ONLINE HATE SPEECH TARGETING THE ENGLAND AND WALES MEN'S FOOTBALL TEAMS DURING THE 2022 FIFA WORLD CUP

Arron Cullen and Matthew Williams

MARCH 2023

HATE LAB

KEY FINDINGS

- 69 per cent of the England team and 26 per cent of the Wales squad received targeted hate speech on social media during the tournament
- The prevalence of abuse and hate speech was highest on the Twitter platform compared to the Reddit and 4Chan platforms
- 88 per cent of the hate speech identified was anti-LGB and 11 per cent anti-black
- 78 per cent of the hateful Twitter posts were sent by users identifying as male and 2 per cent as female (20 per cent were unidentifiable)
- 20 per cent of hate speech posts originated from Twitter accounts in the United Kingdom and 16 per cent from other countries (64 per cent were undisclosed)
- In the United Kingdom, the main hate hotspots were identified as London, Manchester, Milton Keynes, Sheffield, and Cardiff
- Among the accounts posting hateful messages, only 5.9 per cent were bots or fake accounts
- Online hate speech on Twitter received an average of 0–2 retweets or replies, which implies minimal public engagement
- 88 per cent of hate posts remain live on social media platforms

69%
of the England team received hate posts

26%
of the Wales team received hate posts

88%
of hate posts were anti-LGB

11%
of hate posts were anti-black

8%
of hate posts were sent by men

20%
of hate posts were sent from the UK

88%
of hate posts remain live on platforms

SUMMARY

This report presents an analysis of 847,370 social media posts directed towards England and Wales men's national football players during the 2022 FIFA World Cup. The findings provide insights into the nature, prevalence and drivers of online hate speech targeting players on social media during the competition.

Using HateLab's award-winning algorithms for detecting hate speech, a total of 198 posts were identified as either racist, homophobic or transphobic. While this finding highlights the ongoing issue of online abuse directed at professional football players, it is also noteworthy that these hateful messages were outnumbered by 362,163 posts containing positive messages in support of the home nations teams.

The majority of hate posts identified in this study remain online, indicating that platform moderation processes remain ineffective. Non-removal solutions, such as the use of community based counter-speech, remain an effective but underused tool in the fight against online hate speech.

THE STUDY

The aim of this study was to highlight the continuing problem of online hate speech in football. Campaigns spearheaded by Kick It Out, Show Racism the Red Card, Rainbow Laces and EE (Hope United) continue to raise awareness of diversity in the game and promote a culture of tolerance and inclusion amongst football fans. In addition, these campaigns strive to educate fans, players, and the broader public on how to recognise and react to hate speech.

A **HateLab study** in 2021 evidenced that black players from the England men's team were sent thousands of racist hate speech posts in the UEFA Men's Euro 2020 final, and a **second study** in 2022 found that 92 per cent of the England women's national football team received misogynistic and homophobic hate posts during the Women's Euro 2022 tournament.

This present study examines hate speech that targets male footballers in the England and Wales national teams during the 2022 FIFA World Cup competition. It is imperative to understand the prevalence and patterns of hate speech targeting professional footballers in order to develop strategies to combat it to ensure that players are protected from such abuse during future tournaments.

The social media platforms examined for the study were Twitter, Reddit and 4Chan. The analysis reveals the types and frequency of online hate players receive, and pinpoints the main trigger events, both on and off the pitch, during the tournament.

HATE

HateLab adopts Ofcom's definition of hate speech: "hateful, offensive or discriminatory content that targets a group or person based on specific characteristics like race, religion, disability, sexuality or gender identity".[1] Most of the hate posts identified by HateLab AI likely fall into the sub-criminal 'legal but harmful' category of abuse, recently dropped from the UK Online Safety Bill. While these posts do not reach the criminal threshold of grossly offensive or threatening and likely to stir up hatred, they are nonetheless harmful to the targeted players if seen, and the wider community.

