The Alan Turing Institute

Tracking abuse on Twitter against football players in the 2021 – 22 Premier League Season

Bertie Vidgen, Yi-Ling Chung*, Pica Johansson*, Hannah Rose Kirk*, Angus Williams*, Scott A. Hale, Helen Margetts, Paul Röttger, Laila Sprejer[1]

The Alan Turing Institute Public Policy Programme
Online Safety Team | **2022**

Offensive content warning:
This report contains some examples of abuse (all are synthetic, i.e. not real). You might find them offensive.

[1] Bertie Vidgen is the lead author. *indicates equal contribution.

Table of Contents

Foreword from Ofcom 3

Executive Summary 5

Introduction 8

Background 9

Data Description 11

Qualitative Analysis 12

Machine learning for Detecting Abusive Tweets 17

Quantitative analysis 24

Conclusion 32

Appendix 34

Foreword from Ofcom

Introduction from Kevin Bakhurst, Ofcom's Group Director for Broadcasting and Online Content

Football is a game of high emotion, pride and belonging. For more than 150 years it has been a celebrated part of our national culture. Today, more than half of British adults consider themselves a fan.

Sometimes that emotion can cross the line. Over the years, football has made great strides in tackling unacceptable behaviour by small minorities which can blight the game for everyone else – from hooliganism, to contemptable racist or homophobic abuse. But those threats never go away; and sadly, as this report reminds us, abuse now exists far from the stadium on social media. As Ofcom prepares for a new role overseeing online safety, and to help further our work on media literacy in the UK, we wanted to understand the problem of footballing abuse in more detail.

We commissioned The Alan Turing Institute to analyse more than 2.3 million tweets directed at Premier League footballers over the first five months of the 2021-22 season. This allowed us to examine whether large-scale data analysis and machine learning techniques can shed light on the prevalence and nature of harmful content.

Many of our findings will make sober reading for anyone who loves football. Hundreds of abusive tweets are sent every day, affecting around seven in ten Premier League players. Many victims – though by no means all – are from minority-ethnic backgrounds.

The research suggests that personal attacks often happen during footballing flashpoints, including high-profile transfers or a loss on the pitch. They can also be linked to events outside football, including players' personal lives.

This kind of abuse has no place in sport, any more than in wider society. When Ofcom becomes the regulator for online safety, we'll be shining a light on what tech companies are doing to combat harm to their users, and expect them to put safety at the heart of how they design and run their services.

This is the first regulation of its kind in the world, and it will take time to get it right. But abuse is an urgent problem. So social media firms need not wait for the current Online Safety Bill to become law before making their sites and apps safer for users.

And we must not lose heart. We know the vast majority of fans use social media responsibly. Among our sample of tweets, some 57% were positive towards players, 27% were neutral and 12.5% were critical. The remaining 3.5% were abusive, so perpetrators are very much in the minority.

We also found that a large proportion of these tweets came from people who are only very rarely abusive. These users may be crossing the line between acceptable criticism and outright abuse, and more could be done to ensure they understand where that line stands.

Here, as in so many areas, supporters can play a positive role in protecting the game they love. We know that, among the wider population, only one in five people report or flag potentially harmful content or behaviour they encounter online. As the new season kicks off, we are asking fans to report unacceptable, abusive posts to the social media platforms whenever they see them.

Online abuse is a problem across platforms. We chose Twitter for this study because it is a widely-used platform on which many Premier League football players are active; because several players have reported being abused on Twitter before; and because, unlike most platforms, Twitter makes data available for academic research. This research is not intended as a reflection, or commentary, on Twitter's trust and safety practices.

Nor is online abuse only a problem in football, and in sport more widely. It can affect everyone – from others in the public eye, who may be seen as easy targets; to children who fall victim to bullying or harassment.

Reports such as this will help us to understand the problem, hold tech firms to account when we take on our new responsibilities, and ultimately create a safer life online.

Executive Summary

Online abuse against prominent sportspeople, such as football players, is a growing concern. To help understand this issue, we have launched a new project analysing tweets directed at Premier League Footballers with an account on Twitter. The analysis was run over a period of 165 days (~ 5 months), from the start of the 2021/2022 season (13th August 2021) to the winter break (24th January 2022). We did not analyse online abuse in the Women's Super League, the highest league of women's football in England. The dynamics and patterns of abuse experienced by women players require their own interrogation in dedicated research.

Twitter is the focus of this report for three reasons. First, Twitter is a large and widely-used platform, and many Premier League football players are active on it. Second, several players have reported being abused on Twitter before, such as during the Euro 2020 finals, which makes it relevant for this research. Third, unlike most platforms, Twitter makes data available for academic research via its free to use API, making this type of analysis possible. This research is not intended as a reflection or commentary on Twitter's trust and safety practices. We did not investigate who saw each tweet, how many times they were viewed, how long abusive posts stayed online or what safety measures were applied by the platform.

This report is quali-quantitative in nature[2], comprising manual review of 3,000 tweets by experts; creation of a new machine learning tool that can automatically assess whether tweets are abusive; and large-scale data analysis of 2.3 million tweets. This report was produced by The Alan Turing Institute and commissioned by Ofcom.

1. **The majority of tweets we qualitatively analysed are Positive**. Of 3,000 randomly sampled tweets that we qualitatively analysed, 55% are Positive towards players, 27% are Neutral, 12.5% are Critical and 3.5% are Abusive.

2. **Our qualitative and quantitative results give different estimates of the prevalence of abuse**. 3.5% of the qualitatively analysed random sample of 3,000 tweets are Abusive, compared with 2.6% of the 2.3 million tweets we analysed with machine learning.

