




Professional
Footballers'
Association

ONLINE ABUSE

AI Research Study: Season 2020/21

Powered by

Signify 

Executive Summary

This study has been commissioned by the Professional Footballers' Association (PFA). It was carried out by data science company Signify Group and looks at targeted, abusive messages sent via social media to Premier League, Women's Super League (WSL) and English Football League (EFL) players, and former players from across the top divisions of English Football. This study covered the 2020/2021 season.

The study was developed to illuminate the size, scale and gravity of the issue of online discriminatory abuse being targeted at professional footballers in the UK. It also demonstrates the validity of using machine learning to capture, analyse and quantify online abuse as a scale solution.

Signify's machine learning systems analysed over 6 million incoming messages and ran a deeper analysis of over 20,000 flagged posts identifying 1,781 explicitly abusive messages from 1,674 accounts for further analysis and action.

SUMMARY ANALYSIS

44%

of Premier League players received discriminatory abuse – two in every five players.

48%

increase in unmoderated racist online abuse in the second half of the season.

50%

of abusive tweets from UK based accounts.

33%

of verified abusive accounts have an affiliation with UK clubs.

20%

of all detected abuse was sent to just four players.

33%

of all abusive posts contained homophobic abuse.

From Analysis to Action

1,781

Offensive tweets reported to Twitter for removal.

1,674

Offensive accounts evidenced to Twitter for sanctioning.

367

accounts identified as fans, members or season ticket holders reported to UK clubs.

10

accounts deemed to have passed criminal thresholds reported to the police.

Why is this significant?

The 2020-21 season has been a tumultuous campaign with matches played behind closed doors amid the COVID-19 pandemic and players showing an impressive collective commitment to highlight systematic racism by taking the knee before games. Off the pitch, we were proud to support player activism across a number of national campaigns and initiatives. There was also combined efforts across the football and sporting world to stop online hate and abuse culminating in a 48-hour social media boycott.

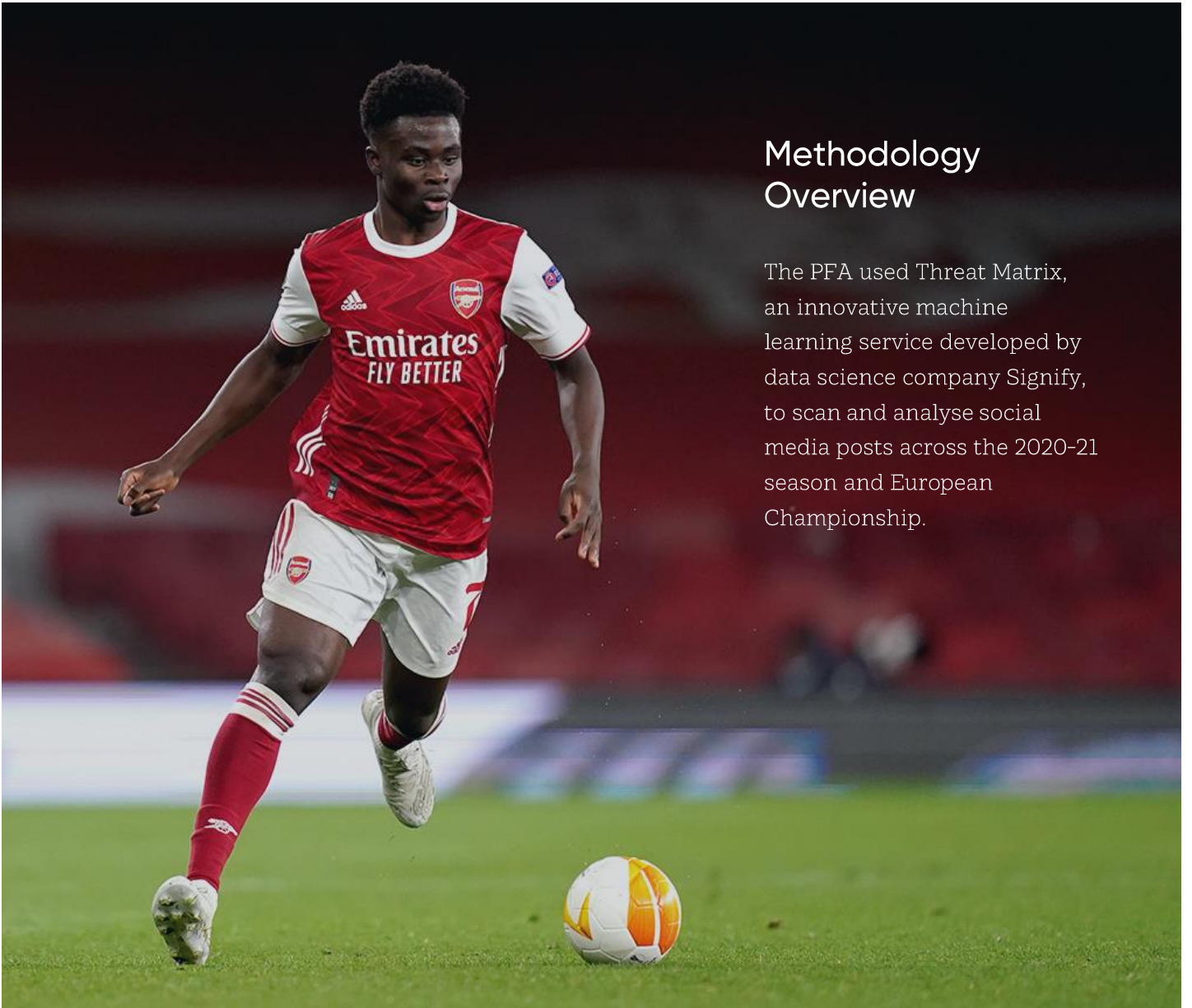
With this backdrop, we ran the most comprehensive study of its kind, covering a full season, across multiple leagues and international tournaments, with a clear focus on targeted discriminatory and abusive posts on Twitter. The results illuminate the true scale of this issue and evidence beyond any reasonable doubt that:

- abusive accounts can be identified and brought to justice
- the issue is not simply an international problem, a majority of the abuse we identified was from UK based accounts where a location was detected
- clubs can be empowered to sanction fans who commit these crimes
- social platforms can and must do more to deal with this problem

Alongside the football authorities and Government, the PFA has been working on this issue across the season. We will continue to work on behalf of our members to effect real change and hold the social platforms to account.

Maheta Molango | CEO
Professional Footballers' Association





Methodology Overview

The PFA used Threat Matrix, an innovative machine learning service developed by data science company Signify, to scan and analyse social media posts across the 2020-21 season and European Championship.

1 Source data

Using AI-powered threat detection algorithm, we scanned millions of public posts on Twitter across the 20-21 season.

2 Clean data

Removing bots, we employ Natural Language Understanding to flag discriminatory abuse being targeted at professional players.

3 Analyse data

Flagged posts are analysed by our team of experts against the FA's Rule E3(2) definition of aggravated breaches, ensuring a level playing field.

4 Evidence + action

Working with the UK Football Policing Unit, social platforms and clubs, we have collated and evidenced the most egregious examples of discriminatory abuse for action.

