WORLD ATHLETICS

ONLINE ABUSE IN ATHLETICS

AI Research Study: World Athletics Championships Budapest 23

THREAT MATRIX

INTRODUCTION

World Athletics has deployed Threat Matrix to protect athletes from the scourge of online abuse at the 2023 World Athletics Championships in Budapest, Hungary. This follows World Athletics' successful use of Threat Matrix to study abuse targeting athletes at the 2020 Summer Olympics in Tokyo and 2022 World Championship in Oregon, USA. This report presents the key findings and insights from this study.

Threat Matrix is an initiative by ethical data science company Signify Group (Signify) supported by sports investigations company Quest Global Ltd (Quest). It uses machine learning and AI Natural Language Understanding to detect abuse and fixated threat online, helping World Athletics understand the issues, delivering actionable, real-world solutions.

World Athletics' deployment of Threat Matrix at the Tokyo Olympics and Oregon World Championships was designed to illuminate the size, scale and gravity of the issue of online discriminatory abuse being targeted at athletes on Twitter/X. At the World Athletics Championships in Budapest, the analysis built on the success of the previous two events to take athlete protection several steps further. Nearly three times the athletes were monitored compared to the World Athletics Championships in Oregon 2022.

This report presents the findings from Budapest, highlighting the tactics used to abuse and threaten athletes online and has formed the basis of recommendations for ongoing monitoring and analysis, providing an updated blueprint detailing solutions to protect athletes at all times, worldwide.

METHODOLOGY

To protect athletes and gain an understanding of online abuse surrounding the 2023 World Athletics Championships, the Threat Matrix team monitored the accounts of 1,344 competitors across social media platforms Twitter/X and Instagram.

The 1,666 active athlete accounts across Twitter/X or Instagram, representing 1,344 athletes, were derived from a list of 2,195 athletes provided by World Athletics ahead of the championships.

Threat Matrix's monitoring began on August 18th and ended August 28th, taking in the duration of the 2023 World Athletics Championships in Budapest, Hungary. The timeframe was designed to detect any abuse in the run-up to and fall-out from the World Championships, as well as abuse during the event itself.

After the Threat Matrix linguistic filters removed any identified spam, the team captured 449,209 posts and comments on Twitter/X and Instagram for review. For the purposes of Threat Matrix, a post on Twitter/X refers to any tweet in which an athlete's @ handle is mentioned. On Instagram, a comment refers to any comment made in response to an organic in-feed upload by an athlete.

The team then used text analysis to search for slurs, weaponised emojis and other phrases that could indicate abuse. Image recognition tools were also deployed to identify offensive images. AI-powered Natural Language Processing was also applied to detect threats by understanding the relationships between words. This allows Signify to determine the difference between, for example, 'I'll kill you' and 'You killed it'.

157,061
Tweets captured for analysis

292,148
Instagram comments captured for analysis

1,344
athletes selected for monitoring with 1,666 active accounts (from a list of 2,195 provided by World Athletics)

1 SOURCE DATA
Using AI-powered threat detection algorithm, Signify scan public posts

2 CLEAN DATA
Removing bots, discriminatory abuse targeting selected athletes is flagged

3 ANALYSE DATA
Flagged posts are analysed by Signify's team of experts

4 EVIDENCE + ACTION
The most egregious examples of abuse are prepared and submitted for action

KEY FINDINGS

Analysis of the monitoring period identified clear instances of online abuse and threat, targeting athletes at the 2023 World Athletics Championships. Threat Matrix detected notable examples of racist and sexualised abuse, with a selection of posts extending into potential action from law enforcement.

The levels of abuse detected in Budapest were noticeably higher when compared with the similar study at the 2022 World Athletics Championships, even allowing for the increased number of athletes monitored, however the number of targeted athletes was only a third higher, meaning those targeted experienced more intense focus. The Budapest championships saw more than four times the abuse targeting an increased dataset of 47 athletes.