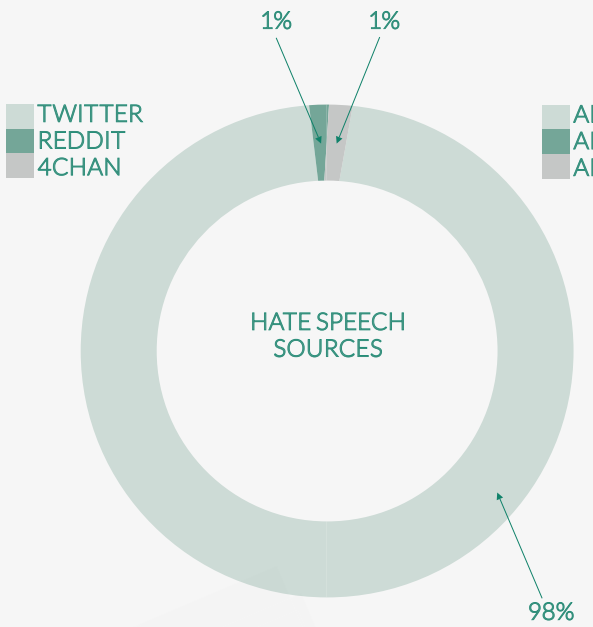

FIGURE 1

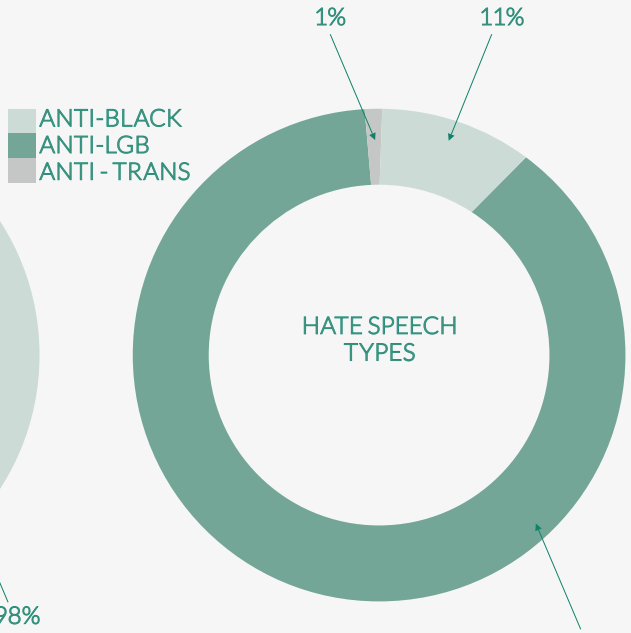

FIGURE 2

A total of 198 social media posts sent during the tournament were identified as hate speech. Figure 1 shows that 98 per cent of these posts were on Twitter, 1 per cent on 4Chan and 1 per cent on Reddit. Figure 2 illustrates the split of hate speech by types, with 88 per cent of posts classed as anti-LGB, 11 per cent anti-black and 1 per cent anti-transgender.

[1] Ofcom (2022) The Online Experiences Tracker (2021/22): Summary Report, Ofcom, London.

"HOMOPHOBIC HATE REMAINS A SIGNIFICANT PROBLEM"