KEY FINDINGS

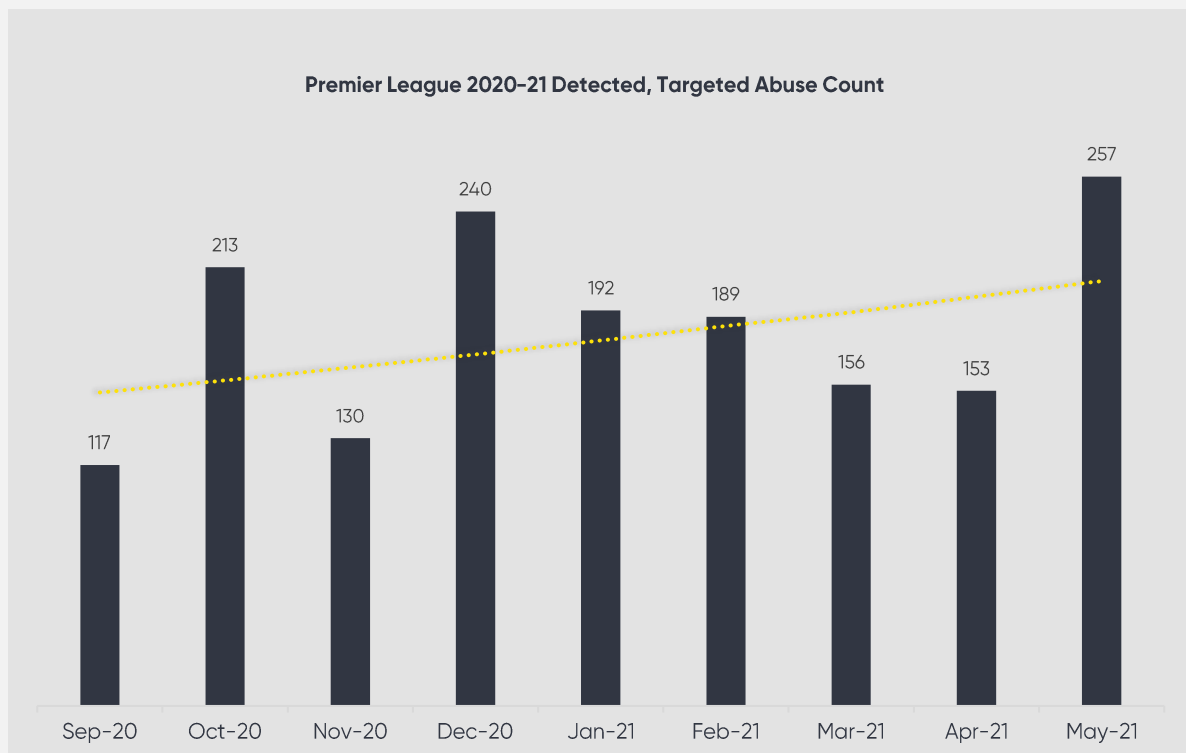
Finding 1: Online discriminatory abuse is getting worse

Our findings show that the problem of discriminatory abuse was worse at the end of the season than at the beginning.

From a source of 6,110,629 posts we identified 1,781 instances of abuse matching our criteria sent from 1,674 accounts.

Across the Premier League we included all 400 players for whom we could find active social media accounts, and we detected abuse towards 176 of them (44%) across the season.

There was no month in which we detected less than 100 pieces of direct, discriminatory abuse and serious threats.

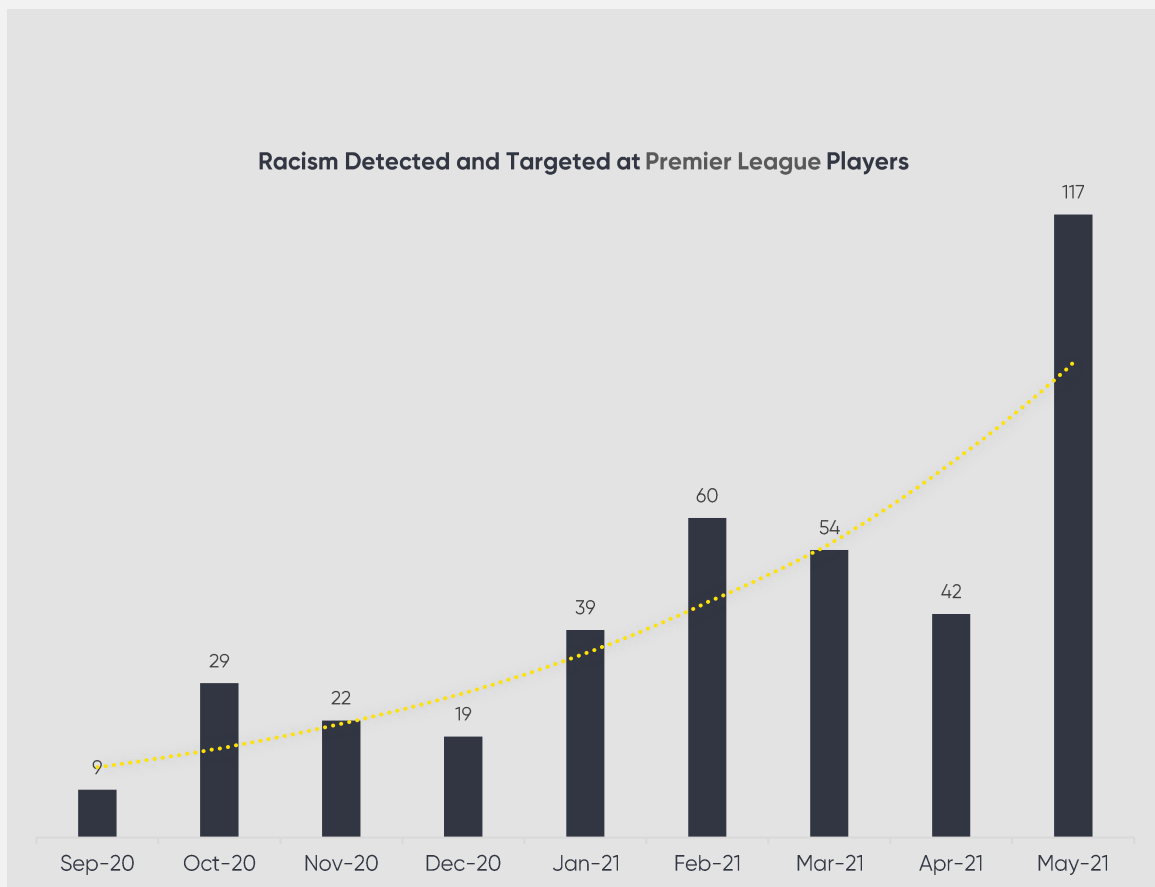


We observed that the abuse we detected fluctuated over time, with May 2021 being the highest point.

Finding 2: Targeted racist abuse peaked in May 2021

May 2021 saw by far the greatest amount of racism detected across our study.

Excluding the final of Euro 2020, the tail end of the domestic season – particularly FA Cup, Europa League and Champions League finals – saw big spikes in racist abuse targeting several players in our study.



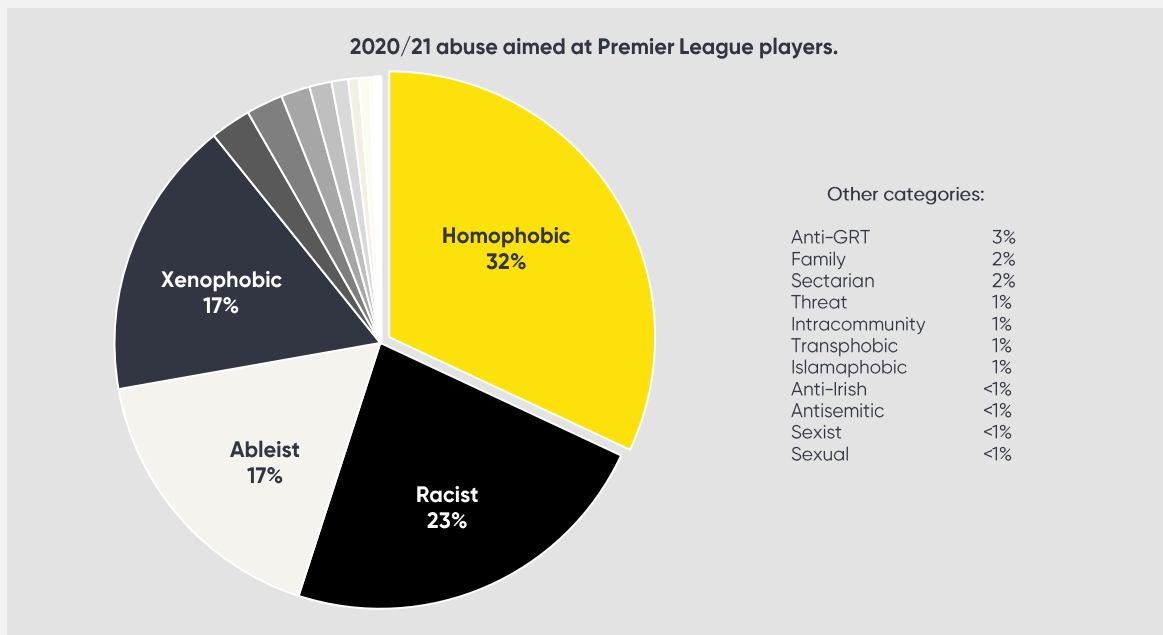
There were several reasons for this, including a core incident involving Chelsea and Leicester City fans associated to the FA Cup Final, and a subsequent League fixture where the two teams met each other. The related incident generated a spike in racist, abusive comments specifically targeting a Leicester City player.