Further comparisons between Budapest 2023 and Oregon 2022 will be made later in this report. However, key findings from this latest study are as follows:

- Racism and sexualised abuse continue to be used to target athletes – over 51% of all abuse detected during the analysis period came from one of these categories
- X was the preferred channel for abusers, accounting for almost 90% of detected abuse – a 500% relative increase from 2022
- Unlike previous tournaments where abuse was driven by events outside the stadium, most abuse was targeted at athletes based on their wider reputations or comments they made at the championships, particularly in relation to the media
- Two athletes were the most heavily targeted athletes in the selection, accounting for around 44% of detected abuse between them.

258
abusive posts were identified coming from 237 unique authors

47
of the 1,344 tracked athletes received targeted abuse

35%
of identified abuse was racist in nature

16%
of identified abuse was sexual or sexist in nature

90%
of detected abuse was made on X (formerly Twitter)

46%
of athletes who were abused in the analytics came from the USA

PLATFORM BREAKDOWN

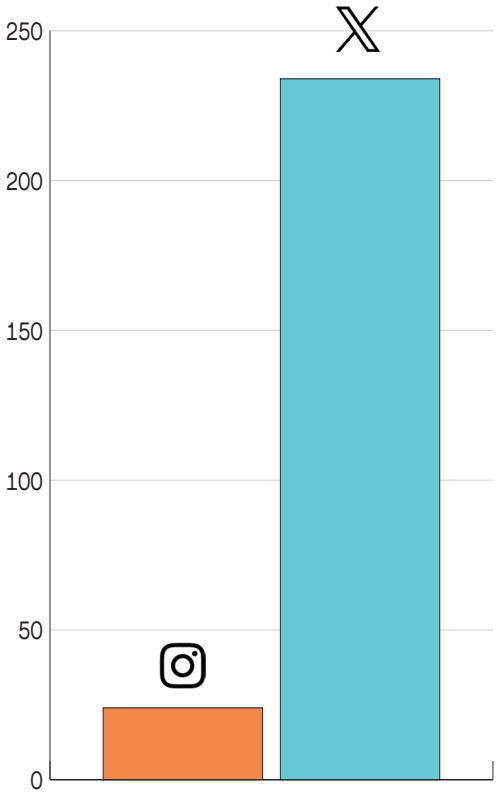
Platform coverage for 1,344 athletes found 1,666 active accounts.

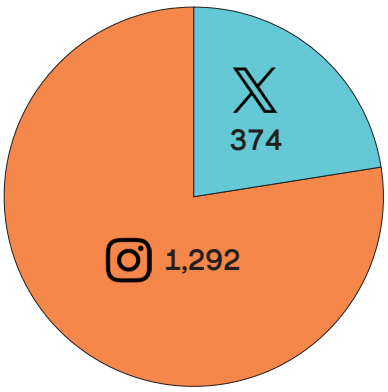
X (formerly Twitter) had 374 active accounts – representing 22.5%

Instagram had 1,292 accounts – representing 77.5%

This split is the inverse of detected abuse.
(See page 9 for comparisons with Oregon 2022).