3. **The percentage of content which is Abusive is low**. Of the 2.3 million tweets we analysed with our machine learning tool for detecting abuse, 2.6% contain abuse (n = 59,871). This is still a large number in total, which creates a serious risk of harm to the players.

4. **Identity attacks comprise a small percentage of all abuse.** Only 8.6% of Abusive tweets, or 0.2% of all tweets (n = 5,148) contain a reference to the player's identity (i.e. a protected characteristic, such as religion, race, gender and sexuality).

5. **The majority of players received abuse at least once.** 68% of players received at least one Abusive tweet during the period (418/618). One in fourteen (7%) received abuse every day.

[2] See: Blok, A., & Pedersen, M. A. (2014). Complementary social science? Quali-quantitative experiments in a Big Data world. *Big Data & Society*, *1*(2), 2053951714543908.

6. **Abuse varies over time, with peaks following key events**. In particular, on two days, there were substantial increases in both the total number and percentage of tweets which are Abusive. For instance, on 7th November 2021, when Harry Maguire sent a tweet about Manchester United's performance, 10.6% of tweets were Abusive (n = 2,903).

7. **A small proportion of players receive the majority of abuse.** For instance, 12 players account for 50% of all Abusive tweets. Cristiano Ronaldo and Harry Maguire receive the largest number of Abusive tweets.

8. **Many users send just one Abusive tweet.** Of 44,907 users who sent at least one Abusive tweet, 82.3% sent only *one* Abusive tweet. The other 17.7% sent more than one Abusive tweet, accounting for 35% of all abuse (n = 7,948).

If you have questions about this report or would like more information about The Alan Turing Institute's research, reach out to Pica Johansson (pjohansson@turing.ac.uk).

The Alan Turing Institute’s Public Policy Programme

The Public Policy Programme[3] works alongside policy makers to explore how data-driven public service provision and policy innovation might solve long running policy problems and to develop the ethical foundations for the use of data science and artificial intelligence in policy-making. Our aim is to contribute to the Institute's mission – to make great leaps in data science and artificial intelligence research in order to change the world for the better – by developing research, tools, and techniques that have a positive impact on the lives of as many people as possible.

The Online Safety Team

Part of The Alan Turing Institute’s Public Policy Programme, the Online Safety Team provides objective, evidence-driven insight into the technical, social, empirical and ethical aspects of online safety, supporting the work of policymakers and regulators, informing civic discourse and extending academic knowledge. We are working to tackle online hate, harassment, extremism and mis/disinformation. There are three core workstreams: (1) Data-centric machine learning, where we are building and critically examining cutting-edge technologies to flag and rate toxic content; (2) The Online Harms Observatory, mapping the scope, prevalence and impact of content and activity that could inflict harm on people online; and (3) Policymaking for Online Safety, where we are working to understand the challenges in ensuring online safety, and supporting the creation of ethical and innovative solutions.

The Online Harms Observatory

The Online Harms Observatory is a new analytics platform from The Alan Turing Institute’s Public Policy Programme. It combines large-scale data analysis and cutting-edge machine learning developed at The Turing to provide real-time insight into the scope, prevalence and dynamics of harmful content online. It aims to help policymakers, regulators, security services and civil society stakeholders better understand the landscape of online harms. Initially, it will focus on online hate, personal attacks, extremism and misinformation. The Observatory is supported by the Department for Digital, Culture, Media and Sport (DCMS).

Funding

This report was commissioned by Ofcom, in relation to its upcoming role as the UK’s Online Safety regulator. It is one output of a larger Turing project utilising the new Online Harms Observatory. In-kind support was given by The Alan Turing Institute.

Acknowledgements

The authors thank Dr. Florence Enock for her feedback.

[3] https://www.turing.ac.uk/research/research-programmes/public-policy

Introduction

Abuse of public figures, from athletes to politicians, is seen as a growing problem online.[4] Online abuse can inflict harm on the people targeted, as well as on others who may see it – in the most extreme cases, online abuse is illegal. Online abuse can also disrupt online discourses and could dissuade people from entering prominent positions. There are concerns that the effects of abuse can be insidious and wide-ranging, and may not be fully understood. Some have suggested that if public figures are subjected to abuse, without it being challenged by society or handled by the social media platforms, it could normalise the use of insulting, threatening and violent messages.

The design and functionality of social media mean that it is easy for large volumes of abuse to reach public figures, presenting a serious threat to their wellbeing. Despite forthcoming regulation, greater public interest in content moderation, and increased efforts by platforms to tackle toxic content, how to ensure online safety remains a problem. Football may be the UK's most popular sport but professional players have often become a locus of hate and abuse online, with some players being subjected to insulting, threatening and even dehumanising messages. This attracted substantial attention during the Euro 2020 Final when non-white players were widely targeted by racist abuse, both online and offline.[5] Consequently, one man was sentenced in early 2022 for using racist and derogatory emoji in tweets directed at Rio Ferdinand[6], many fans were arrested[7], and the Prime Minister met with social media companies to discuss the abuse received by professional footballers.[8]

There is an opportunity to better understand the dynamics and patterns of abuse targeted at players, thanks to new developments in machine learning and data analytics techniques. With a better understanding of the true extent and scope of abuse, we can start to develop more effective ways of mitigating their harmful effects through interventions, regulation and more effective content moderation. This report is one part of a larger Alan Turing Institute project, aiming to create the reliable and transparent evidence needed to tackle abuse. The report is structured as follows. First, we provide a brief summary of previous work in this area. Then, we summarise our qualitative analyses of the tweets. In the third section, we introduce the new machine learning tools that we have trained to detect abuse against footballers. In the fourth section, we present our quantitative results from our large-scale data analysis. Methods are introduced and explained in each relevant section. Finally, in the Conclusion, we outline our next steps and key lessons learnt.