Notably, May 2021 began with the social media boycott, placing the issue of online abuse at the top of the agenda. The boycott weekend itself did see a drop in abusive posts due to the number of players abstaining from platforms for 48hrs, however, the numbers quickly shot back up, and thanks to the incident mentioned, peaked for the season in the same month.

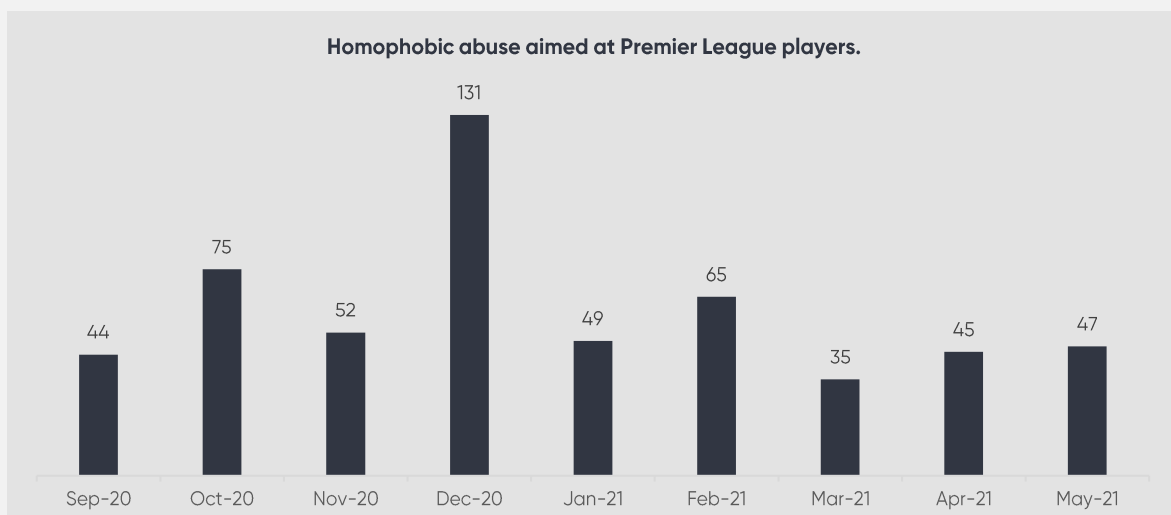
Finding 3: Homophobia is the most common form of online discriminatory abuse targeting UK professional footballers

Homophobic abuse represents nearly a third of all detected abuse.

We identified 543 posts containing homophobic abuse and 392 racist examples. There were also many instances of ableist and xenophobic abuse. However, homophobic content was the largest category of discriminatory abuse.



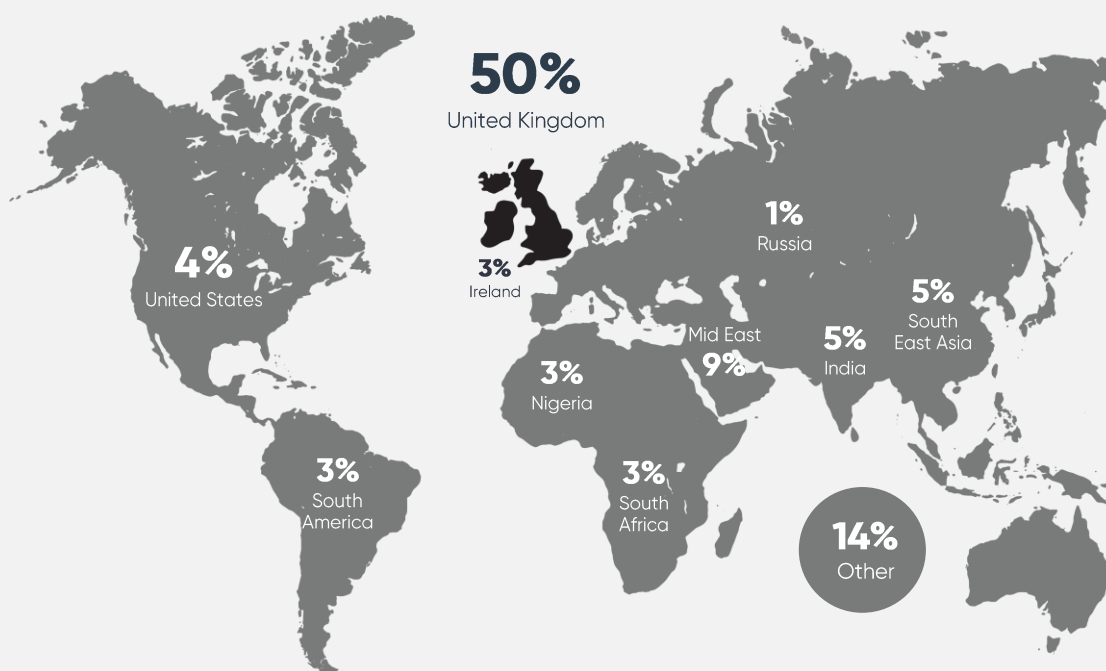
Homophobic abuse – while disconcertingly present throughout the year – peaked in December 2020 with spikes often corresponding with campaigns against homophobia, including the Rainbow Laces / rainbow armband campaigns, with individual players targeted for their support of these causes.



Finding 4: More than 50% of the problem is home grown

From the posts where we were able to clearly identify geographic region of origin of abusive accounts, 50.4% came from the UK.

This is in contrast to narratives that suggest most discriminatory abuse comes from outside the country. Whilst there is a widespread international problem, we detected abusive posts from 53 different countries. The map below indicates volumes from different jurisdictions:



The levels of bots requiring filtering from the data set was relatively low and we saw no evidence of any state-sponsored / coordinated campaigns.

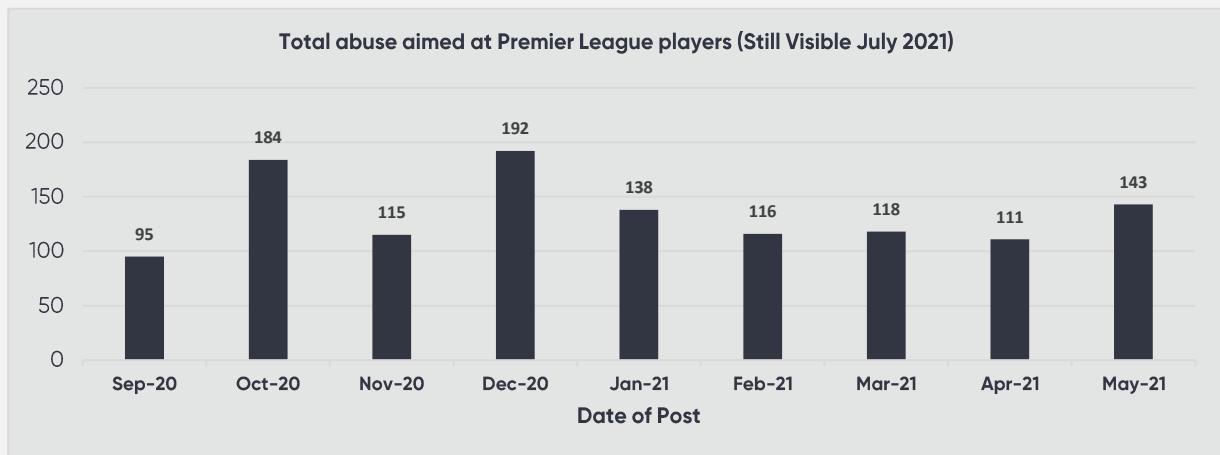
The predominance of abusive accounts within the UK, should mean that there is more potential to take action via the UK legal system (where criminal thresholds can be evidenced to have been passed).

This data also empowers UK football clubs to identify and sanction account owners who may be associated to their club as fans, members or season ticket holders.