Breakdown of detected abuse by platform





Breakdown of active accounts by platform

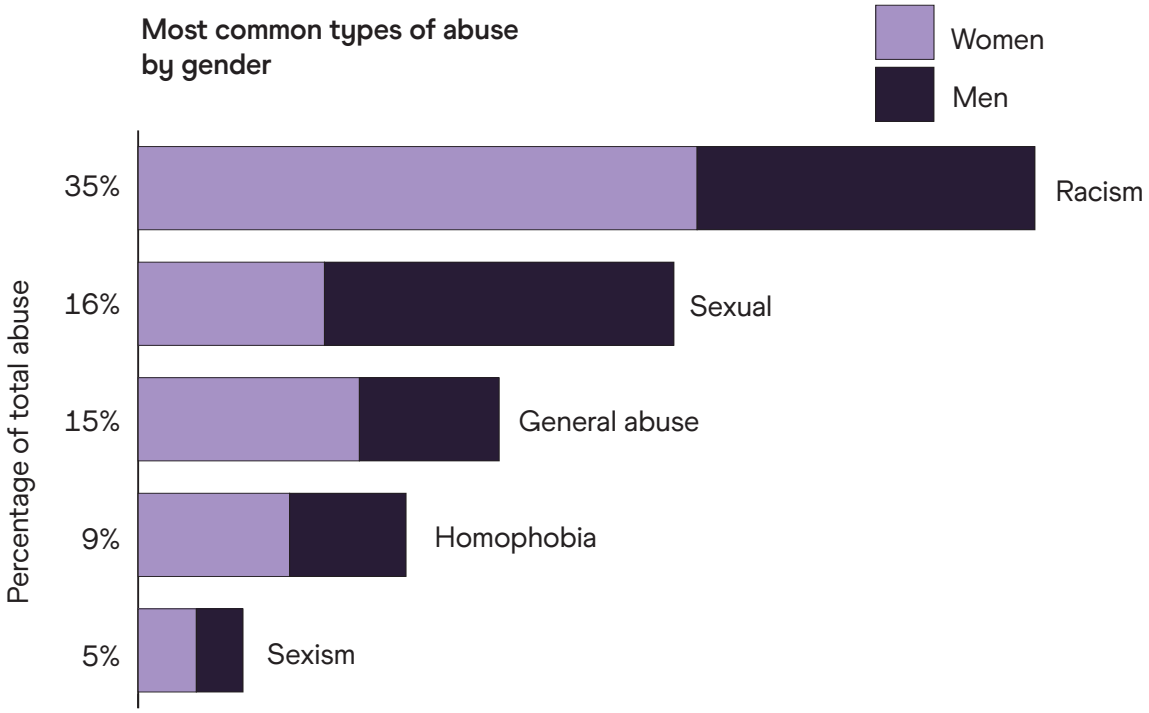
ANALYSIS OF ABUSE

In the analysis of detected abuse, the most frequent types of abuse were racism and sexualised comments – comprising over 56% of all abuse during the monitoring timeframe.

Together, sexualised and sexist abuse – overwhelmingly targeted at female athletes – made up over 21% of all detected posts.

In an evolution of previous analysis, racism was not only a form of abuse but an allegation levelled at athletes, particularly black athletes. This was frequently used as a mask for the posters' own racism.

Unfounded doping allegations were used as the springboard and justification for other forms of abuse.



TRIAGE ANALYSIS OF ABUSE

As part of this analysis, a detailed triage was applied to all flagged abuse, with the content and context of each message assessed and categorised into a set of tiers according to the gravity of abuse and the recommended actions.

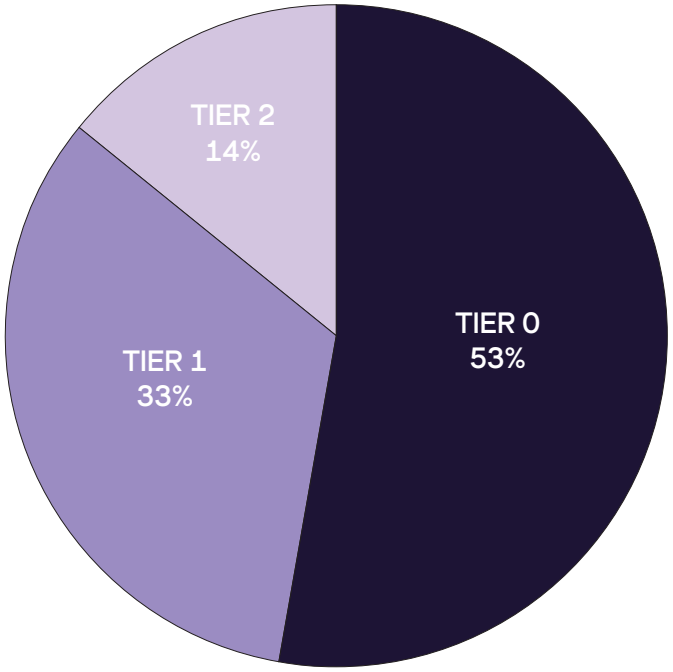
Similarly, the Threat Matrix service assesses all flagged posts and comments for threat. Across this dataset, there was nothing deemed as imminent threat (by contrast to previous events), in accordance with Threat Matrix proprietary indicators for fixated threat. No cases were identified for immediate escalation to law enforcement.

