Players and coaches (1688 active accounts)
- 129 Officials (31 active accounts)
- 32 Team accounts (126 active accounts)
- 45 Ex-Players / Media (76 active accounts)

Identified posts are based on text, with layered word, emoji and phrase categorisation and an AI-empowered threat detection algorithm.



A BESPOKE PROCESS: BUILT FOR FOOTBALL

Build a proactive net around players: AI scan for abuse of player's accounts across social media platforms.

Instantly hide abusive messages in real time on opted-in team and player accounts.

Proactively identify accounts used by abusers at scale - using a combination of AI tech and the nuance of security and intelligence experts.

Unmask abusers using specialist Open Source Intelligence forensic tools to **de-anonymise and identify abusers to an evidential standard.**

Report abusive accounts to social media platforms, Member Associations and Law Enforcement, removing the expectation on players to report abuse.

THREAT MATRIX FIFA SOCIAL MEDIA PROTECTION SERVICE WORLD CUP QATAR 2022 – ANALYSIS

BREAKDOWN OF FINDINGS

The initiative in place during FIFA World Cup Qatar 2022™ illuminated several trends and behaviours across online abuse and threat.

On-pitch events triggered abuse and threat in a manner similar to previous tournaments, with the knockout stage creating some key flash points as teams secured qualification or faced elimination.

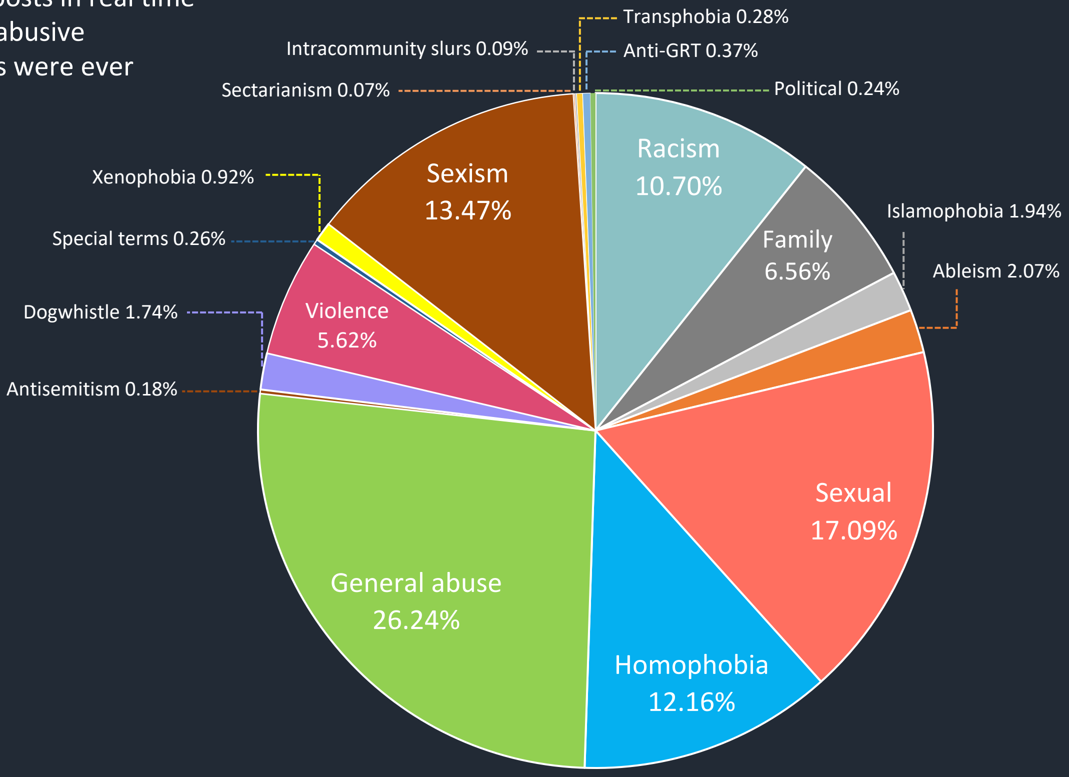
Players' political allegiances or personal circumstances were often identified as catalysts for inciting vitriol against them.

Social media companies' responses to abuse and threat published on their platforms evolved throughout the tournament but still indicated many blind spots, particularly outside of English language content.

Conversely, FIFA's Social Media Protection Service (SMPS) created a protective net around participating players, coaches and officials, reporting thousands of abusive posts in real time and hiding tens of thousands of abusive comments before players or fans were ever exposed to them.

KEY INSIGHTS

- Targeted individual racism was high volume with more than 300 players being targeted and a few individual high-profile players receiving a large proportion of targeted abuse across the competition.
- Platforms are dealing better with some incidents but there is still a highly sporadic approach to tackling and removing reported content.
- Homophobia was prolific and platform responses seemed blurred by the cultural differences which seemed to bar action.
- Violence and threat became more extreme as the tournament progressed with players' families increasingly referenced and many threatened if players returned to a particular country – either the nation they represent or where they play football.
- In the final stages of the tournament, there was more pronounced targeting of individuals, due to performance, incidents or penalty misses.