[4] See: Demos (2020). Public Figures, Public Rage (2020, October 5) and World Athletics (2021, November 25) and *World Athletics publishes Online Abuse Study covering Tokyo Olympic games.*

[5] Jamieson, A. (2021, July 12). Saka, Sancho and Rashford racially abused online after England defeat. *The Independent*.

[6] Man who racially abused Rio Ferdinand on Twitter after England's Euro 2020 final loss is sentenced. (2022, March 1). *Sky News*.

[7] *Hate crime investigation following Euro 2020 final leads to 11 arrests*. (2021, August 5). *NPCC*.

[8] Evans, S. (2021, July 13). English football faces up to global nature of online hate. *Reuters*.

Background

Policymaking and Regulation

Across the globe, countries are introducing new regulations, policies and initiatives to tackle content and activity online that creates a risk of harm. Notably, in the UK, the Online Safety Bill is currently going through Parliament. It is a landmark regulation that will create new requirements for platforms to protect users from certain types of content, to be clear about their policies, and (in respect of larger and riskier services) to publish transparency reports.[9] Other UK laws also now include provisions for online abuse, including the Police, Crime, Sentencing, and Courts Act 2022[10] which, amongst many provisions, extends existing legal protections against physical violence and disorder associated with live football matches to fans who commit offences online.[11] The Professional Footballers' Association (PFA) announced in March 2022 that it had opened more than 400 cases against people who had abused footballers online.[12] Other countries are also announcing new steps to tackle online abuse, such as NetzDG in Germany; the work of the eSafety Commissioner in Australia; the Digital Services Act and the Audiovisual Media Services Directive in the EU; new regulation in Canada; and efforts to reform Section 230 in the USA.

Activism and Campaigns

Numerous initiatives have drawn attention to the problem of online abuse targeted at footballers, such as the four-day boycott of social media by sports bodies, football clubs and players in spring 2021.[13] During the boycott, the charity Kick It Out set four demands for social media companies to tackle abuse: (1) improved prevention, (2) account verification, (3) sufficient punishments, and (4) government intervention by fast-tracking the Online Safety Bill through Parliament[14]. Other charities and civil society organisations have also launched campaigns against online abuse, such as Hate Won't Win and Show Racism the Red Card.[15]

Research and Monitoring

Understanding the prevalence, scope, dynamics and patterns of the abuse directed at football players online in a timely manner is a difficult task. A range of academic studies have investigated the problem of online abuse in football, showing both its growing importance and impact.[16] However, these studies are generally limited in scale and, because of academic publishing cycles, tend to use data that can be several years out of date. A small number of measurement

[9] GOV UK, 'World-first online safety laws introduced in Parliament', (2021), Accessed 30 March 2022.
[10] GOV UK, 'Police, Crime, Sentencing and Courts Act 2022', (2022), Accessed 12 July 2022
[11] The Sportsman, 'Online Hate Crimes Will Now Result in Football Banning Order After Law Change', 29 June 2022. Accessed 12 July 2022.
[12] Sport Techie, 'Premier League Investigates 400 Cases of Online Abuse Against Players, Managers', 16 March 2022. Accessed 30 March 2022.
[13] The Premier League, 'English football announces social media boycott', 24 April 2021. Accessed 30 March 2022.
[14] Sky Sports. 'Social Media Boycott Sent "powerful and United Message" as Sports World Reacts', 4 May 2018.
[15] MacInnes, Paul. 'Kick It Out to Work with Facebook on Scheme to Tackle Football Racism'. The Guardian, 11 October 2020.
[16] See: Kilvington, D., & Price, J. (2019). Tackling Social Media Abuse? Critically Assessing English Football's Response to Online Racism. *Communication & Sport*, *7*(1), 64–79.
Kilvington, D., & Price, J. (2021). The 'beautiful game' in a world of hate: Sports journalism, football and social media abuse. In *Insights on Reporting Sports in the Digital Age*. Routledge.

analyses have been conducted which are quantitative and are more responsive. In June 2021, The Guardian and Hope Not Hate reviewed 585,000 tweets sent five hours after England's three group games in Euro 2020.[17] They filtered 4,505 tweets, which contained keywords and emoji associated with abuse, such as slurs, and then had trained journalists manually review them. 2,012 abusive tweets were identified, of which 102 contained hate speech. According to this study, Harry Kane and Raheem Sterling were the two most abused players. In August 2021, the research agency Signify, on behalf of the PFA, tracked 6,110,629 tweets directed at 750+ player accounts. Using their text analysis flagging algorithm, based on 500 keywords, phrases and emoji, they flagged 16,000 posts for review to see if they contained abuse. The analysis found 1,781 instances of abuse, and showed that two in every five Premier League players received online abuse during the 2020/21 season. They also reported that racist online abuse increased by 48% as the season progressed.[18] In April 2022, the Australian A-League announced that it had launched a new tool for tracking abuse against football players, following successful trials earlier in the year.[19]

Other monitoring projects have tracked abuse against athletes in other fields, such as World Athletics' reports of online abuse during the Tokyo Olympics[20], studies of abuse experienced by National Football League (NFL) players[21] and by female athletes.[22] The breadth, depth and quality of coverage provided by this monitoring varies considerably, and to date there is still need for a bespoke monitoring tool to be made available.

[17] Barr, C., MacInnes, P., McIntyre, N., Duncan, P., & Cutler, S. (2021, June 27). Revealed: Shocking scale of Twitter abuse targeting England at Euro 2020. *The Guardian*.

[18] *Online Abuse—AI Research Study (Season 2020/21)*. (2021, August 4). Professional Footballers' Association.