The majority of cases contained language that was abusive but required no further action. 33% of the abuse was escalated to the platforms for additional action and 14% of the abuse was deemed suitable for both platform action and assessment by World Athletics of possible further action up to and including assessment of appropriateness of jurisdictional law enforcement involvement.

TIER 0
Flag to World Athletics

TIER 1
Report to platform + World Athletics

TIER 2
Report to platform + World Athletics + assess for law enforcement



EVENT COMPARISON

Threat Matrix's studies for the 2020 Summer Olympics in Tokyo and World Athletics Championships in Oregon 2022 analysed almost 250,000 and 428,000 posts respectively targeting a representative sample of 161 athletes and officials/media personalities in Tokyo and 458 in Oregon.

For Budapest this was increased to cover 1,344 athletes and officials/media personalities with accounts and resulted in 449,209 posts and comments being analysed.

Given the different range of accounts monitored it is hard to draw detailed comparison between the three events. However, there are useful trends that can be detected.

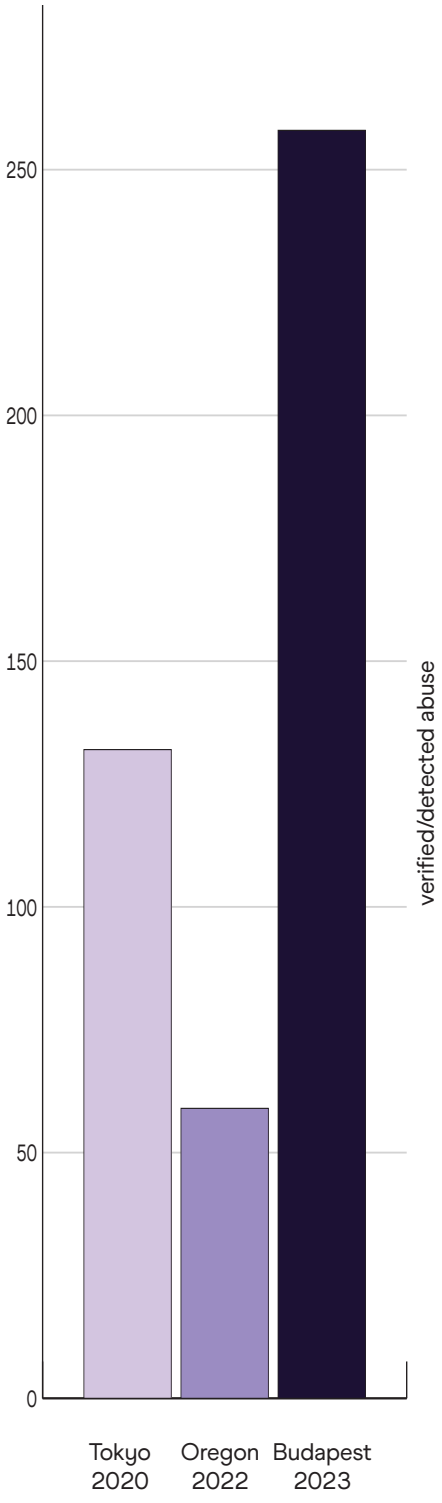
In Tokyo the process detected 132 posts flagged as abusive from 119 authors, targeting 23 athlete accounts.

In comparison, the volume of abuse targeting athletes at the 2022 World Athletics Championships appears much lower. 59 abusive posts were detected against a much larger set of athlete accounts.

However, the figures for the 2020 Olympics are influenced by two athletes receiving almost 65% of all abuse during the Tokyo Games. When the abuse targeting those athletes is stripped out, similar volumes can be observed across the two events.

At Budapest in 2023 nearly 260 cases of abuse were detected, almost double the number of incidents in Tokyo and 5x the number detected in Oregon. However, there are several factors to consider in drawing conclusions.

Both Tokyo and Budapest saw two athletes account for the majority of abuse and highlight the fact that abuse often follows abuse with more abusers joining in once an athlete has been singled out. Secondly, Tokyo saw over half as much abuse from a far smaller sample set showing the profile that an Olympic Games brings.