Detected abusive messages by category: All platforms

THREAT MATRIX FIFA SOCIAL MEDIA PROTECTION SERVICE WORLD CUP QATAR 2022 – ANALYSIS

MONITORING KEY STATS + INSIGHTS

The SMPS's proactive monitoring capability scanned over 20 million messages on Twitter, Instagram, Facebook, TikTok and YouTube throughout the tournament.

The system flagged 433,696 posts containing language that had been or could be used in an offensive way. This dataset was then reviewed by a team of expert analysts with a double human triage process to ensure that no posts were incorrectly attributed, removing the presence of false positives. This led to a final set of 19,636 posts confirmed as abusive, discriminatory or threatening, which have been reported to platforms for breaking their community guidelines.

Key insights:

- Twitter had by far the highest volume of identified, targeted abusive content, with 23% of posts removed by the platform following FIFA's report.
- Where there was additional escalation to Twitter after failure to act on immediate automated reporting, we observed a much higher takedown rate - highlighting the value of direct platform engagement. This mechanism was not made available by all platforms.
- Some platform engagement did improve across the tournament and there was clearly more resource applied to this issue across the final itself.
- Meta's initial response to direct reporting was often an automated response confirming; *"due to the high volume of reports we receive, our review team hasn't been able to review your report"*.
- More generally takedowns still have significant blind spots: whilst English language racism generated higher takedown rates, overt racism or homophobia in other languages remained live for longer.
- Human verification of AI-assessed material ensured that risk was minimised in all content reported, allowing for regional and cultural nuances to be incorporated into the review.

19,636

Verified instances of abuse or threat reported to platforms *

Reports to Twitter	13,105
Reports to Instagram	5,370
Reports to Facebook	979
Reports to TikTok	69
Reports to YouTube	113

* Abuse totals based on the number of reported breaches of platform's own community guidelines.

REAL WORLD ACTION

These activities are linked to real-world action that starts with submission to platforms and follows through to account investigations and detection of the owners of abusive accounts.

The SMPS Open Source Intelligence (OSINT) team have reviewed and triaged every detected abusive message, where possible identifying the account owners and gathering any available evidence to take action to prevent further targeted online abuse.

The figures (right) relate to the two-stage investigations process.

STAGE ONE: TRIAGE OF ABUSE BY TIER

The SMPS operates on a three tier categorisation of abusive comments:

- **Tier 3 -** where content breaches platform terms of service or community guidance and should be removed.
- **Tier 2 -** an additional threshold where Member Associations or domestic clubs may wish to act via education initiatives or ticket sanctions.
- **Tier 1 -** where jurisdictional law enforcement action may be appropriate.

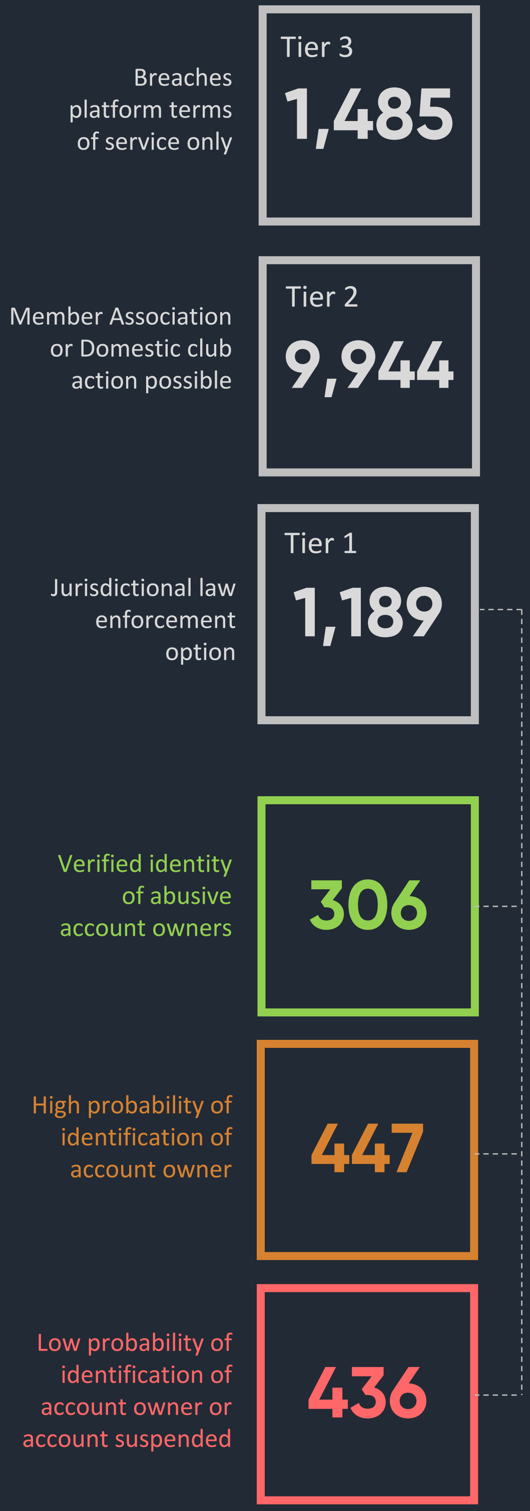
STAGE TWO: INVESTIGATION OF ACCOUNT (TRAFFIC LIGHT)