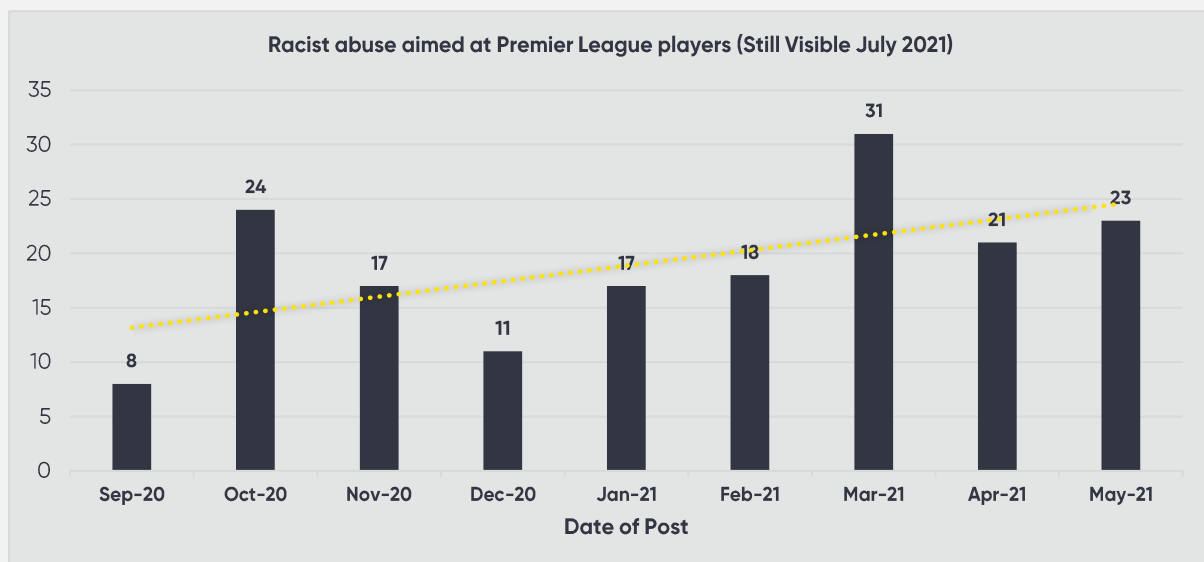
Finding 5: Twitter's moderation services are not working

One increased area of focus for social media platforms has been moderation. Twitter claim to be more effective at detecting and moderating discriminatory posts.

All abusive content (still live as of July 2021) in the tables below were more than one month old, with some abusive posts going back to the beginning of the campaign.



The graph below suggests that racist abuse of Premier League players is getting worse, despite Twitter's efforts at improving moderation capabilities across the season.

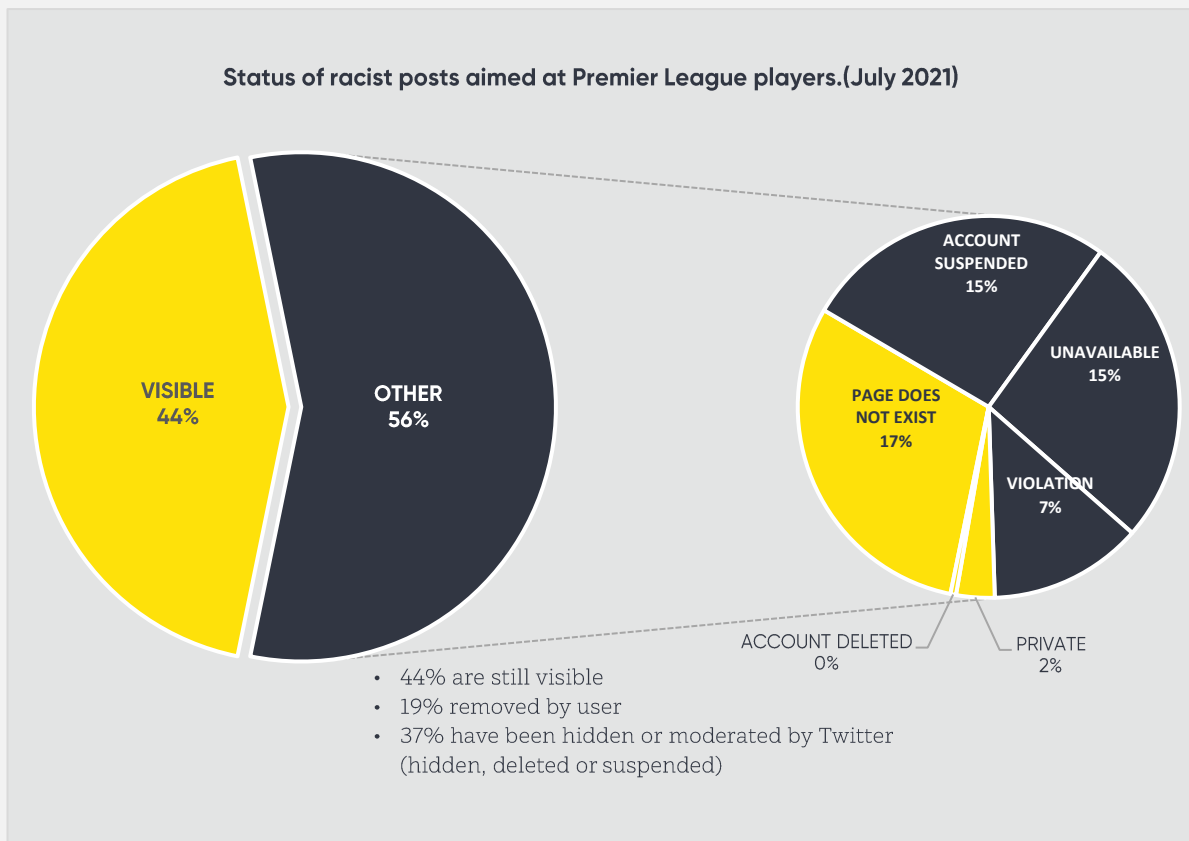


1. We observed Twitter's moderation efforts by selecting posts containing discriminatory abuse (many of which have been reported to the platform). We then checked the status of this content on a monthly basis.
2. Covering the full 2020-21 season until 31st May (two clear days after the Champions League final and Championship play-off final - to allow for any fallout).
3. We analysed which posts were still present on the platform vs those no longer visible.
4. Where unavailable, we analysed the reason for this (Twitter action vs user action).

Finding 6: Moderation of racist posts is not enough

Twitter's activities in moderating abusive and discriminatory posts appear to be heavily focused on posts rather than the source accounts sending them.

44% of identified abusive racist posts from across this season are still visible on the platform. As of July 2021, a little over half (56%) of the racist posts we detected during the season are no longer available. Of the posts that have been removed, a further analysis indicates that only 37% of these were likely to have been actioned by the platform, with 19% likely to have been removed directly by the user.



The vast majority of accounts sending these posts are still live and have remained unsanctioned.

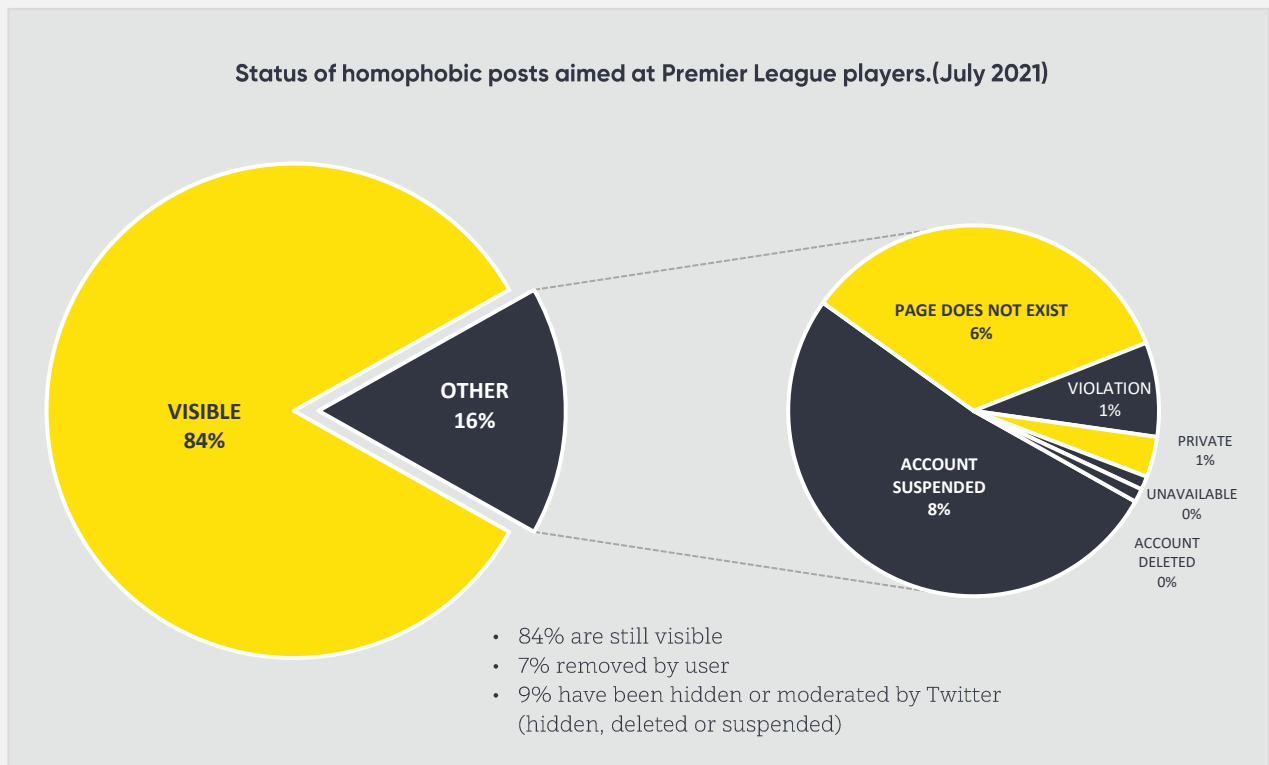
As of July 2021, from the 359 accounts (who sent 391 abusive posts), 288 accounts were still on the platform.