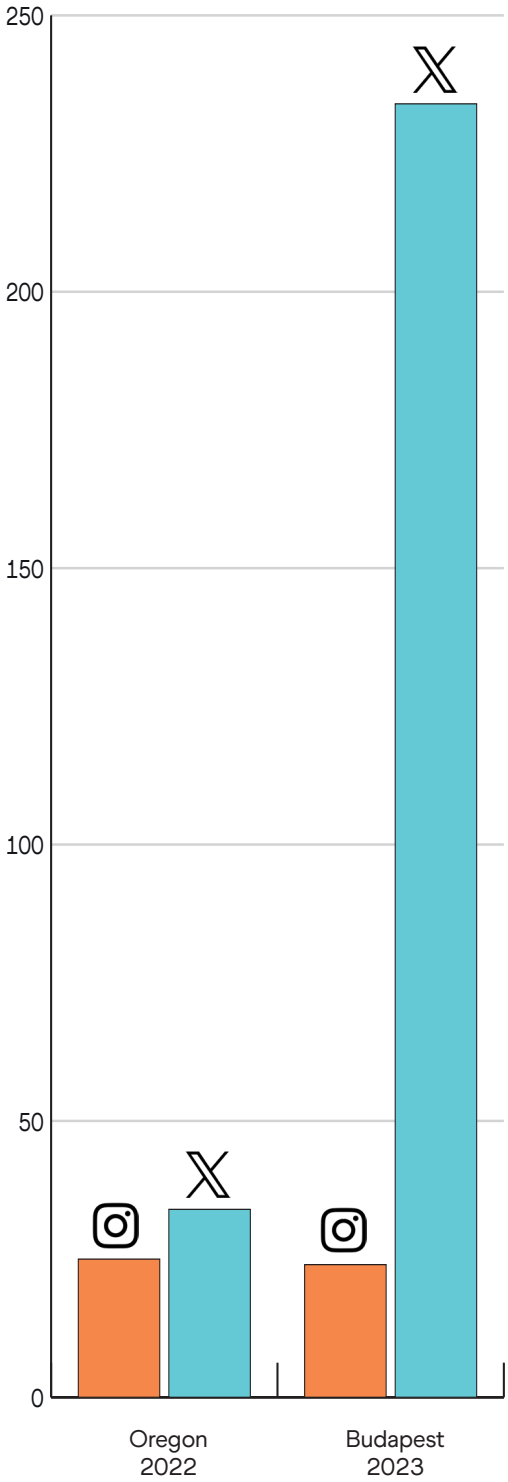
Considering the relevant number of athletes monitored for each competition this represents a small rise but is a consistent level of abuse with the previous World Athletics Championships. Budapest also saw a far higher number of individual athletes receive abuse highlighting the importance of monitoring all athletes.

Analysis of platform breakdowns between Oregon and Budapest shows a stark increase in Twitter/X cases but a consistent level of Instagram abuse. This illustrates two points:

1. Twitter/X remains the pre-eminent platform for watching live events and therefore abusing athletes in real time, and Instagram use is evolving amongst athletes.
2. Far fewer athletes post during competition time lending an atmosphere of stasis to their accounts – those that do often only do so to celebrate success, and many employ some form of comment management where platforms allow.

As the study shows abuse and threat are still getting through to athletes on Instagram, but it is not the primary platform of choice for attacking athletes, which the data indicates remains Twitter/X.

This is relevant to note as it suggests World Athletics faces a consistent issue of online abuse targeting athletes at major events. However, the ability to detect and report these to their respective platforms demonstrates the availability of solutions and protective measures that can be deployed.



CONCLUSION

This study represents the third deployment of Threat Matrix's AI-powered methodology following previous reports covering the Tokyo 2020 Olympic Games and the 2022 World Athletics Championships in Oregon. Alongside continuing to highlight the prevalence of abuse targeting athletes at a range of events across multiple years the report is also the first that allows us to draw like-for-like comparisons on two editions of the same event.

The expansion of athlete coverage for Budapest 2023 has shown that the number of athletes targeted during a competition can run deep, however, the competition saw a handful of athletes singled out to receive the majority of abuse.

This is significant as it represents both a need and opportunity to shape the way in which athlete welfare is handled during and after competitions. There is an opportunity to pre-emptively intervene with athletes who are being targeted – ahead of a spike of abuse/pile-on occurring.