The SMPS process segments Tier 1 accounts by likelihood of author identification:

- **Green** = identity of account verified
- **Amber** = accounts under investigation
- **Red** = accounts likely to require a disproportionate level of resource to fully identify

All investigations are focused upon Tier 1 and Tier 2 accounts.

12,618 accounts were recorded sending abuse or threat across the tournament. The full tournament tiering plus traffic light numbers at time of publication for Tier 1 are shown (right).



Breaches platform terms of service only

Tier 3
1,485

Member Association or Domestic club action possible

Tier 2
9,944

Jurisdictional law enforcement option

Tier 1
1,189

Verified identity of abusive account owners

306

High probability of identification of account owner

447

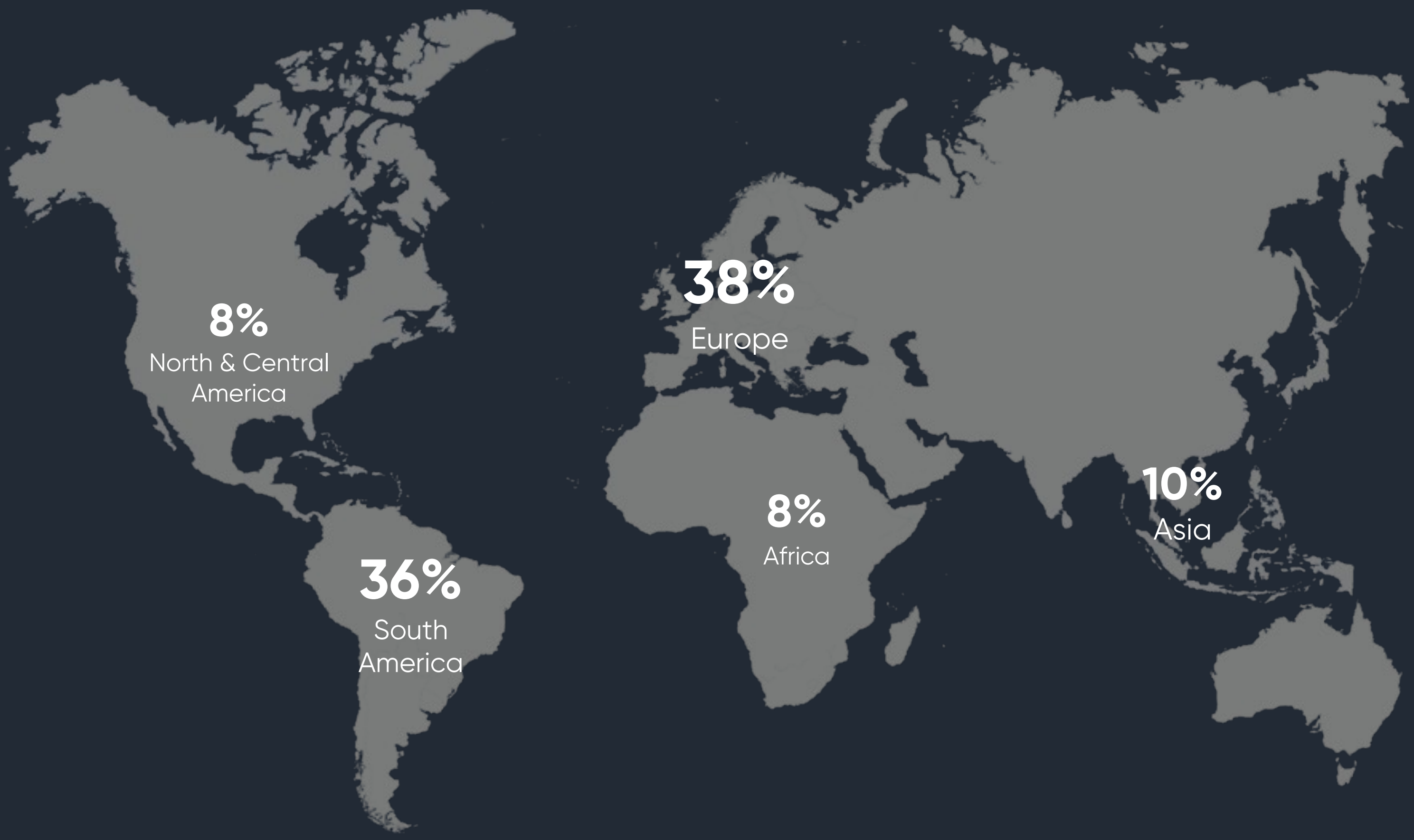
Low probability of identification of account owner or account suspended