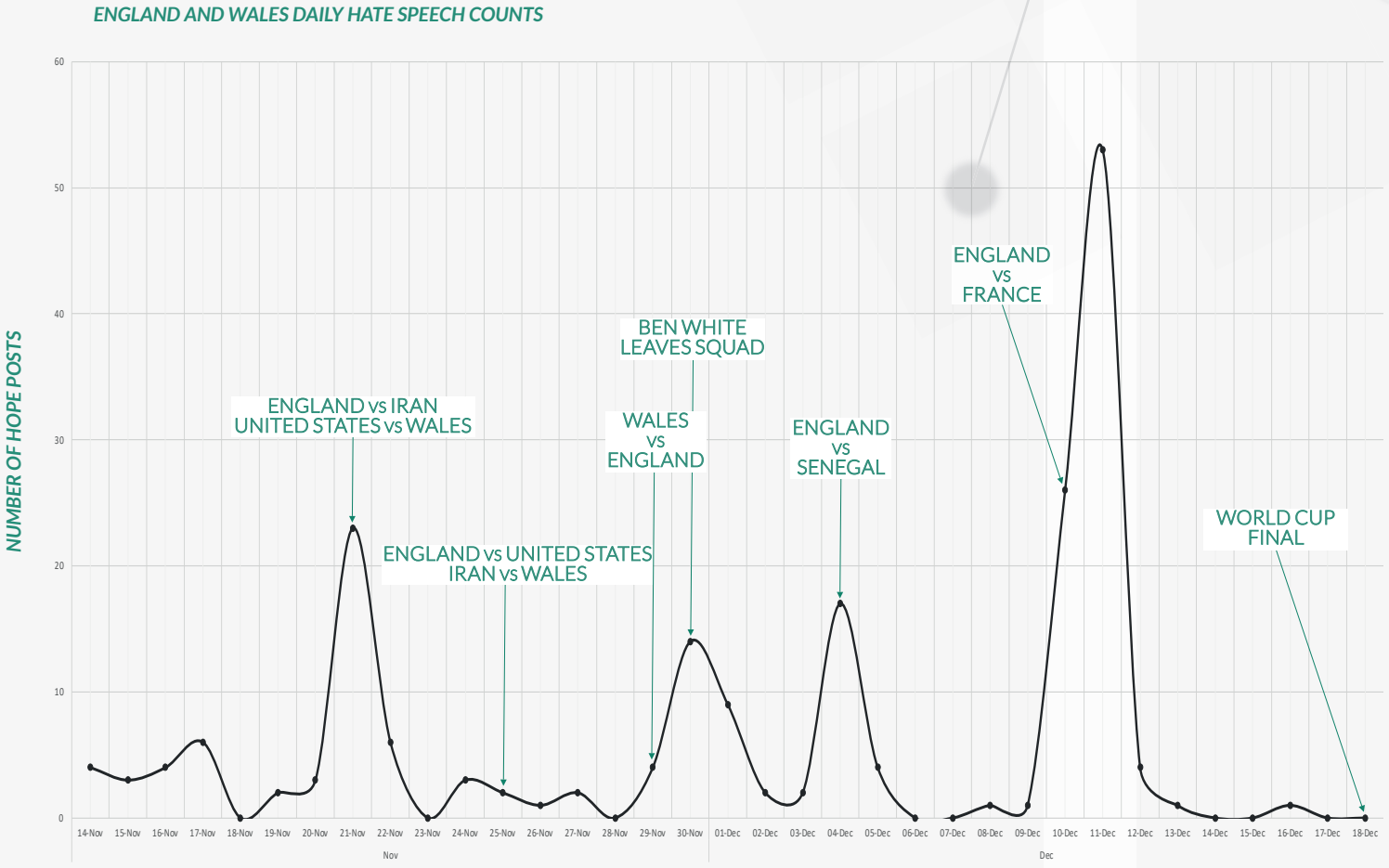

FIGURE 3

The posting of hate speech tended to increase during tournament matches, mirroring HateLab's previous football monitoring exercises (see Figure 3). Notable peaks emerged during the England versus France game, especially in the immediate aftermath, and when England defender Ben White left the camp to return home. This indicates that hate speech continues to be driven by temporal forces, in particular matches and off pitch activity that act as 'trigger events'.

Figure 4 shows the number of hate posts by targeted player. 69 per cent of the England team and 26 per cent of the Wales team were sent hate speech on social media during the tournament. The players who received the most hate speech across the analysis period in the England team were Harry Kane, Ben White, and Mason Mount. Gareth Bale received the most hate in the Wales team.