This demonstrates that whilst Twitter is attempting to moderate some racist posts, it is not sanctioning and banning the source accounts behind them.

Finding 7: Other forms of discriminatory abuse

As of July 2021, more than four out of five targeted homophobic messages identified across the season are still live on Twitter.

While the standard of Twitter's moderation around racist abuse remains lacking, other forms of discriminatory abuse also appear to go unchecked.



Racism is one of many forms of discrimination that affects UK footballers (23% of abusive posts we detected targeting Premier League players across the season were racist in nature). We have analysed Twitter's moderation activities to identify how focused they are on additional issues like homophobia. As displayed in the chart above, the platforms hit rate for removing homophobic content is much lower (16% vs 56% for racist content). We see that moderation has been similarly ineffective for other major categories such as ableism and xenophobia.



Above: Examples of tweets targeting Premier League players with homophobic content left live on platform

EXPLAINER

How effective are Twitter at taking down offensive tweets and accounts?

This study has analysed Twitter's capability to remove offensive tweets and accounts by reviewing the seven possible statuses displayed when attempts are made to access posts that are no longer available. This methodology allows us to hold the platform to account in terms of actions being taken.

STATUS LIKELY ACTIONED BY TWITTER

- **ACCOUNT SUSPENDED:** The poster has been temporarily or permanently banned from the platform, whether for this or other posts.
- **VIOLATION:** The Tweet has been removed for violating Twitter's rules.
- **UNAVAILABLE:** The post is hidden from view. This happens often for posts mentioning certain words that can be seen as slurs. This appears to be done automatically without any human oversight.

STATUS LIKELY ACTIONED BY ACCOUNT OWNER

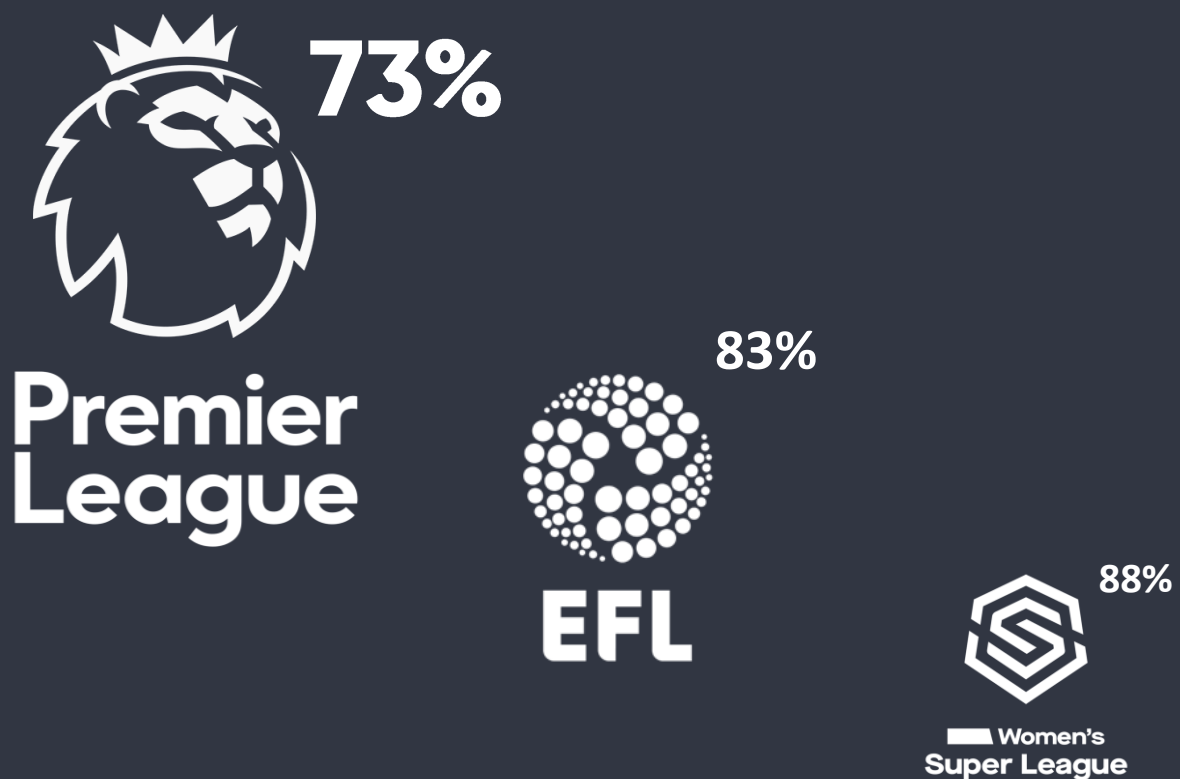
- **VISIBLE:** Post is still online, and entirely visible to the public at time of checking.
- **PRIVATE:** The post is unavailable because the poster has made their own account private.
- **PAGE DOESN'T EXIST:** The post is unavailable, with no given reason. This includes posts deleted by the user themselves.
- **ACCOUNT DELETED:** The post is unavailable as the poster has deactivated their own account.

Finding 8: Twitter is only focused on the Premier League

The PFA represents professional footballers across the England. This study has assessed online abuse targeting players across the EFL and WSL, as well as those playing in the Premier League.

Twitter appears to apply a hierarchical order to its moderation activities with the following action taken against discriminatory abusive posts that we identified across the three leagues:

- 27% of abusive posts targeting Premier League players were detected and are no longer visible
- This drops to 17% for posts targeting EFL Players
- Only 12% for those targeting players from the WSL



Percentage of targeted abusive posts identified in this study, still visible on Twitter (July 2021)

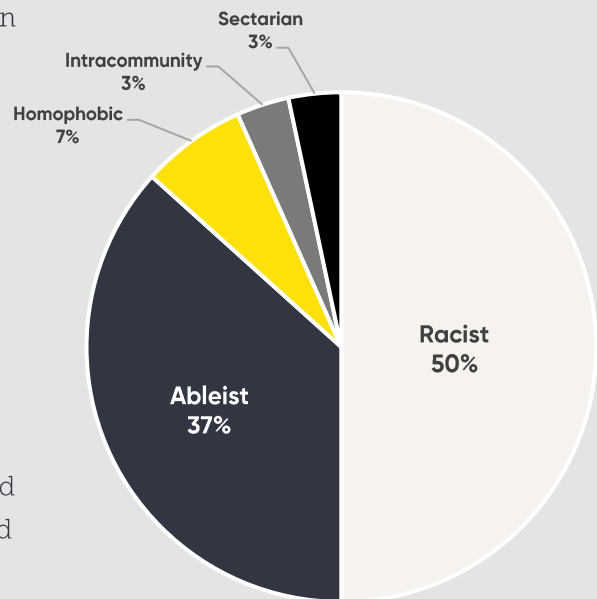
Finding 9: Different leagues face different challenges

Players in the EFL and WSL are subject to different types of discriminatory abuse.

With fans not in stadiums for most of the season and relatively diluted television coverage, we saw less attention directed at EFL players and linked to that was a lower level of abuse.

Our EFL study encompassed 97,806 posts at 53 selected EFL players, of which 29 were deemed to meet our threshold for abuse.

Half of all identified abuse targeting EFL players was racist. This is in comparison to the high percentage of homophobic abuse identified across the Premier League, illustrating a marked difference in issues faced by professional footballers playing in the EFL.