436

GEOGRAPHIC LOCATION

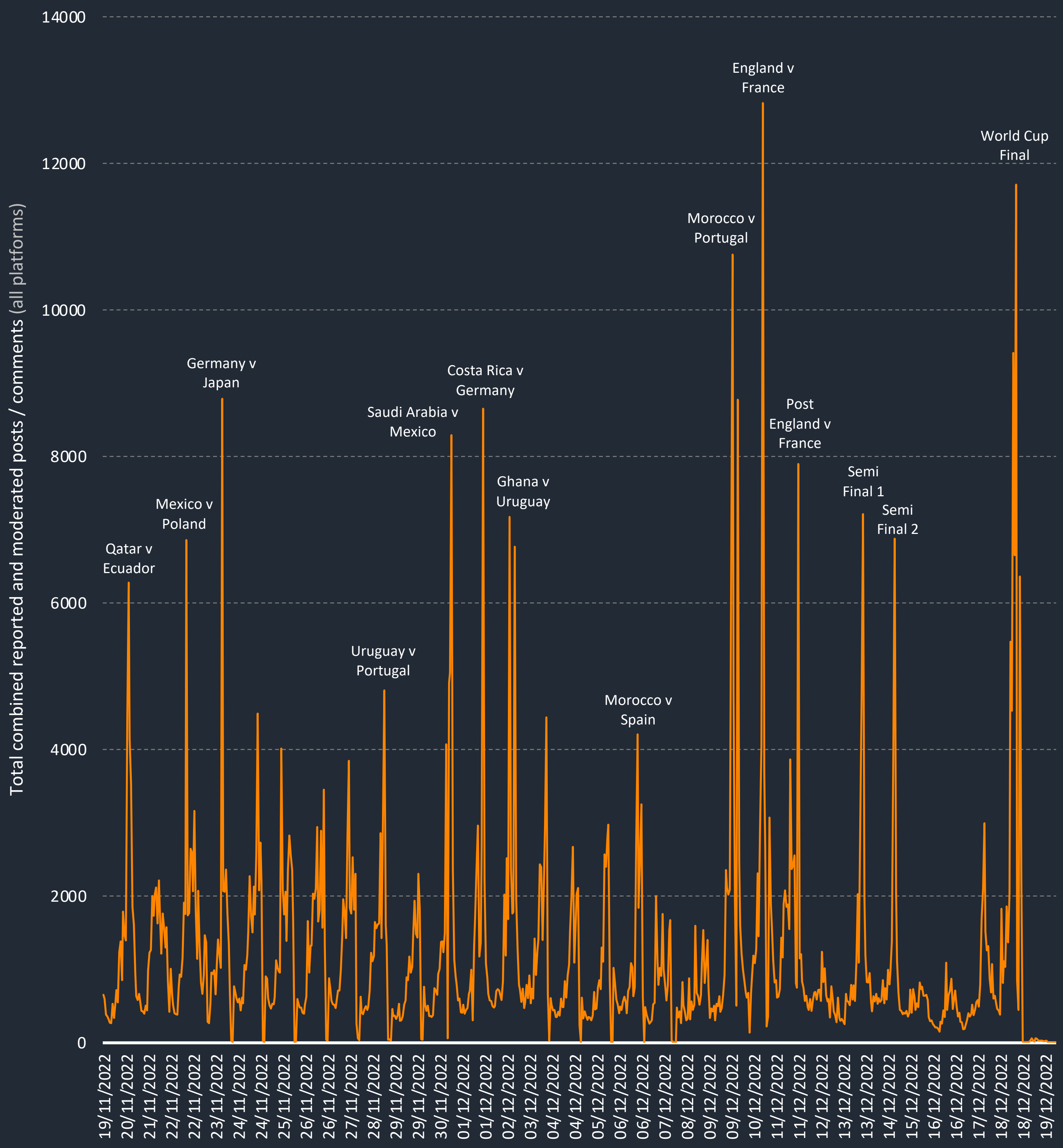
Verified geography of abusive accounts (sources of posts / comments).

This chart is a representation of accounts where author location has been verified, across the world, drawn from abuse reported to social media platforms at the FIFA Qatar World Cup 2022TM.