[19] *Online abuse targeting footballers to be tackled by 'world first' AI software*. (2022, April 4). *The Guardian.*

[20] *World Athletics publishes Online Abuse Study covering Tokyo Olympic Games.* (2021, November 25). World Athletics.

[21] *The Shocking Truth Around NFL Online Trolling*. (2022, January 24). Action Network.

[22] *List: Top 10 Most Trolled Professional Female Athletes On Twitter*. (2022, March 27). V1019.

Data Description

Data collection began on 13th August 2021, the official start of the Premier League 2021 season. Live data is continuously being collected through The Turing's Online Harms Observatory – but for the purposes of this report we finished data collection on 24th January 2022, the start of the Premier League's 2021 winter break. We used an actor-based approach to sampling, rather than a keyword- or hashtag- based approach. We started with a list of all Premier League football players (including development squads and junior teams).[23] We then searched for players with Twitter accounts and manually checked any player who is not Verified. We filtered out any account that did not appear authentic, which left 618 players. They tweeted 9,487 times during the period studied, of which 5,521 were standalone tweets (58%), 1,628 were quotes (17%) and 2,338 were replies (25%). The players themselves retweeted 3,915 tweets in total, and were retweeted 5.8 million times by other Twitter users. We do not collect any direct messages (known as "DMs"). Previous research shows that they are often used to attack people online[24], and access to this data would enable important additional analyses – however, it is not made available for research and cannot be used here.

For the purposes of this report, we focus on a category of content that we call "Audience Contact". This comprises tweets which tag a player and are very likely to be explicitly directed towards them. We include (a) a standalone or quote tweet which tags the player and (b) direct replies to the players' content in Audience Contact. We do not include longer chains of engagement, such as replies to replies to replies (etc.) because these are often not explicitly directed towards any of the players that have been tagged. We collected 3.4 million Audience Contact tweets during the period studied, from which we filtered out tweets with no text content (e.g. tweets which only contain a URL) and tweets not in English, leaving us with 2.3 million tweets, of which 1.1 million were standalone tweets (48%), 1.0 million were direct replies (44%), and 0.2 million were quote tweets (8%). The Audience Contact tweets (3.4 million) were themselves collectively engaged with 5.7 million times, in the form of 4.1 million retweets (73%), 1.1 million replies (19%), and 0.5 million quote tweets (8%).

We pre-processed the tweets by replacing the player usernames, club usernames, other usernames and URLs with generic tokens (@PLAYER, @CLUB, @USER and [URL] respectively) to minimise biases when reviewing the tweets. Hashtags and emoji were not replaced as they encode important semantic information. All data was collected from the Twitter API using an Academic Licence[25], and then was processed and stored in custom-built secure infrastructure in Azure.

[23] https://www.premierleague.com/players

[24] *Hidden Hate: How Instagram fails to act on 9 in 10 reports of misogyny in DMs*. (2022). Centre for Countering Digital Hate.

[25] https://developer.twitter.com/en/docs

Qualitative Analysis

To understand the nature of content directed at players, two of this study's authors each qualitatively analysed 3,000 randomly sampled tweets in the "Audience Contact" category (see Data Description). We removed 24 tweets which were not in English or which did not directly address a player, leaving us with 2,976 tweets (see: Annotation Process). This work directly informed the machine learning tool, and the 2,976 tweets served as our gold standard test set.

Note on presenting tweets: Following the advice of Williams et al. (2017) we do not present verbatim tweets from our dataset of Audience Contact as they largely come from individuals who are not public figures.[26] Instead, we construct synthetic examples which closely resemble the originals. The full dataset is available to download and use for research.

Annotation Framework

Offensive content warning: This section of the report contains some examples of abuse (all are synthetic, i.e. not real). You might find them offensive.

We created a framework for labelling tweets directed towards players, as shown in Figure 1. It is hierarchical, with two levels. In Level 1, tweets are labelled as Abusive or Not Abusive. Abuse is defined as a tweet which "threatens, insults, derogates, dehumanises, mocks or belittles a player" (see below). A key challenge in this research is drawing the line between Abusive and Not Abusive entries, particularly in cases where players are criticised. We sought to consistently apply our definitions and guidelines to all tweets, following what Röttger et al. describe as a "prescriptive" paradigm of data annotation, where the goal is to "encod[e] one specific belief, formulated in the annotation guidelines".[27] However, this is a very difficult task and a degree of subjective decision making must be taken when drawing the line between subjective categories, such as Abuse. Often further context is needed to fully understand the intent of the speaker, which is not available with our data. This is a limitation of this paper (and all research in this domain), which we continue to explore in our ongoing research projects.

In Level 2, Abusive tweets are labelled as either (a) only a personal attack or (b) a personal attack with an identity attack. Not Abusive tweets are labelled as either Criticism, Positive or Neutral. The framework is mutually exclusive and collectively exhaustive, which means that each tweet can be assigned to one and only one of the Level 1 and the Level 2 categories. For example, a tweet cannot be both Abusive and Not Abusive; and cannot be both Positive and Critical. Definitions of each category are given in Table 1. We also created detailed annotation guidelines to clarify and explain the categories. The annotation framework builds on our previous work[28] and was developed by iteratively assessing a sample of tweets, discussing the results as a team and then making updates to the categories and definitions. For example, we originally had a category

[26] Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, *51*(6), 1149-1168.

[27] Röttger et al., (2022). Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics,* 175–190

[28] Vidgen et al.,. (2020). Detecting East Asian Prejudice on Social Media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 162-172) and Vidgen, B., Burden, E., & Margetts, H. (2021). Understanding online hate: VSP Regulation and the broader context. *The Alan Turing Institute*.

for Counter-speech but we dropped it due to a lack of examples and conceptual overlap with the other categories. In addition to the hierarchical annotation framework, we also inductively analysed the tweets, the results of which are presented below.