Threat Matrix analysis has also shown similar issues continue to plague athletes in terms of abuse from Tokyo 2020 and Oregon 2022. There continues to be a significant portion of racist abuse, while female athletes continue to face sexualised and sexist comments online.

Homophobic abuse continues to be prevalent but has evolved to encompass more than just targeting those who have publicly declared their sexuality, perceived by the authors as one of the most insulting allegations that can be thrown at an athlete, particularly male athletes.

Detecting abuse among the millions of tweets and comments directed at athletes requires AI-enabled solutions and a specialist expertise. Having these abusive posts taken down by platforms is, however, just the first step. Threat Matrix has also continued to demonstrate through its triage, analysis, and investigative work that accounts behind this abuse can be uncovered and – in many cases – deanonymised, paving the way for real-world action that, in the long-term, will build deterrence and show abusers they cannot behave in this way without facing consequences.

Whilst the data clearly shows that Twitter/X remains the pre-eminent platform for those engaging 'live' with events such as athletics, seeing by far the most abusive content during or on the immediate wake of events, Instagram contains a higher proportion of serious abuse, often coming later and in a more calculated fashion than Twitter/X abuse.

It is also clear that despite many athletes not posting frequently on social media during competition time – if at all – and limiting or managing who can comment, serious abuse, is still getting through and requires monitoring to ensure it is not left unchallenged.

Feeding back to the platforms and acting in partnership with them will be essential in enacting the goals of this report.

Threat Matrix and World Athletics will continue to work with the platforms to be able to highlight this and work towards ending the scourge of abuse targeting athletes on their channels.

APPENDIX: DETAILED METHODOLOGY

CHANNELS AND MEDIA FORMATS

This study focuses on discriminatory abuse and threats captured on Twitter/X and Instagram in a range of media formats:

- Text, including word matches denoting abuse or threat
- Emojis
- Images, whether symbolic or text within images
- Voice notes

DEFINITION OF ABUSE

Our definition of abuse is based on the inclusion of a reference, whether express or implied, to any one or more of the following: ethnic origin, colour, race, nationality, religion or belief, gender, gender reassignment, sexual orientation or disability.

As a priority, we look for any threatening comments and those which take a similarly abusive or threatening aspect towards an athlete or an athlete's family.

For this study, we also incorporated additional categorisations of possible abusive terminologies, associated specifically to Athletics. These include narratives around doping allegations and transphobia.

SCOPE

The primary focus of this study has been public posts on Twitter/X and public comments made to selected athletes on Instagram.

Guided by World Athletics, we have worked with a selection of 2,195 athletes and officials / media personalities, from which 1,344 had at least one of an active Twitter/X or Instagram account.

SCALE + COVERAGE

We examined 449,209 posts across Twitter/X and Instagram matching our inclusion criteria.

- Posts mentioning an athlete by handle
- English language or Emoji
- A small amount of Spanish language abuse was also flagged by Threat Matrix
- Posts mentioning 2 or fewer separate handles, to filter out spam and long conversations which often have little relevance towards the individual tagged people

We examined 449,209 posts across Twitter/X and Instagram matching our inclusion criteria.

Most of the abusive posts were detected by our text analysis algorithm, which flags posts on the basis of over 500 keywords, phrases and emojis.

This flagged over 8,321 posts for further review, which were then individually assessed to see if they met abuse criteria.

In addition to this, we use an AI-empowered threat algorithm to determine posts that contained explicitly threatening language.

LANGUAGE

The Threat Matrix service operates a multi-lingual filter set, with specialist keywords (as well as emojis), representing different caterisations of abuse, discrimination and threat. The service includes 32+ languages (including the following):

Argentine Spanish	Filipino	Italian	Serbian
Australian Slang	French	Japanese	Slovakian
Brazilian Portguese	French Canadian	Korean	Spanish
Chinese	Georgian	Norwegian	Swedish
Croatian	German	Persian	Turkish
Danish	Haitian Creole	Polish	Vietnamese
Dutch	Hebrew	Russian	Xhosa
English	Hungarian	Samoan	Zulu