Based on 12,618 social media accounts
with 7,204 identified locations

DETECTED ABUSE TIMELINE – FULL TOURNAMENT



THREAT MATRIX

FIFA SOCIAL MEDIA PROTECTION SERVICE
WORLD CUP QATAR 2022 – ANALYSIS

TARGETED PLAYERS + TEAMS

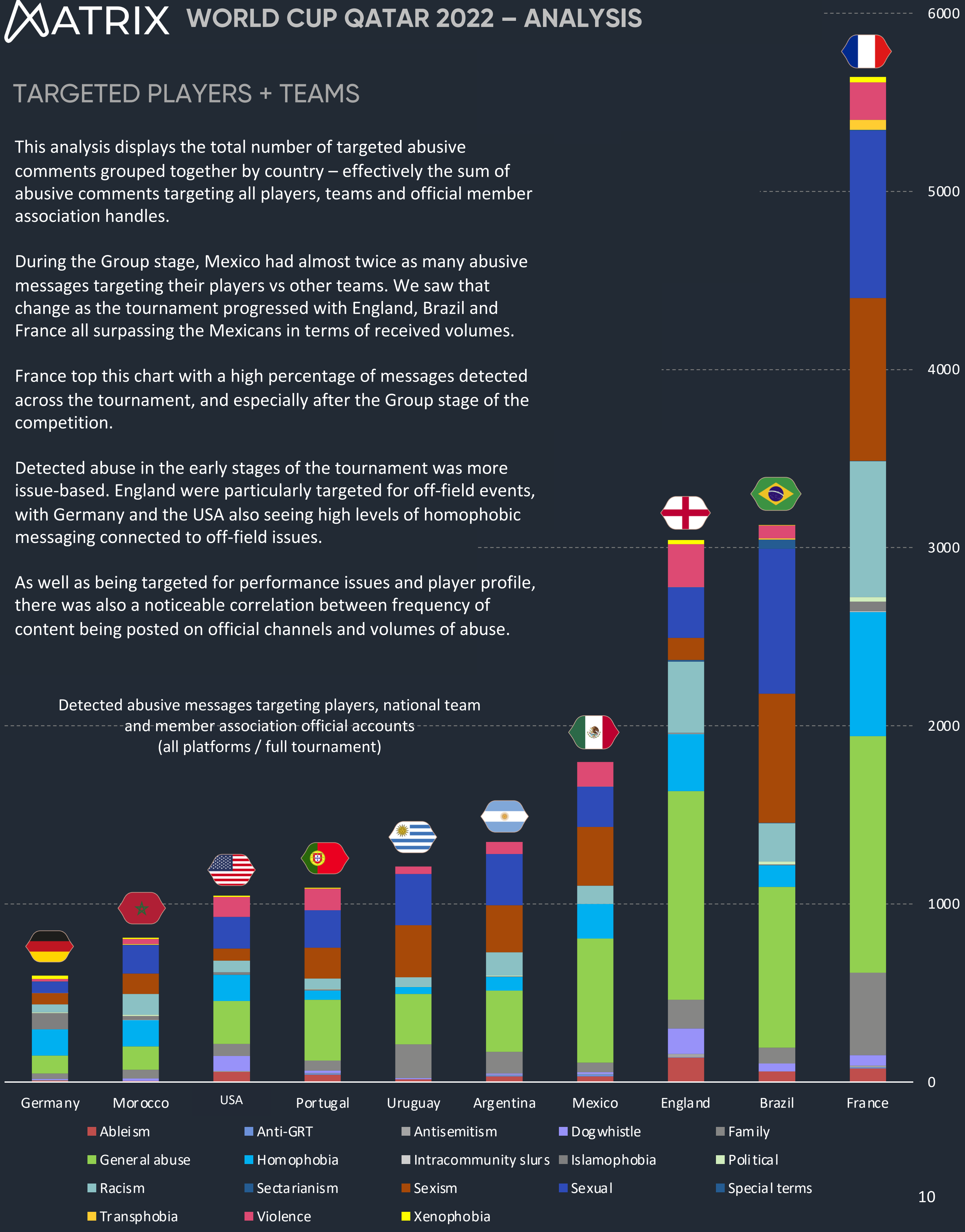
This analysis displays the total number of targeted abusive comments grouped together by country – effectively the sum of abusive comments targeting all players, teams and official member association handles.

During the Group stage, Mexico had almost twice as many abusive messages targeting their players vs other teams. We saw that change as the tournament progressed with England, Brazil and France all surpassing the Mexicans in terms of received volumes.

France top this chart with a high percentage of messages detected across the tournament, and especially after the Group stage of the competition.

Detected abuse in the early stages of the tournament was more issue-based. England were particularly targeted for off-field events, with Germany and the USA also seeing high levels of homophobic messaging connected to off-field issues.

As well as being targeted for performance issues and player profile, there was also a noticeable correlation between frequency of content being posted on official channels and volumes of abuse.



Detected abusive messages targeting players, national team and member association official accounts (all platforms / full tournament)

MODERATION ACTIVITIES

FIFA's SMPS intercepted and hid abusive, discriminatory and threatening comments (along with spam) across 22 languages throughout the World Cup on Instagram, Facebook and YouTube.

Comments were hidden on official team channels as well as player accounts.

Levels of abuse can often be driven by frequency of posts and content so some Member Association accounts / players will attract more targeted messages than others.

How does moderation work?

An individual or their representative can opt-in to moderation of their account which allows abusive, discriminatory or threatening content to be detected and hidden in real time.