Types of abuse at EFL players, 2020-21

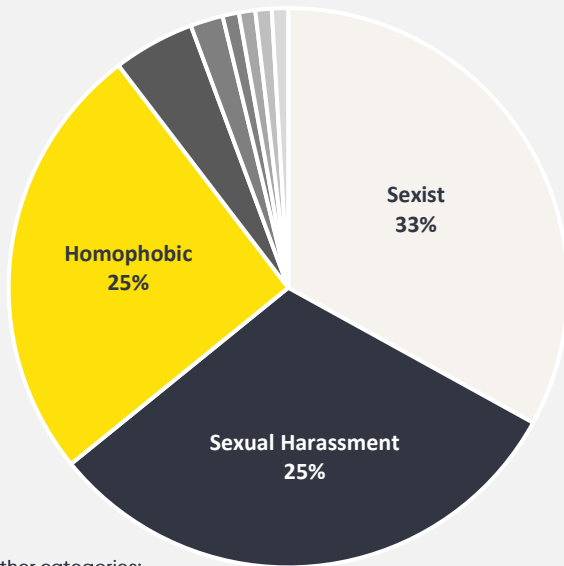


Now is the time for change. If we have this kind of technology at our disposal, why aren't social media companies using it to eliminate racist and discriminatory abuse?



Rio Ferdinand | Football Pundit | Former Manchester United & England Player

Types of abuse targeted at WSL Players



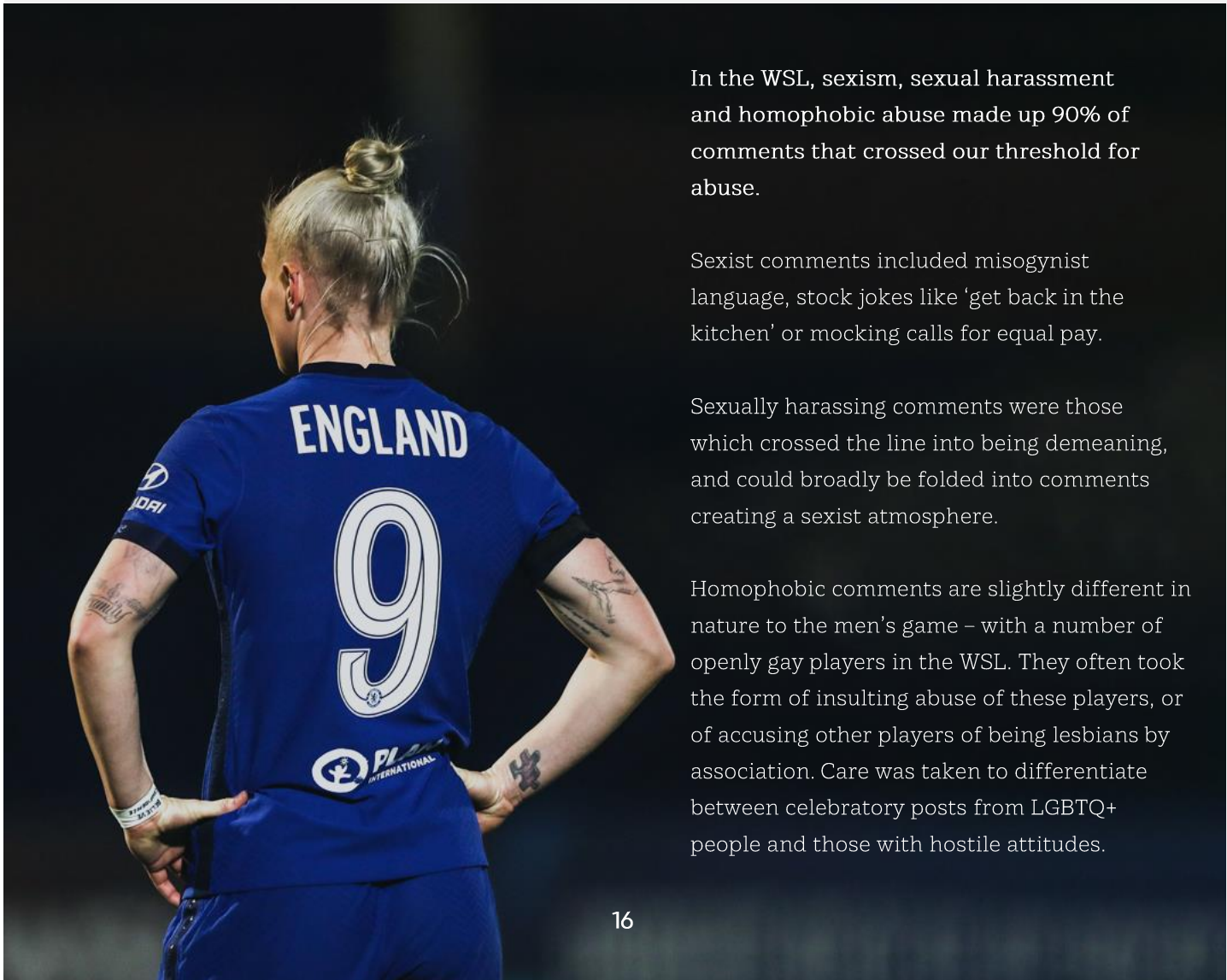
Other categories:

Ableist	5%	Racist	1%	Transphobic	1%
Sectarian	2%	Xenophobic	1%	Family	1%

Over the period of study, we detected 105 pieces of abuse matching our criteria from a source data set of 249,103 tweets.

Players at 11 of the 12 clubs in the WSL were affected by discriminatory abuse online.

All 246 WSL players with an identifiable Twitter account are covered in the study, of which we found 38 (15%) received some kind of targeted abuse.



In the WSL, sexism, sexual harassment and homophobic abuse made up 90% of comments that crossed our threshold for abuse.

Sexist comments included misogynist language, stock jokes like 'get back in the kitchen' or mocking calls for equal pay.

Sexually harassing comments were those which crossed the line into being demeaning, and could broadly be folded into comments creating a sexist atmosphere.

Homophobic comments are slightly different in nature to the men's game – with a number of openly gay players in the WSL. They often took the form of insulting abuse of these players, or of accusing other players of being lesbians by association. Care was taken to differentiate between celebratory posts from LGBTQ+ people and those with hostile attitudes.

Finding 10: Racism waits for its moment

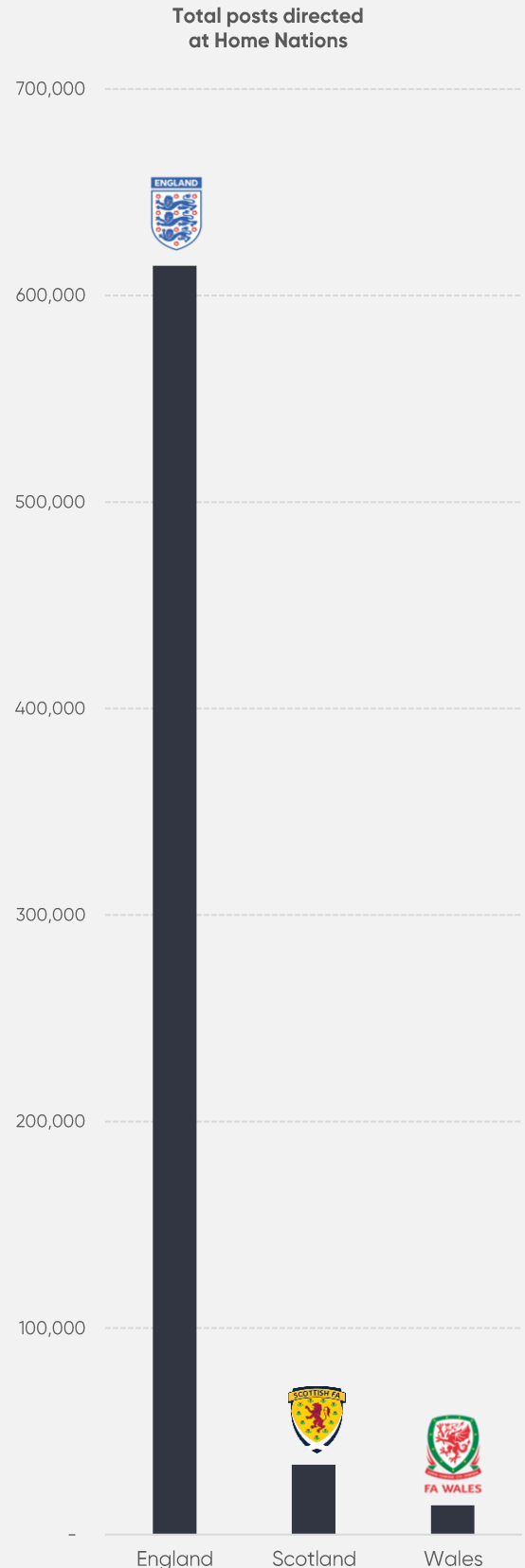
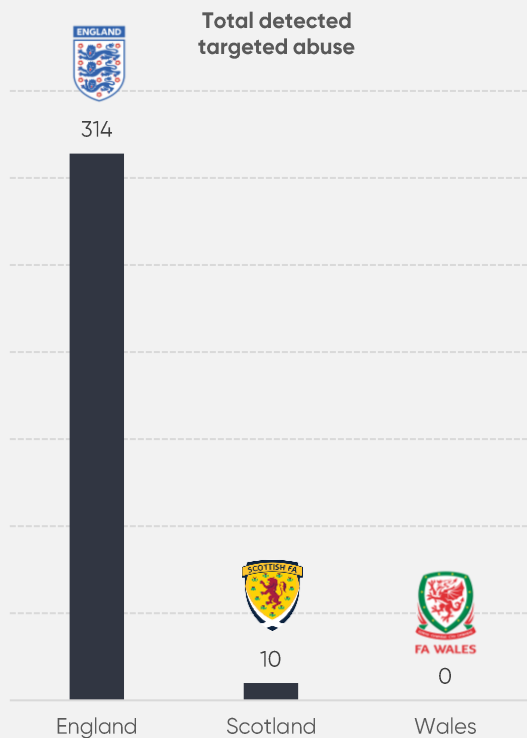
England, Wales and Scotland at Euro 2020