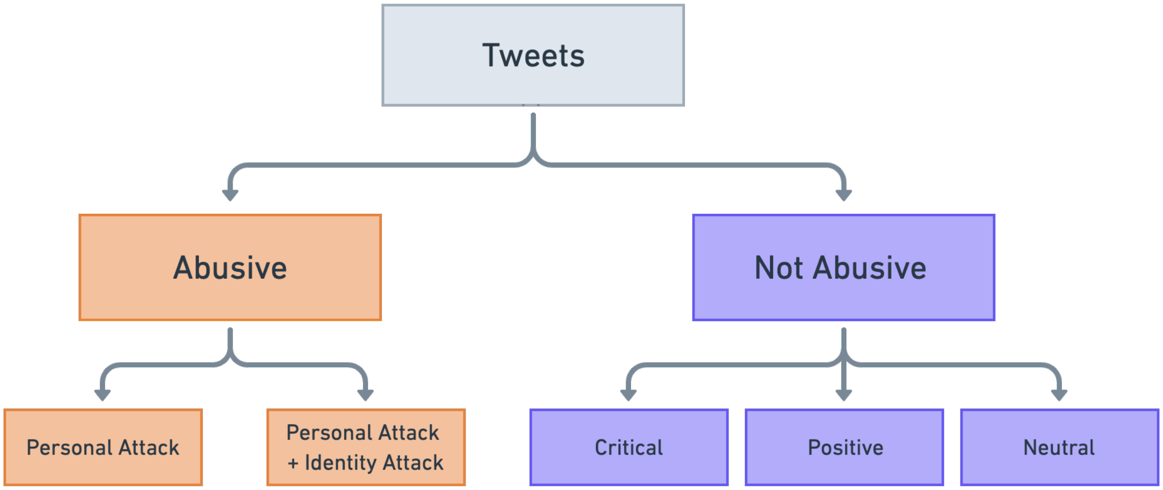

Figure 1: Annotation framework for labelling tweets.

Alt-Text: *The figure shows a flowchart of the annotation framework which is described in the paragraph above. The top-level categories are Abusive and Not Abusive. For Abusive tweets, the secondary categories are Personal Attack and Personal Attack plus Identity Attack. For Not Abusive tweets, the secondary categories are Critical, Positive and Neutral.*

Abusive Tweets

103 of the tweets are Abusive (3.5%). The Abusive tweets mostly contain insults and aggressive language, and profanities are frequently used to express strong dislike towards a player. Abusive tweets often attack players' personalities, character traits or beliefs rather than their performance on the pitch. In some cases, tweets use name-calling (e.g. using 'idiot', 'loser'), casual insults ('get a life', 'shut up') or other demeaning terms and emoji to convey their dislike. We did not identify any cases of threatening and inciting language or language which implied a desire to inflict physical harm. However, given the small sample size, it does not mean that such content does not exist – only that it is likely to be rare overall.

Only six tweets contain identity attacks (0.2%), which we define as attacks against protected characteristics, such as religion, race, gender and sexuality. For example, "[PLAYER] That was stupid 😡 you gey guy" or "Jew in the mud [PLAYER]". The small number of identity attacks is surprising, given concerns about the spread of racial and ethnic-based hatred against footballers during the Euro 2020 final.[29] However, the low prevalence of identity attacks might be because our research design focused on a representative sample of tweets directed at all players – and identity attacks are likely to (a) follow specific events, such as Euro 2020 and (b) be directed at specific players. For example, if we were to review a comparably sized sample that contained tweets sent immediately following a high-profile event or were to only review tweets directed at non-white players, we could identify a higher proportion of Abusive tweets.

[29] Langran, C. (2021, July 17). Why was my tweet about football labelled abusive? *BBC News*.

Table 1: Definitions of the four categories in our annotation framework.
This table shows that 57% of tweets are positive, 27% are neutral, 13% are critical, and 3% are abusive.

Category	Definition	Examples	# (%)
Abusive	The tweet threatens, insults, derogates, dehumanises, mocks or belittles a player. This can be implicit or explicit, and includes attacks against their identity. We include use of slurs, negative stereotypes, excessive use of profanities and angry emoji, as well as abuse which is conveyed through jokes and sarcasm.	"[PLAYER] Fu*ing disgusting man! Shame on u " "[PLAYER] you are a fucking cheat " "[PLAYER] You gay or something?"	103 (3.5%)
Critical	The tweet makes a substantive criticism of a player's actions, either on or off the pitch. It includes critiquing their skills, their attitude and their values. Often, Criticism is less aggressive and emotive, although this is not a defining feature.	"[PLAYER] missed today - he was in bad form" "[PLAYER] [CLUB] Pathetic performance. Please come back better."	373 (12.5%)
Positive	The tweet supports, praises or encourages the player. It includes expressing admiration for a player and their performance, and wishing them well.	"[PLAYER] is an amazing footballer" "[PLAYER] [PLAYER] you all are great members of the team" "I fucking love you [PLAYER] 💙"	1,696 (57.0%)
Neutral	The tweet does not fall into the other categories. It does not express a clear stance. Neutral statements include unemotive factual statements and descriptions of events. Ambiguous tweets are considered Neutral.	"[PLAYER] #ManU has seven BME players this season" "[PLAYER] was playing today, I watched down at my local pub." "Saw [PLAYER] on that pitch at Old Trafford today."	804 (27.0%)
Total			**2,976**

Not Abusive Tweets

Positive tweets

The majority of tweets (n = 1,696, 57%) are Positive. This finding is an important check on the public narrative that online content directed at footballers is overwhelmingly negative. We identified three main types of Positive tweets: (1) tweets which celebrate the players' performance, such as scoring goals; (2) tweets which wish them well or send generic positive messages, often by using emoji such as 🏆, 👏 and ❤️; and (3) tweets which express concern about players' wellbeing. To a lesser extent, we identified tweets which commented on the physical attractiveness of players; challenged abuse from other people; and requested replies and shoutouts. Generally, Positive tweets are heartfelt and emotive in nature, indicating that they were sent by supporters who feel a personal connection to the players.