This service was made available to all players across FIFA World Cup Qatar 2022™.

Twitter does not currently offer functionality for a user to fully hide a reply to one of their tweets so that platform was not covered by the moderation element of the SMPS, while TikTok does not currently allow automatic, API-driven moderation of comments.

286,895

Abusive, discriminatory and threatening comments hidden targeting player / coach / team accounts

167,108

Moderated comments on Facebook participant / team accounts

118,413

Moderated comments on Instagram participant / team accounts

1,374

Moderated comments on YouTube participant / team accounts

4%

Proportion of reviewed messages / comments hidden

THREAT MATRIX FIFA SOCIAL MEDIA PROTECTION SERVICE
WORLD CUP QATAR 2022 – ANALYSIS

WORLD CUP FINAL ANALYSIS + ACTION

Played out over 120 mins and culminating in a thrilling penalty shoot-out, the final of FIFA World Cup Qatar 2022TM featured numerous incidents that traditionally serve as triggers for targeted abuse of players, coaches, and teams.

During and following the World Cup Final, FIFA's SMPS service detected and immediately reported messages targeting players involved in the match to both the Member Association and relevant platforms for immediate removal.

FIFA's SMPS initiative provided all players with unprecedented levels of protection online during the final – and for the duration of the tournament. The service ensured almost 1,255 abusive posts were reported to platforms for removal during the 24 hours of the final.

In one example on Instagram, racist abuse was detected coming from an account where even the name of the account contained identifiably abusive and racist terms, clearly breaching Meta's terms of service.

This flagged a vulnerability in the platform's own review process – as the offensive account remained live for more than 4 months after the tournament ended - despite being reported on the day of the final.

The numbers displayed below are the detailed actions, per platform, carried out by FIFA's Social Media Protection Service on the day of the World Cup Final.

On the day of the World Cup Final (and 24hrs after)…

1.6m
Posts + comments captured for analysis

44,710
Posts/comments flagged by AI monitoring for human review

1,255
Verified instances of abuse or threat reported to platforms following review*

2,946
Comments hidden

MONITORED + REPORTED

1,004	Reports to Twitter
5	Reports to TikTok
16	Reports to Facebook
226	Reports to Instagram
4	Reports to YouTube

MODERATED + HIDDEN

876	Hidden comments on Instagram
5	Hidden comments on YouTube
2,065	Hidden comments on Facebook

* Abuse totals based on the number of reported breaches of platform's own community guidelines.

WORLD CUP FINAL COMPARISONS

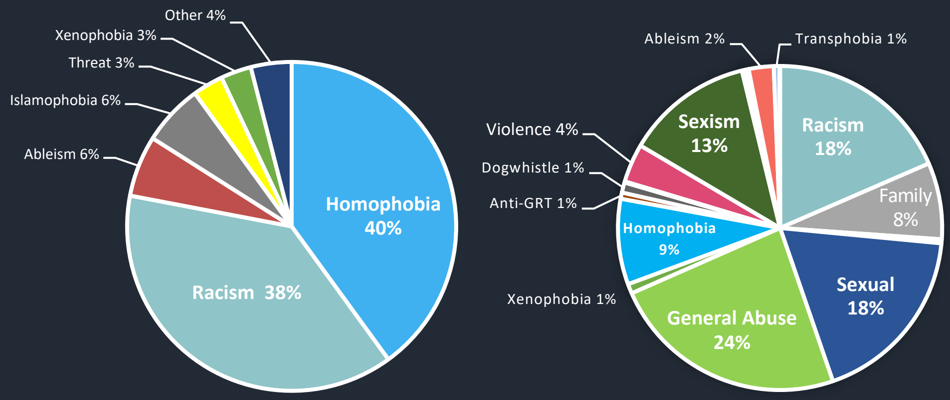
COMPARISONS + LEARNINGS: AFCON 2021 / EURO 2020

FIFA's development and deployment of the Social Media Protection Service (SMPS) was informed by its study of abuse targeting players at the 2022 Africa Cup of Nations and EURO 2020 finals.

This retrospective study examined over 400,000 social media posts across Twitter and Instagram targeting players involved in the respective finals. It found over 55% of players involved received some form of discriminatory abuse during those matches. Crucially, with no live monitoring or moderation solution in place, such abuse went unchecked on the platforms at those events.

The FIFA World Cup Qatar 2022™ final was notable for the range of abuse types targeting players online. The finals of AFCON 2021 and EURO 2020, by contrast, were targeted most heavily by racist and homophobic content with 78% of all detected abuse falling into one of those two categories.