To ensure the most comprehensive, full-season dataset and insights, this study was extended beyond the domestic season and into the European Championship tournament held one year late in the Summer of 2021.

With a remit to cover all three home nations teams taking part in the tournament, we set up further monitors for the English, Scottish and Welsh squads.

This included over 650,000 social media posts. Within this data, we found 324 abusive posts that crossed the threshold of being discriminatory or threatening after reviewing over 1,500 flagged posts mentioning players' handles.

These were split across national teams similarly.



The vast majority of abusive posts contained either racism or homophobia, or in some cases both (85%).

Homophobic Abuse

As a continuation of what we observed across the Premier League season, homophobic abuse was constant form of abuse identified throughout the Euro's - appearing in our results to some degree on 30 of the 43 days covered in this part of the study.

The biggest spike occurred when players were engaged in their pre-tournament media responsibilities..

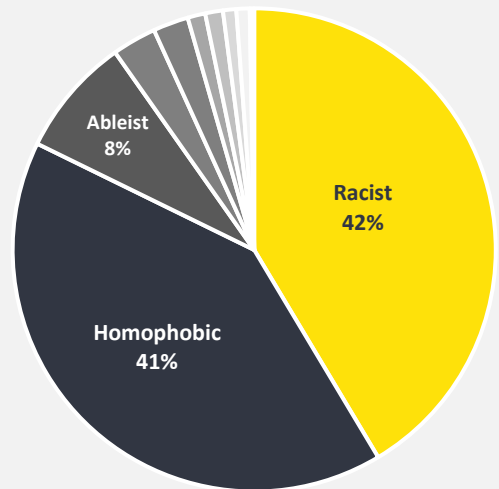
Racist Abuse

Racist incidents tended to occur in larger spikes, evident on nine separate occasions and increasing as the tournament progressed.

The first racist incidents we detected towards home nations players were around the England v Scotland game.

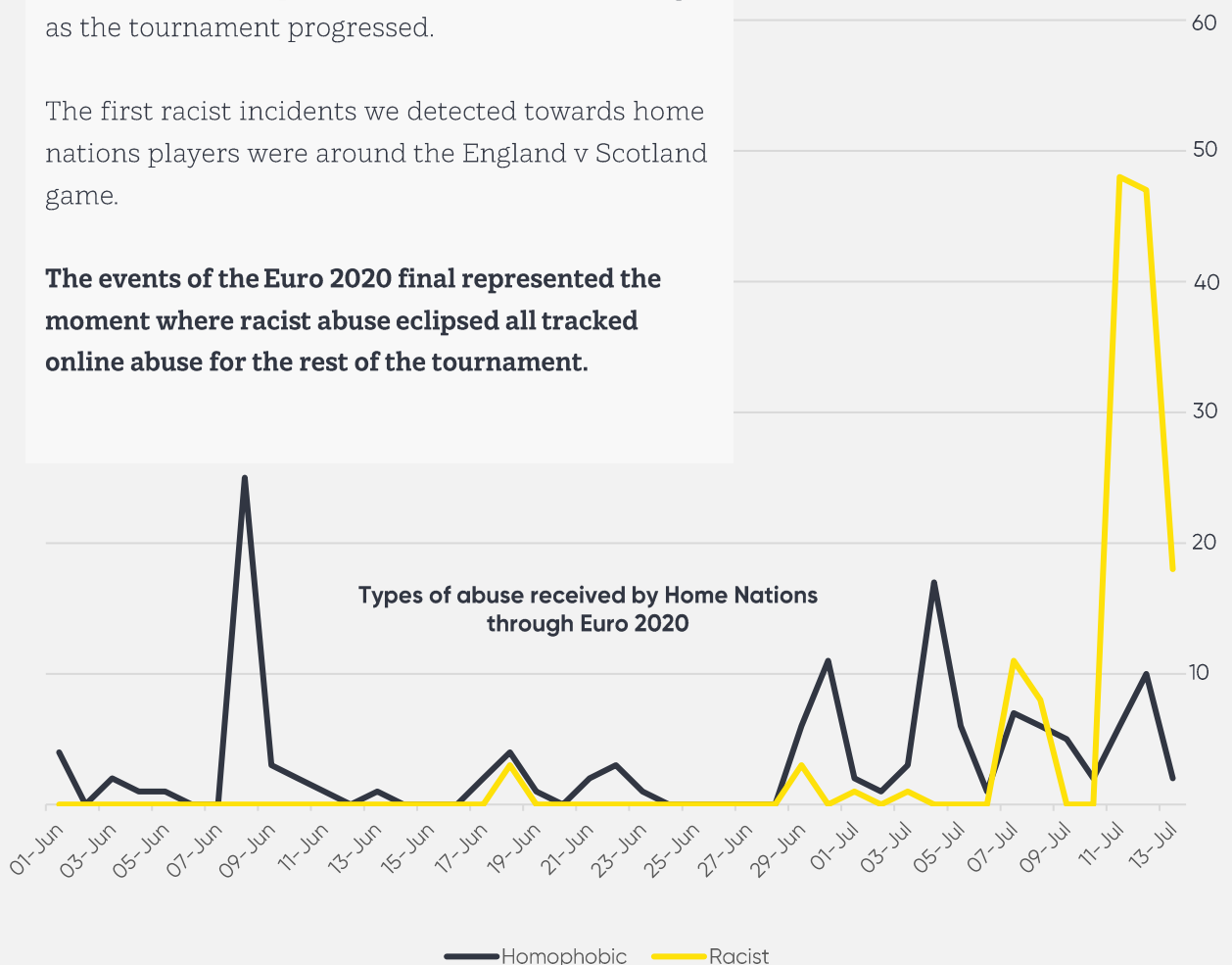
The events of the Euro 2020 final represented the moment where racist abuse eclipsed all tracked online abuse for the rest of the tournament.

Types of abuse at Euro 2020



Other categories:

Threat	3%	Xenophobic	1%
Sectarian	2%	Anti-GRT	1%
Anti-Irish	1%	Family	<1%
Transphobic	1%		



TAKING ACTION

Over the course of the season, we have been taking action on behalf of our members and working with our partners across the football authorities. Using this data, we hope to continue our engagement with the UK Government, police and football authorities to help tackle this issue. This includes investing in AI research and monitoring, as well as education initiatives and player support.

Social media platforms

We have worked with the social media companies to help them understand the pain and damage being caused by content coming from their platforms. Racist abuse causes trauma. It impacts the targeted players, their teammates and we know it will also affect their peers. It causes hurt to all other fans who view online hate, and it will inevitably live with the aspiring generation of young players.

The data in this report evidences how the interventions from social media companies are insufficient, allowing racist abuse to thrive on its platforms for months, in some cases even after being reported.

We have called for social platforms to permanently ban all offending accounts and proactively compile evidence to give to the police to pursue prosecution. We know more can be done – and through this report and our activities, we have proven it.