Critical tweets

373 tweets are Critical (12.5%). Critical tweets can be split into (1) criticisms directed at players' performances on the pitch and (2) criticisms directed at players' activities off the pitch. Criticisms of players' performances are the most common, and addressed both 'lousy' performance in specific games and underperforming over the whole season. These tweets include a lot of football specific terminology (a table of key terms and phrases is given in the Appendix). Captains are subject to more scrutiny than other players, with many references to "the armband" and whether it is "deserved". Criticisms often appear to be against players within each person's favoured team; but we also identified criticism against players in opposing teams, for example remarks on "divers" and players who were seen to foul. There is also some criticism of contracts and pay rates, particularly for new team members. Criticisms of players' activities off the pitch mostly focus on the fact that they engage in non-football activities publicly, such as Marcus Rashford's food activism. Players were routinely told to "stick to football", with some comments addressing the players' tweeting habits and public image. In many cases, players were told to "focus more on your performance".

Neutral tweets

804 tweets are Neutral (27%). This category captures all tweets which do not exhibit a clear stance (i.e. Positive, Critical or Abusive), as well as tweets which are ambiguous. Ambiguous tweets are entries where more context is needed to decipher the true meaning (e.g. "I can't stand what has been done to this player") or which contain URLs which need to be understood to assess the tweet (e.g. "@PLAYER watch this!! 👇👇👇"). Other common types of Neutral tweets include cases of spam and/or marketing, and tweets which mention a player but do not direct the content at them. This includes, for example, detailed religious passages (e.g. "@PLAYER It is written in the Holy Bible that God created everything in 6 days and rested on His eternal throne on the 7th day. Next, Brahm misguided everyone.") as well as tweets which tag many users and players, making it unclear who the real recipient is ("I wonder… hoping I'm lucky @USER @PLAYER @PLAYER").

Discussion of Qualitative Analysis

Deciding the correct category for tweets is generally straightforward as tweets are short-form text (under 280 characters) and our dataset comprises entries which are nearly all aimed specifically at a player. However, in some cases, tweets are edge cases and it is difficult to decide the correct category, such as when they straddle a decision boundary between two categories. For example, "@PLAYER shut up dummy. You can't talk of what is good and what isn't because you're an overrated diver" could be considered both Abusive and Critical. It could be labelled Abusive because, on the one hand, the tweet (a) tells the Player to "shut up"; (b) calls them a "dummy" and (c) describes them as an "overrated diver". However, all of these points could also be considered indicators that the tweet is Critical; (a) "shut up" might be considered an informal way of addressing the player; (b) "dummy" could be seen as rude but not actually abusive given the non-aggressive tone of the tweet and (c) "overrated diver" could be seen as an analysis of their playing. In cases like this, it is difficult to reach a final decision, especially given we only have single tweets to analyse and cannot take into account the full context of the speakers' prior statements.

Some tweets mention two (or more) players with different sentiments. For instance, a tweet might direct abuse at one player but express positivity towards another ("@PLAYER you are the GOAT!!!! So much better than that scumbag @PLAYER"). Although we created rules for these cases (such as any tweet which contains abuse should be labelled Abusive, even if other sentiments are also expressed), in many cases the correct label had to be decided on a case-by-case basis. We sought to apply the annotation guidelines neutrally and avoid making value judgements when assessing the tweets. For instance, we replaced player handles with a generic @PLAYER token to minimise the risk of biases about the players affecting our analysis. However, several incidents were very specific to particular individuals so it was not possible to avoid reviewers being aware of their identity.

Machine Learning for Detecting Abusive Tweets

Our quantitative analyses, presented in the next Section, assess 2.3 million tweets of Audience Contact (see: Data Description). Given this volume of content, we cannot feasibly annotate all of the tweets. Therefore, to enable large-scale analysis of the data, we train a new machine learning tool which can automatically detect whether tweets contain abuse. We set up the machine learning task as a binary assessment, using the Level 1 labels (Abusive and Not Abusive).

To build a high-performing, robust and efficient model, we draw on our previous academic research at The Turing which has demonstrated the strengths of data-centric techniques for machine learning, such as active learning and adversarial data generation,[30] as well as granular labelling frameworks and high-quality annotation processes.[31] Although often construed as primarily an engineering task, machine learning is best understood as a socio-technical problem where human intervention and guidance is needed, and should be explicitly considered, across the entire process.

Iterative Training Process

We started with the pool of 2.3 million unlabelled Audience Contact tweets. We randomly sampled 3,000 tweets for the test set and 1,000 tweets for the validation set.[32] We do this before we acquire and label data for model training so that our evaluation sets (test and validation) are a truly random sample and are not affected by any biases that active data acquisition may introduce. The 3,000 test set was labelled as part of the Qualitative analysis, presented in the previous Section.