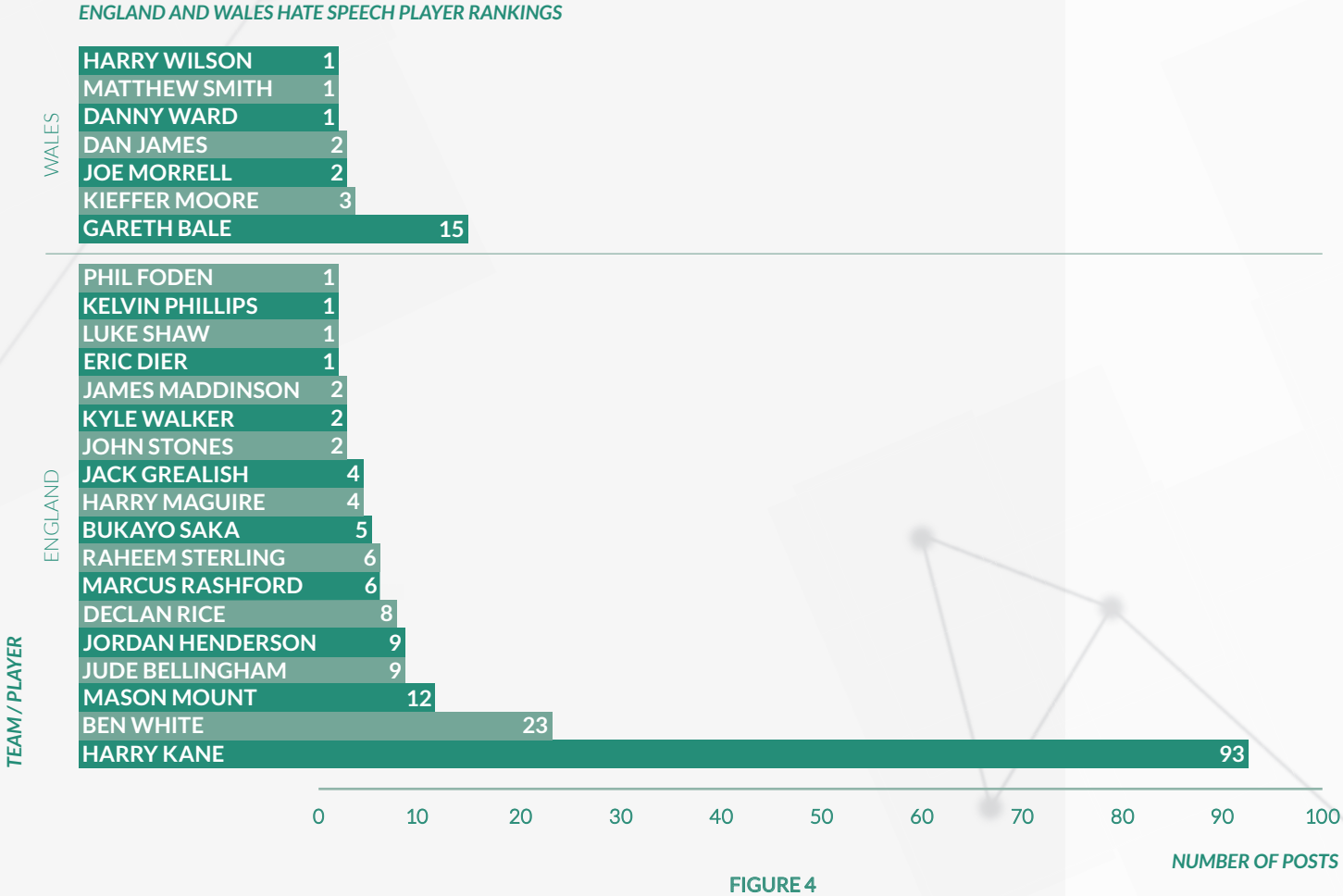

FIGURE 4

FIGURE 5

The hate speech narratives aimed at both teams were broadly similar throughout the tournament. Figure 5 shows that taking the knee, which both England and Wales teams promoted as a sign of solidarity in the fight against discrimination, triggered the posting of racist hate speech. But predominantly, narratives were homophobic in nature, clustering around press coverage of anti-LGBTQ+ laws in Qatar and the banning of the 'OneLove' armband.

England Captain, Harry Kane received the majority of homophobic posts. It is likely Kane was most targeted due to his missed penalty and his prominent role as skipper. The data also suggest some fans felt Kane's focus on the OneLove armband ban early in the tournament detracted from the football.

Kane does not identify as gay or bisexual, indicating the use of homophobic language against him was not targeted at his identify, but was instead being used to generally insult him. While some players may not be severely impacted by the posting of homophobic language, its casual use creates a hostile environment for LGBTQ+ people online that can induce anxiety and fear in some fans.

HateLab research has found that social media users are increasingly adopting the use of emojis to convey hatred in 'code' that can evade automated moderation systems used by platforms. For some time, Facebook, Instagram, and Twitter algorithms and moderation teams failed to recognise the use of certain emojis in certain contexts as potentially abusive. For example, the presence of a primate emoji in a post, while not always racist, can become read as racist by the receiver given a conductive context, such as a trigger event. Furthermore, the addition of emojis that convey disgust or anger, can hammer home the intended hateful meaning. HateLab machine learning algorithms recognise emojis in context, allowing us to classify hate speech more precisely and accurately.

FIGURE 6

Figure 6 presents the most commonly used emojis in hate speech posts. The LGBTQ+ rainbow flag was present in many homophobic posts, often accompanied by laughing face, the middle finger, feeling sick, vomiting, angry face, and clown emojis. The abundance of emojis in hate posts is a worrying issue, as studies indicate that their usage can intensify emotional response and arousal beyond the effects of text alone.

"THE ENGLAND TEAM RECEIVED THE MOST HATE SPEECH"

HATE ACCOUNTS

Metadata from hate posting accounts was extracted to identify account creation date, sex of the user, location, tweet count and follower/followee count. Figure 7 shows that 96 of 174 hate posting accounts were created between 2020 and 2022. The remaining 78 were created between 2008 and 2019. This finding suggests that most of the hate posting accounts on Twitter are new to the platform, but that some longer-established accounts were still involved in perpetrating hate towards players.