UK Football Policing Unit

The PFA cooperate closely with the UK Football Policing Unit (UKFPU) who have been working to bring offenders to justice. We have shared our findings from this study with their team and continue to highlight evidence and examples of abusive accounts (and their tactics) on a regular basis.

1,781

Offensive tweets reported to Twitter for removal.

1,674

Offensive accounts evidenced to Twitter for sanctioning.

367

accounts identified as fans, members or season ticket holders reported to UK clubs.

10

accounts deemed to have passed criminal thresholds reported to the police.

Working with Clubs

As clubs standardise their processes to incorporate and deal with abusive account holders who are associated to their clubs, we stand by our partners across the football pyramid, sharing data and ensuring education and safeguarding is in place for players on the specific issue of online abuse.

Supporting our members

For any current or former players who receive discriminatory abuse, the PFA are here to help. From providing wellbeing support to guiding you through the process of reporting an incident, our Equality, Diversity and Inclusion team have vast experience in this area. For more information, contact: enough@thepfa.com

Working with Government and Legislators

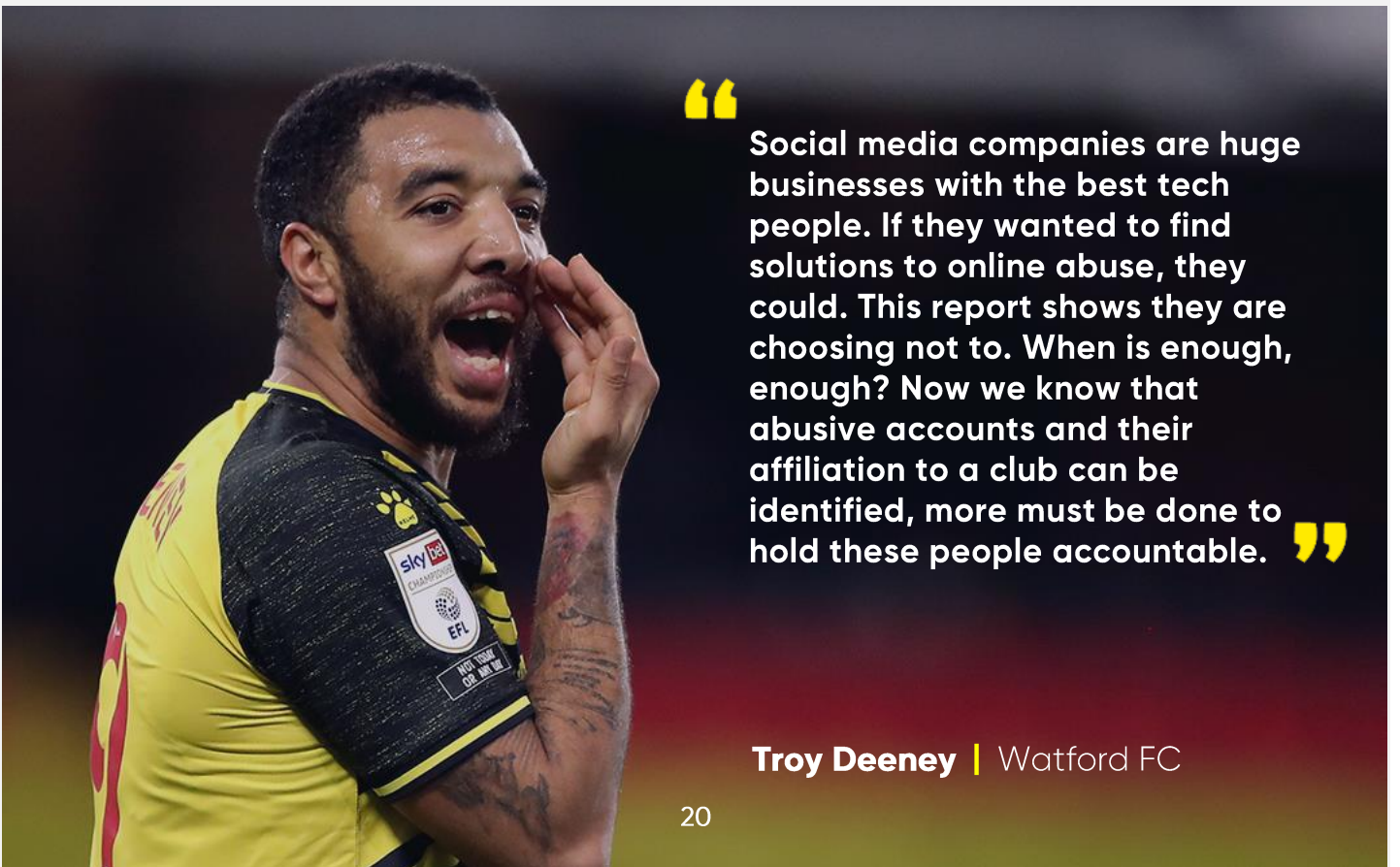
The PFA continue to support and shape the legislation needed to ensure that perpetrators face real-life consequences for online abuse and that. We have been feeding our data insights and knowledge to UK policymakers via DCMS and Government representatives.

As the Online Safety Bill passes through Parliament we will continue to provide insights from studies such as this one, alongside testimonies from our members to underline the impact of online abuse.



Social media companies are huge businesses with the best tech people. If they wanted to find solutions to online abuse, they could. This report shows they are choosing not to. When is enough, enough? Now we know that abusive accounts and their affiliation to a club can be identified, more must be done to hold these people accountable. ”

Troy Deeney | Watford FC



CONCLUSIONS

The findings in this report present a worrying picture and validate key concerns raised by players throughout the season; despite more attention and calls for action on online abuse, the reality (evidenced in this study), is that instances of targeted online abuse have increased.

The measures being taken to deal with this issue are yet to prove effective. A lack of transparency and consistency in the application of these solutions is making it harder for workable solutions to be implemented.

Much of the debate around tackling online abuse seems to be founded on the assumptions that... you cannot identify account owners; that real world action is complicated; offending accounts are mainly outside the UK; abusive messages are not published by 'real' fans and that racism is the only discriminatory issue. The findings of this report show that over the duration of an entire season these assumptions are not borne out.

In this study we have demonstrated:

- how to detect targeted abuse and verify those who are publishing it, empowering authorities, and clubs to apply real world consequences for online action.
- that significant amounts of online abuse originate in the UK, including from season ticket holders, regular attendees, and members of fan organisations, all of which have access and contact with players, club facilities and staff.
- that there are multiple forms of serious abuse targeted at players and that tackling the impact of online abuse will require action across multiple types of discrimination.

The PFA will use these findings to continue to press platforms, partners and government to take action and help change the direction of travel, improving the online experience for players and fans alike.

**For football.
For life.**

APPENDIX A: METHODOLOGY

This study focuses on discriminatory abuse and threats in the format of:

- 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 FA's Rule E3(2) which defines an aggravated breach:

A breach of Rule E3(1) is an "Aggravated Breach" where it includes 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.

To this, we added threatening comments and those which take a similarly abusive or threatening aspect towards a player's family.

This benchmark of what is determined to be abusive is essential for our work ensuring a threshold that is currently applicable to every Professional footballer in the UK.

Scope

The primary focus of this study has been Twitter, due to the availability of public data on the platform. Instagram data has also been used in parts and has contributed to our investigatory work and evidence gathering.

Across the Premier League, WSL and the EFL we included over 700 accounts of players for the 2020/21 season.

We included a further 55 accounts for players in home nations teams attending Euro 2020.

Scale + Coverage

We examined over 6 million posts matching our inclusion criteria.

- Posts mentioning a player by handle
- English language or Emoji
- Posts mentioning 2 or fewer separate handles, to filter out spam and long conversations which often have little relevance towards the individual tagged people.

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 16,000 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 pick out the meaning in threatening language.

Much of the terminology in football could be used in a threatening sense (e.g. 'shoot') but will not be so in the vast majority of cases.

Our threat algorithm detects the linguistic relations in a sentence and can help us find the difference between 'I'm going to scream if X doesn't shoot' and 'I'm going to shoot X'.

Once abuse is detected, this was overlaid with a club affiliation algorithm to help identify abusers who show particular support to a UK based football club, likelihood to attend or even own a season ticket.

Report team

This report was commissioned by the PFA. The data gathering and analysis was conducted by ethical data science company Signify Group (www.signify.ai). With a specialist capability in the identification of hate speech and social media abuse, Signify have worked with governing bodies and clubs in professional football, and have a proprietary AI driven monitoring service to protect clubs, players, officials and fans from online abuse www.threatmatrix.ai.