We then started the iterative process of model training. To kickstart the iterative loop of model training we created a "Round 0" dataset for the first model to be trained on, comprising tweets identified with a mix of random sampling (1,500 tweets) and keyword sampling (1,500 tweets).[33] We then launched active learning with three rounds of 2,000 tweets (see Box 1). At the end of each round, a new model is trained, which is then used to acquire the next round of entries using a mix of random, diversity, and uncertainty sampling techniques. We then added one round of adversarial data generation to curate a final batch of 500 tweets (see Box 2). We used adversarial

[30] Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2021). Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1667-1682).
Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., & Hale, S. A. (2022). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics,* 1352–1368
[31] Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2021). Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics,* 175–190
Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, *15*(12), e0243300.
[32] In line with standard machine learning practices, the test set is what we use to evaluate the model's performance *after* training is completed. The validation set is used *during* training to fit the model parameters.
[33] A keyword list of 323 abusive terms and identity terms was compiled from existing academic sources. We then manually labelled each term as either a profanity keyword (e.g. "f*ck", "sh*tty") or an identity-based keyword (e.g. "n*gger", "f*g", "immigrants"), and remove terms which could be football words (e.g. "roma", "balls"). 283 out of these 323 appeared in our pool of 2.3 million tweets. Of the 1,500 entries selected by keywords, 750 entries were from matches to any word on the profanity keyword list (n = 169 words), and 750 entries were matches to any word on the identity keyword list (n = 114 words).

learning because during the process of data labelling, we manually reviewed thousands of tweets and were able to qualitatively identify key model weaknesses. These include: (1) separating abuse directed at people who were not players from abuse directed at the players; (2) identifying abusive use of emoji; and (3) identifying abuse in longer tweets which tagged multiple players. In total, across the rounds, 13,500 tweets were sampled from the pool and labelled. We discarded entries that are not in English and do not address players. Accordingly, we kept 13,418 entries for modelling.

Box 1: Overview of active learning process

Active learning starts with a large pool of unlabelled data (e.g. our pool of 2.3 million tweets). To initialise the process, a model is trained on a small starting dataset. This model is then applied to the pool of unlabelled data to select the next batch of entries to label. A data acquisition algorithm is used to select entries which 'confuse' the model (uncertainty sampling) or entries which are unexpected given what the model has seen before (diversity sampling). These are then sent to human annotators for labelling. The model is retrained on the labelled data and the process is repeated. *In effect, the model chooses itself what data it needs to learn from.* The active learning process is shown in Figure 2.

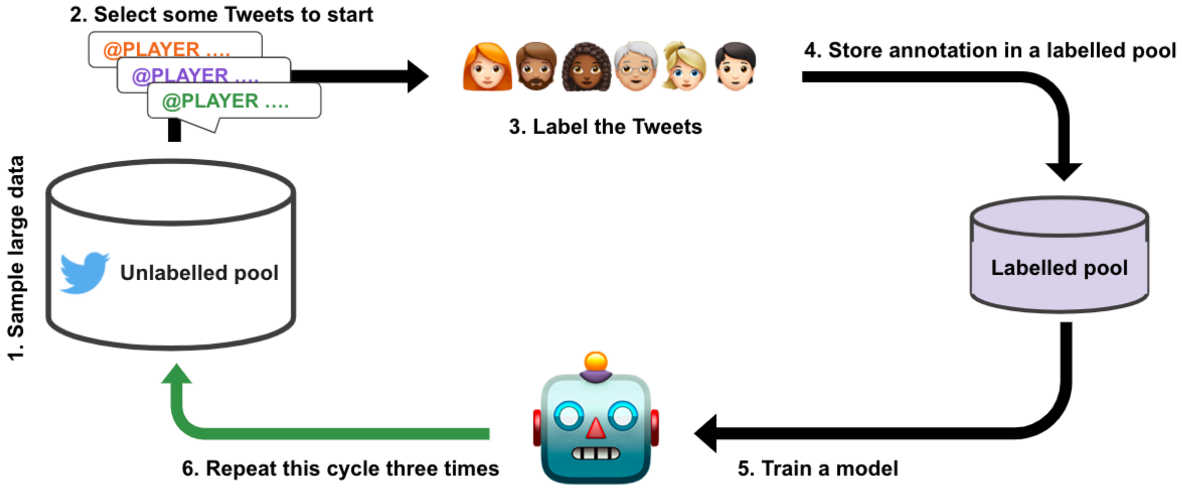

Figure 2: Illustration of active learning process.

Alt-Text: *The diagram shows the iterative cycle of active learning which is described in the box above. Moving clockwise, the cycle runs from the unlabelled pool, to trained human annotators, to the labelled pool, to the AI model then back to the unlabelled pool.*

Box 2: Adversarial data generation

Similarly to active learning, adversarial data generation starts with a trained model. We then task annotators with creating synthetic entries which humans can label correctly but the model will mislabel. For instance, many hate speech models overfit to identity referents, and can be tricked by statements such as “I hate the black tiles in my kitchen”. The model is retrained on the adversarially generated data and the process is repeated. The adversarial data generation process is shown in Figure 3. This method is very effective as annotators find and exploit the model’s key limitations. *In effect, the model learns from what it is worst at classifying.*

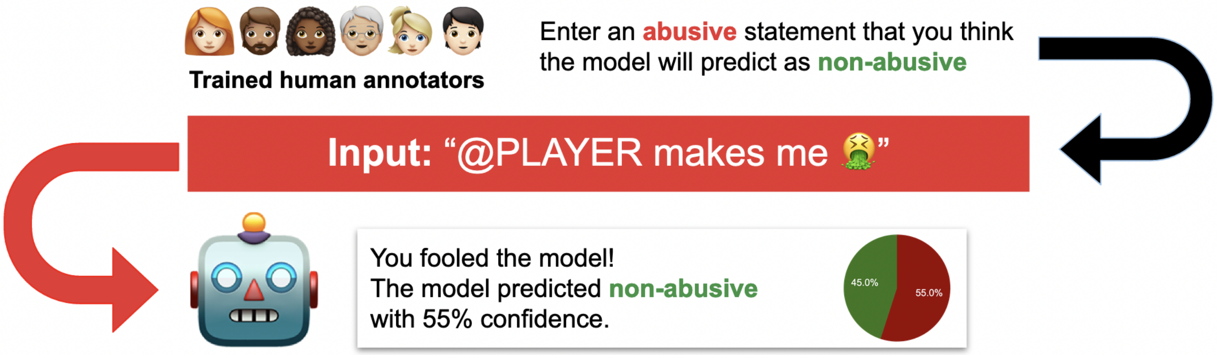
Trained human annotators
Enter an abusive statement that you think the model will predict as non-abusive
Input: “@PLAYER makes me 🤮”
You fooled the model!
The model predicted non-abusive with 55% confidence.
45.0%
55.0%