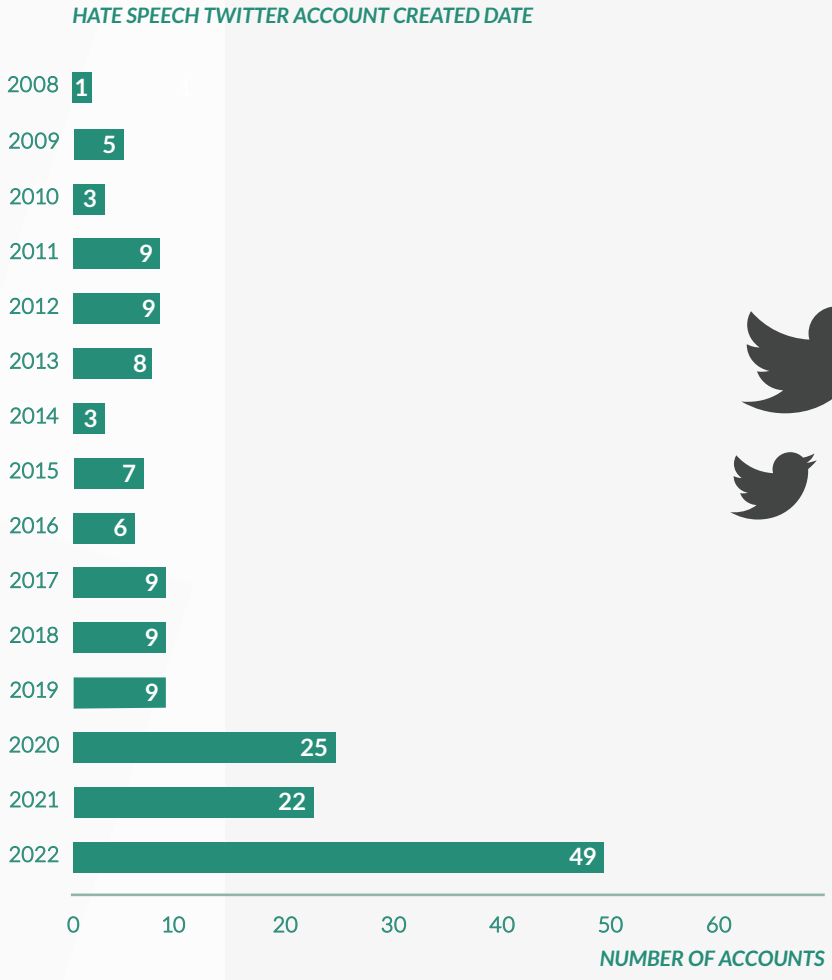

FIGURE 7

Figure 8 shows that of the total number of Twitter hate posting accounts, 78 per cent were run by users with a male name/photo, 2 per cent were run by users with a female name/photo, and 20 per cent could not be classified as any sex based on name/photo.

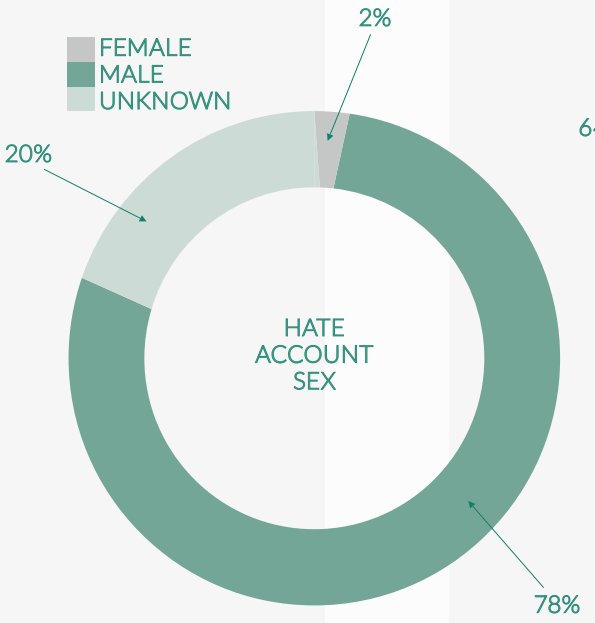

FIGURE 8

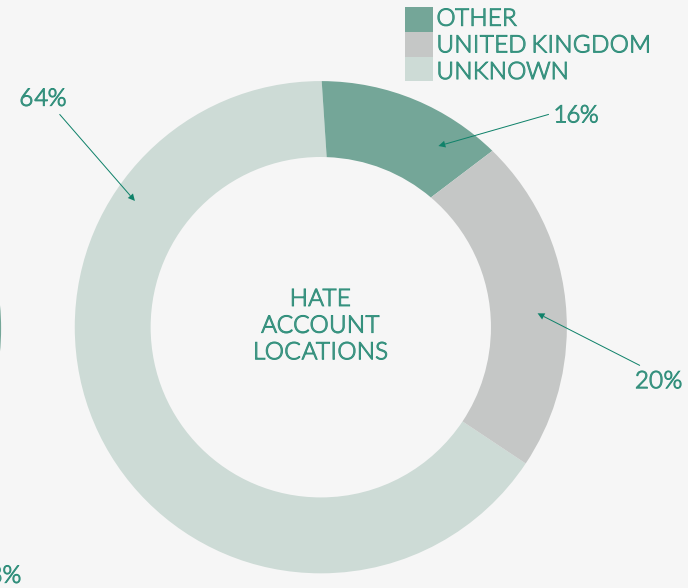

FIGURE 9

Of the total Twitter hate posting accounts, 20 per cent claimed to be based in the United Kingdom, 16 per cent claimed to be from other locations such as North America, Europe, and Africa, and 64 per cent did not reveal their location (see Figure 9).

Within the United Kingdom hotspots for hate posting were located in London, Manchester, Milton Keynes, Sheffield, and Cardiff (see Figure 12).

A **study conducted by HateLab** previously investigated the prevalence of hateful abuse directed towards ethnic minority players during the Euro 2020 championship, as well as in the Premier League over the past decade.