Categorisation of detected abuse in Finals

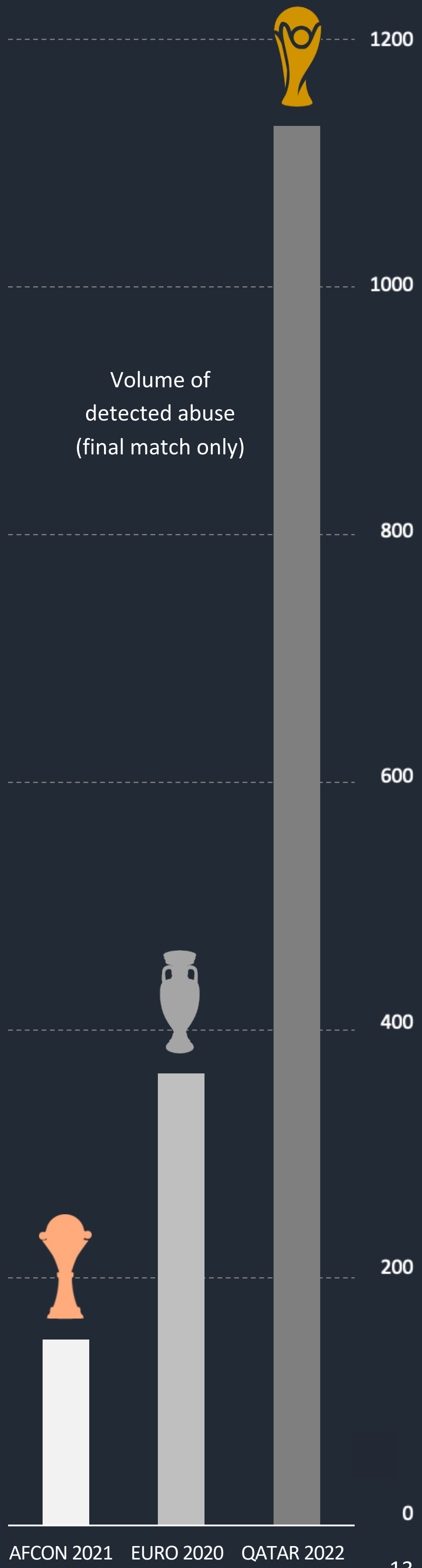


AFCON 2021 / EURO 2020 finals

FIFA World Cup Qatar 2022™ final

Racist and homophobic abuse is typically the most egregious and more easily identifiable / actionable by platforms. Whereas more nuanced abuse (such as targeting of family members, more subtle dogwhistles or abuse sent in non-English languages) makes traditional detection more complex. This is a powerful learning and progression from the Qatar World Cup study as FIFA's SMPS initiative implemented a wide-ranging focus on abuse in different languages, enabling a much higher capture, reporting and moderation success.

These learnings will further enhance FIFA's SMPS moderation filters to improve capture and action ahead of the FIFA Women's World Cup Australia and New Zealand 2023™.



THREAT MATRIX FIFA SOCIAL MEDIA PROTECTION SERVICE WORLD CUP QATAR 2022 – ANALYSIS

THE MENTAL HEALTH TOLL ON PLAYERS

The hatred and abuse that exists online is a social crisis that touches individuals all over the world and cannot be simply ignored or easily shrugged off. In professional football, this toxic online environment is a difficult and risky place to be in for players. Footballers are heavily exposed to the continuous vitriol of online trolls and abusers, with high-profile players being the most vulnerable to large-scale abuse, as evidence shows.

The risks and mental health challenges associated with being victim of online abuse are real and have a direct and immediate effect on players. Hatred and discrimination in the online environment can be damaging at both a personal and professional level, negatively impacting the players' ability to be and perform at their best.

Abusive comments are personal attacks on the identities and characteristics of players that can have detrimental effects on their overall well-being and can also lead them to hide and withdraw from who they are and want to be.

The consequences of being exposed to continuous abuse online for professional football players across the globe are underestimated and oftentimes understated. Footballers can feel discouraged to admit feeling impacted by social media abuse and may choose to act as if all is well with them. These self-protective strategies can place players at even greater risk of not addressing the impact of online abuse on their wellbeing and of not receiving the support they need.

> *"I know there's a competitive edge in the game, but I think the things that have been said online go far beyond that, especially when you're talking about families or something that cuts a little bit deeper than just actual play. I think it needs to stop and our voice needs to be heard that we're people as well."*
>
> *Kellyn Acosta (USA)*

INSIGHTS PROVIDED BY FIFPRO

> *"Players don't speak out much about how online abuse affects us. There's stigma. As professional athletes, we don't want to admit this could hurt us, or that we even notice it. But we're all human. It's not as easy as "just tuning it out." It hurts. It hurts our families."*
>
> *Mark-Anthony Kaye (Canada)*

REAL WORLD SYMPTOMS

The horrific abuse that we see across social media and all online platforms exposes players across the globe to serious risks.

Being victim of online abuse and continuous hatred online can lead to a range of real-world consequences and symptoms, including (but not exhaustively):

- anxiety attacks
- depression
- accumulation of trauma
- low self-worth
- sleep disturbances
- change in eating patterns
- feelings of inadequacy
- social withdrawal and isolation
- and in extreme cases, death by suicide.