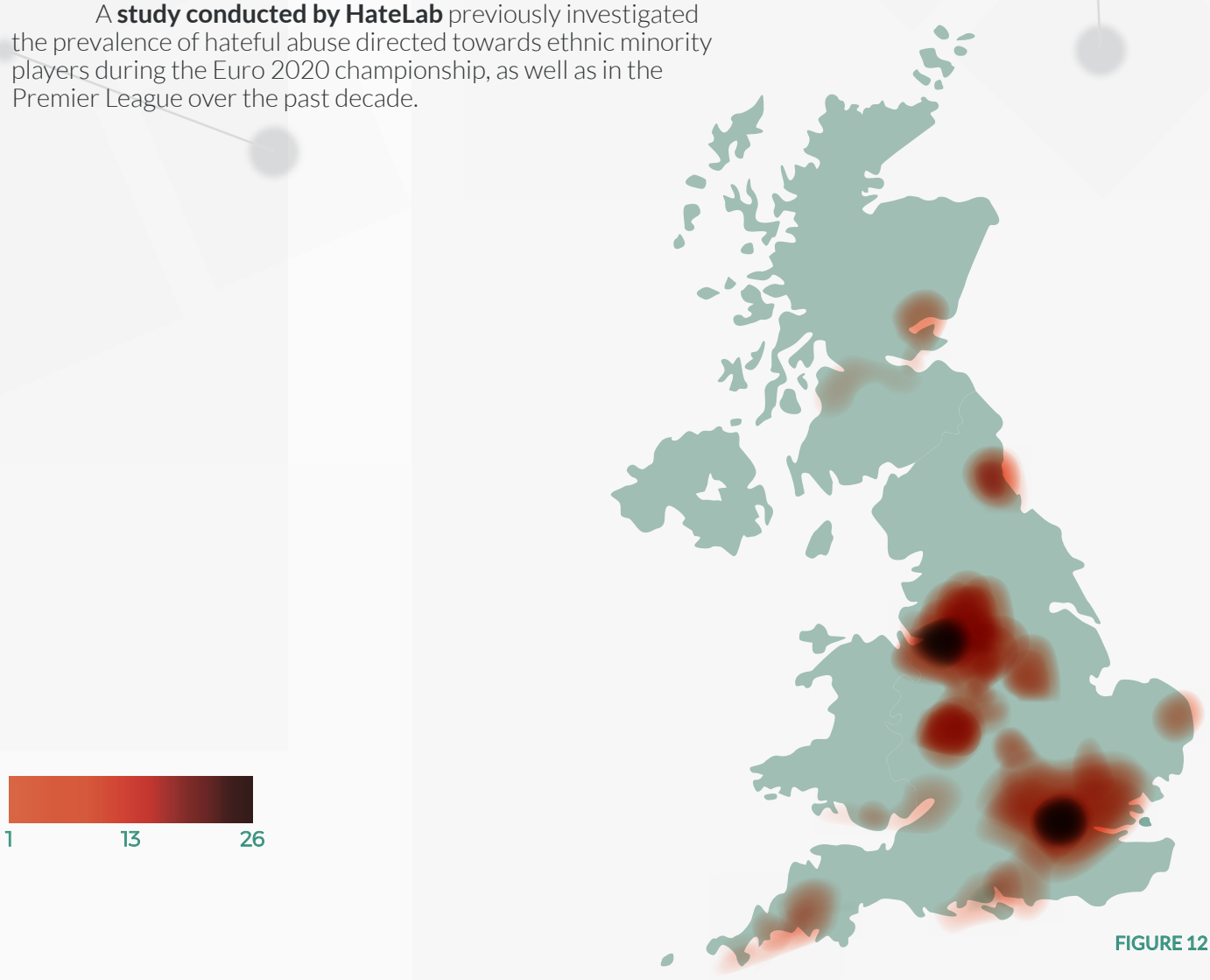

FIGURE 12

The findings of the study showed a similar pattern in both cases. Approximately half of the abusive posts received by ethnic minority players during the Euro 2020 final, and around 40% of the posts directed towards ethnic minority players between 2012 and 2021 in the current Premier League, were attributed to accounts claiming to be located in the UK.

Table 1 shows follower, following and tweet count data that suggests many hate accounts are well established, with the majority having several hundred followers and high tweet counts. The total number of followers in the hate posting network was approximately 100,000, similar to the Women's Euros hate posting network. While this seems large, it is important to note that the number of followers was not equally distributed throughout the network. A few users had extremely high follower counts, meaning a small group had a big influence over the whole network. This suggests that all hate posts did not have an equal chance of being seen by other Twitter users. However, as we only collected hateful posts that @mentioned players, it is likely that the intended receiver saw the post if they were using Twitter in the period shortly after England and Wales matches.

Further analysis of hate accounts highlights that only 5.9 per cent were bots or fake accounts. This suggests that most accounts spreading hateful messages on social media platforms are real people who have found a platform on Twitter to amplify their views and reach a wider audience to target victims.

HATE SPEECH TWITTER ACCOUNT INFORMATION

	M	SD	MIN	MAX	TOTAL
FOLLOWERS	589	3,221	0	41,473	102,508
FOLLOWING	659	2,759	0	35,965	114,750
TWEET COUNT	8,425	19,831	2	138,402	1,465,946

TABLE 1

Table 2 shows that public engagement with hate posts remains relatively low. Compared to HateLab's analysis of the Women's Euros, hate posts sent during the World Cup tended to gain slightly more engagement in terms of quote tweets (.09 to .04), retweets (.18 to .06) and likes (4.39 to 1.72), but lower engagement in terms of replies (.68 to 1.13). The lower mean reply count may indicate Twitter users engaged in less counter-hate-speech during the World Cup, compared to the Women's Euros.

HATE SPEECH TWITTER POST ENGAGEMENT

	M	SD	MIN	MAX	TOTAL
QUOTE COUNT	0.09	0.51	0	5	16
RETWEET COUNT	0.18	0.80	0	7	31
REPLY COUNT	0.68	2.99	0	27	118
LIKE COUNT	4.39	39.04	0	515	764

TABLE 2

“88% OF HATE POSTS REMAIN LIVE ON PLATFORMS”

CONTENT MODERATION

Hate speech posts were recollected after the World Cup competition to determine if they remained visible to footballers and the wider public. We found that 88 per cent of hate speech posts remained live on the social media platforms at the time of writing. This is far higher than the number of racist posts that remained online that targeted players in the Premier League (44 per cent) in the 20/21 season, and slightly lower than the posts remining following the Women's Euros (94 per cent).

Platforms' AI and moderation teams failed to pick up the majority of homophobic, racist and transphobic hate speech posts sent directly to players. Almost all of the posts picked up by our machine learning algorithms would be considered as either offensive or grossly offensive by the average person on the street due to their hateful content (see methods). Furthermore, the targeted nature of the hate speech we identified arguably increases its severity and potential impact on the intended victims. Therefore, it is likely that existing AI solutions and moderation team processes are not working as intended to combat these forms of hateful abuse.

RESPONSE

HateLab and Mishcon de Reya published an **extensive overview** of the current legal and operational responses to online hate speech. These responses are limited in their effectiveness. For example, imposing large fines on platforms that refuse to remove hate speech will only work in a limited number of cases. The hate speech would likely have to be clearly criminal in nature for a take-down notice to be issued, leaving a wide array of offensive content untouched. Imposing a 24-hour time frame on social media companies for the removal of illegal hate speech, as has been suggested by some, also means the damage is likely already done to the victim and the wider community. Even improvements in policing and prosecutions are unlikely to deter the most hardened haters, or those who post in the heat of the moment.

This issue is not a technical or legal one, but a social one. Fans must make a stand against hate.

COUNTER-SPEECH

We have the ability to coordinate in powerful ways to stop online hate. In the face of hate and abuse, counter-speech that reinforces community standards can change online behaviour, and perhaps the minds of those behind the screens. Counter-speech is any direct or general response to hateful or abusive speech which seeks to undermine it. Every social media user can favourably influence discourse through counter-speech by having a positive effect on the speaker, convincing them to stop propagating hate speech or by having an impact on the audience – either by communicating norms that make hate speech socially unacceptable or by 'inoculating' the audience against the speech so they are less easily influenced by it.

Combating hate speech with counter-speech has some advantages over law enforcement and platform sanctions: i) it can be rapid, ii) it can be adaptable to the situation; and iii) it can be employed by any internet user. Counter-speakers are often first at the online scene who witness the hate bubbling up. They are the 'online first-responders'.

Six forms of counter-speech can be considered:

- Attribution of prejudice moral suasion
  e.g. "Shame on you for spreading sexist tropes like that! Imagine if someone said that about your daughter."

- Claims making and appeals to reason
  e.g. "This has nothing to do with immigration! Take a look at these statistics."

- Request for information and evidence
  e.g. "How does this have anything to do with religion?? Do you have any proof?"

- Jokes/comedy and reintegrative shaming
  e.g. Oh no, this trans woman sounds REALLY scary. I'm going to join a radical feminist group immediately - you guys. LOL!

- Mimicry and sarcasm highlighting issues with logic and consistency
  e.g. Hate speech: "I'm officially scared of butch lesbians. #NotHomophobic #JustScared"
  Mimicry: "I'm officially scared of bigoted men. #StereotypingMuch? #JustStupid"

- Reductio Ad Absurdum (an argument pushed to its absurd extremes to identify its inherent problem)
  e.g. "I guess what you're saying is you would feel more comfortable if women didn't exist? Fascinating. Do you want the human race to go extinct?"

“FANS MUST MAKE A STAND AGAINST HATE”

When engaging in counter-speech, or advising others on its use, the following principles should be followed to reduce the likelihood of the further production of hate speech:

1. **Avoid using insulting or hateful speech**
2. **Make logical and consistent arguments**
3. **Request evidence if false or suspect claims are made**
4. **State that you will make a report to the platform, police or a third party if the hate speech continues and/or gets worse (e.g. becomes grossly offensive or includes threats)**
5. **Encourage others also to engage in counter-speech**
6. **If the account is likely a fake or a bot, contact the social media company and ask for it to be removed**

Initial research suggests the wide adoption of counter-speech would see a reduction in hateful communications on large platforms. Those most susceptible to the stemming effects of counter-speech are users who engage in hate speech only occasionally (for example, around 'trigger' events – defensive, retaliatory and thrill-seeking posters).[2] Emerging evidence suggests that counter-speech that uses moral suasion and empathy induction, delivered by multiple members of an ingroup, is most likely to succeed in changing the behaviour of hateful posters.[3] As a community-based measure, counter-speech could have a significant impact on hate speech if well-orchestrated on a large scale.