FIFPRO FOOTBALL PLAYERS WORLDWIDE

Analysis on the mental health toll on Players contributed by FIFPRO.

NEXT STEPS – LEARNINGS + PREPARING FOR FUTURE EVENTS

FIFA World Cup Qatar 2022™ represented the first of several events and tournaments that the FIFA SMPS will be activated for. This initial deployment provided many valuable learnings and potential advancements of the service.

1. Insights & Analysis

This study and analysis covering FIFA World Cup Qatar 2022™ provided a rich dataset allowing for the further strengthening of the SMPS service. The work and actions generated by this data have continued beyond the event and include:

- Issue analysis – deep dives on trigger points and spikes of abuse
- Keyword / emoji filter consolidation and strengthening recommendations
- Investigations into identified accounts
- Wider participant analysis to include FIFA legends, media targets and match officials
- Recommendations for strengthened solutions

2. Platforms status analysis

- Review performance with platforms and enhance understanding of successes and failures (why have some categories and languages been effectively dealt with whilst others are not)
- Improving process and communication with social media platforms to ensure faster and more effective action - feeding into platform proactive measures and tools for long-term success
- Work with platforms to implement learnings from this study

3. Investigations + action

- Working with Member Associations to agree strengthening of protection and any action available
- Work with jurisdictional law enforcement to submit evidence packs to build cases where abusive messages have passed criminal thresholds

4. Player support + welfare

- FIFPRO review of learnings and insights
- Communicate successes of SMPS to players and Member Associations
- On-boarding for future events

5. Future events preparation

- On-going moderation protection for on-boarded players
- Coverage of further FIFA events including FIFA Club World Cup 2023, FIFA U-20 World Cup and other events
- Expansion of social media channel / platform coverage
- Expansion of language database to incorporate teams not represented at FIFA World Cup Qatar 2022™

6. Preparation for FIFA Women's World Cup Australia and New Zealand 2023™

- Detailed analysis of slurs and abusive comments historically targeted at Women's football
- The development and strengthening of categories and filters pertaining specifically to Women's football
- Work with players, ex-players and experts to analyse and verify terms and filters, including new weaponised emojis and identified tactics
- Engagement with host nation(s) officials and representatives to incorporate nuanced regional issues

TACKLING ONLINE ABUSE

WHAT WE'RE DOING

FIFA does not tolerate abusive, discriminatory or threatening behaviour targeting officials, players, coaches, staff or their family members.

The SMPS service **moderates** (i.e. to hide or delete, varying by mechanics of the platform in question) abusive, discriminatory and threatening content including but not limited to text, emojis and images.

Where platforms do not support moderation, FIFA also proactively **monitors** social media for abusive, discriminatory and threatening content.

FIFA has and will continue to take action against individuals or groups who are evidenced to have produced or disseminated abusive, discriminatory and threatening social media content, including **reporting** offenders to social media platforms and working with regional law enforcement where appropriate.

Content will be subject to moderation, monitoring or reporting where it includes a reference – whether express or implied – to any one or more of the reasons listed by Article 4 of the FIFA Statutes, where the context may be reasonably concluded to be harmful. Additionally, any content which may be deemed to include threat of harm to the subject or their family members will be automatically included for assessment.

Abusive, discriminatory and threatening behaviour has no place in football.

WHAT WE WON'T DO

FIFA will not use private data or seek to compromise platform-based privacy settings.

FIFA will not deploy surveillance technology to monitor individuals – both moderation and monitoring are issue-based, not individual-based.

FIFA is a strong advocate of free speech and this policy is only designed to tackle abusive, discriminatory and threatening content.

ARTICLE 4 OF THE FIFA STATUTES

> *"Discrimination of any kind against a country, private person or group of people on account of race, skin colour, ethnic, national or social origin, gender, disability, language, religion, political opinion or any other opinion, wealth, birth or any other status, sexual orientation or any other reason is strictly prohibited".*

GLOSSARY

- **Abuse / Abusive posts and accounts:** Refers to content that includes verified discriminatory, egregious and aggravated terminology.
- **Discriminatory flags:** Posts flagged for content that may include racist, homophobic, sexist etc. terms.
- **Dogwhistle:** An abusive message clearly exploiting a racist trope without explicitly using directly racist language.
- **Fan identification:** Info on profile that indicates with high likelihood a user supports a particular country / domestic football club.
- **Flagged posts:** Posts flagged for content that may include personally abusive or discriminatory content. Personally abusive content can include calling someone a a**hole, etc.
- **GRT:** Gypsy Roma and Traveller communities.
- **Inclusion criteria:** A post will have mentioned one of the monitored player handles or one of two key terms (or variations thereof). Multiple reports can be generated from a single post / comment if more than one offence is detected.

THREAT MATRIX

FIFA / FIFPRO

SOCIAL MEDIA PROTECTION SERVICE

FIFA WORLD CUP QATAR 2022^{TM}

TOURNAMENT ANALYSIS

respondology | Signify