[2] Williams, M. (2021) The Science of Hate: How Prejudice Becomes Hate and What We Can Do To Stop It, London, Faber and Faber.

[3] Munger, K. (2017) 'Tweetment effects on the tweeted: Experimentally reducing racist harassment' Political Behavior, 39:3. Munger, K. (2021) 'Don't@ me: Experimentally reducing partisan incivility on Twitter' Journal of Experimental Political Science, 8:2. HateLab (2019) 'A study of cyber hate on Twitter with implications for social media governance strategies', Conference on Truth and Trust Online. 4th – 5th October, London, UK. Siegel, A. A. and Badaan, V. (2020) '#No2Sectarianism: Experimental approaches to reducing sectarian hate speech online' American Political Science Review, 114:3.

METHOD

HateLab facilitates access to all open-source online communications, including Twitter, Reddit, 4Chan and Telegram, for the monitoring and countering of online harms, including abuse, threats, identity-based-hate and divisive disinformation. For this report social media posts were sourced from Twitter, Reddit, and 4Chan between 14th November and 18th December 2022. In total 847,370 English language social media posts were collected. For Twitter, a query was created for all tweets @mentioning player account handles to ensure that posts directly targeting the squads were captured. Twitter is a reliable platform for the examination of communications sent to players as all the national teams have established profiles for public engagement purposes. As players do not have public-facing accounts on Reddit and 4Chan, queries were set up to obtain posts which mentioned their full names. We used bespoke machine learning classifiers within the HateLab Platform to classify hateful content.

Our classifiers are trained on gold-standard human-annotated data. To build the training datasets, four human coders independently evaluated posts to determine if they were offensive or grossly offensive based on their hateful content. Only posts that achieved 75 per cent agreement (3 out of 4 coders) were included in the hate class of the training datasets. Content flagged as hateful by our trained algorithms was subject to independent manual inspection by two hate speech experts to reduce false positives.

Metadata from accounts found posting hateful content were collected to perform additional analysis on user sex and location, account creation date, and the number of followers and followees.

In accordance with Cardiff University's Ethics Committee Standards, we have avoided directly quoting hate speech posts to preserve the anonymity of social media users. Instead, we present content in aggregate form via the visualisation of post content in tables, charts, word clouds, and emoji clouds.

THE BENEFITS OF HATELAB'S APPROACH

- **Blended AI and HI:** We partner with experts in civil society, government and law enforcement organisations to source online harms training data allowing us to continually update our AI to the highest standard
- **Minimal error:** We do not rely on simple keywords to identify online harms. Our natural language processing machine learning techniques are top performing, award-winning and verified in international peer-review open science journals (e.g. IEEE, ACM, WWW)
- **Deep knowledge:** We are world-class experts in hate speech, hate crime and cyberisk, and our founders are in the top 3 most cited in their fields
- **Rapidly adaptive:** Our AI + HI approach ensures we remain up-to-date with changes in online behaviours that can avoid detection in fully automated systems
- **Predictive:** Network and information propagation statistical modelling allow us to project the spread and survival of online harms, enabling enhanced threat assessment and mitigation

HateLab is a nonprofit with an ambitious civic mission to democratise the latest AI and data science capabilities amongst civil society organisations so that they can reliably monitor and counter online hate speech, abuse, threats and divisive disinformation.

Arron Cullen is Head Analyst at HateLab. He holds an MSc in Terrorism, International Crime and Global Security and a PhD in Criminology. His research focusses on the nature of online hate speech, responses to online harms, and police use of social media. His latest published work appears in the *British Journal of Criminology*.

Matthew Williams is Professor of Criminology at Cardiff University and is widely regarded as one of the world's foremost experts in hate crime and online hate speech. He advises and has conducted research for the UK Home Office, the Foreign, Commonwealth & Development Office, the US Department of Justice, Google, Deutsche Telekom, EE and BT among others. Williams is also founder and director of HateLab, and he has conducted the largest dedicated study of hate victimisation in the UK. His research has appeared in documentaries for BBC One *(Panorama, Crimewatch)*, BBC Two, BBC Radio 4 *(File on 4)*, ITV *(Exposure)*, Channel 4, CBS, Amazon Studios, and Netflix, and in major publications including the *Guardian*, the *Observer*, the *Independent*, the *Times*, the *Herald*, the *Los Angeles Times*, *Scientific American* and *New Scientist*. In 2021 he published the popular science book *The Science of Hate: How prejudice becomes hate and what we can do to stop it*, with Faber and Faber. @MattLWilliams

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
UKRI Economic and Social Research Council