Figure 3: Illustration of adversarial data generation.
Alt-Text: *The diagram shows adversarial data generation which is described in the box above. Human annotators input an Abusive statement that the model mispredicts as Not Abusive.*

Language Modelling

At the end of each round of data acquisition, we used state-of-the-art methods for language modelling by finetuning pre-trained transformer models on our data. These models are large neural networks which have been pre-trained on billions of online posts and can be used for downstream tasks, such as abuse detection. They are very expensive to train from scratch and are very complex, with billions, or even trillions, of parameters.[34] For this project, we use DeBERTa v3[35], which has achieved state-of-the-art performance in a range of language modelling tasks. We use default hyperparameter values for each round’s model then optimise hyperparameters using a grid-search at the end of the data acquisition process. See the Appendix for more information on the implementation of model training.

Annotation Process

Researcher wellbeing and safety: We follow best practice guidelines for ensuring annotator wellbeing and safety, which we developed in our prior work.[36] Researchers take regular breaks whilst working, fully understand the goals and rationale for the research, and do not solely annotate data for this project to ensure a varied workload. We had a Slack channel in-case any

[34] See: Simon, J., (2021). Large Language Models: A New Moore's Law? *HuggingFace Blog.*
[35] He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543.*
[36] See a set of guidelines for annotator wellbeing that we released publicly in 2019.

safeguarding issues appeared (none did), and made support services available (they were not needed). Researcher wellbeing is continually monitored after project completion.

All of the 13,500 tweets were first annotated by 3-5 crowd workers. Annotators were tasked with annotating (1) whether the tweet is Abusive or Not Abusive; (2) if the tweet is Not Abusive, the category (Critical, Positive or Neutral); (3) whether the tweet is in English; and (4) whether tweets are "Identity directed" and "Person directed" (irrespective of category). To ensure data quality, we flagged tweets for expert review based on whether they had; (1) fewer than 3 annotations; (2) been labelled as non-English by any annotator; (3) less than 100% agreement on the Abusive/Not Abusive category; (4) less than 100% agreement on whether they are an identity attack or personal attack. Two of the study authors acted as expert annotators. Nearly half of the entries in each round required expert annotation, reflecting the complexity of abuse and the importance of experts. The resulting labelled dataset is summarised in Table 2.

Table 2: Labelled data statistics.

This table shows how the 13,418 tweets are split between Not Abusive and Abusive, alongside the number of entries in the training, test and validation sets.

Breakdown of labels	Total	Test	Validation	Training				
				R0	R1	R2	R3	R4
Total	13,418	2,976	993	2,980	1,987	1,991	1,991	500
Not Abusive (%)	82.2%	96.6%	96.7%	71.1%	87%	75.9%	81.9%	42%
Abusive (%)	17.8%	3.4%	3.3%	28.9%	13%	24.1%	18.1%	58%

Model Performance

We evaluate model performance on the 2,976 tweet test set with F1 Score, Precision and Recall. For the purposes of comparison, we show the performance of a DeBERTa v3 model which has not been fine-tuned on any of our labelled data (often called "zero-shot learning"), which we refer to as the "Base model". For each tweet, the models give a predicted probability score of Abusive, which lies between 0 and 1. To binarize these predicted probabilities into one label, we set a cut of 0.5 (any tweets with a score < 0.5 are labelled as Not Abusive, and ≥ 0.5 as Abusive). We also benchmark our results against Perspective API, a widely-used content moderation tool created by Google Jigsaw, which is available 'out the box' for anyone to use, and has not been trained on any of the data we collected.[37]

We do not present results for model accuracy (a measure of the number of entries in the test set which have been correctly predicted by the model). It is an intuitive but ultimately unhelpful measure of performance when the test set is highly imbalanced, as is the case here (96.6% of the

[37] Note that Perspective API is a generic tool for detecting toxicity, identity attacks and insults but has not been trained on any "in-domain" data, i.e. footballer-specific abuse.

test set is Not Abusive). With this imbalance, a model which always predicts the majority class of Not Abusive (which we call a “No Skill” model) can achieve 96.6% accuracy.

F1 is a widely-used metric in machine learning which combines Precision and Recall by taking their harmonic mean. It is useful when datasets are imbalanced. Precision measures how many of all the entries the model has predicted to be Abusive actually are Abusive. Recall measures how many of all the true Abusive examples are correctly identified by the model. We calculate the Macro-F1 score, which takes the average of the F1 scores for both the Abusive and Not Abusive classes. Figure 4 shows the F1 of the models trained at different rounds. The first green dot is the “Base” DeBERTa v3 model. The orange dot shows our model trained on R0 data and the three purple dots show the model trained at the end of R1, R2 and R3 through the active learning process. The pink dot shows our model trained on the adversarially generated data (R4). The first dotted line shows the performance of the “No Skill” model, and the second dotted line shows the Perspective model. The Base model has equivalent performance to the No Skill model (both have a Macro-F1 score of 0.5). Through active learning, model performance improves substantially, up to a Macro-F1 score of 0.82 in R4. Even without seeing any in-domain data, Perspective performs well, with a Macro-F1 score of 0.7.

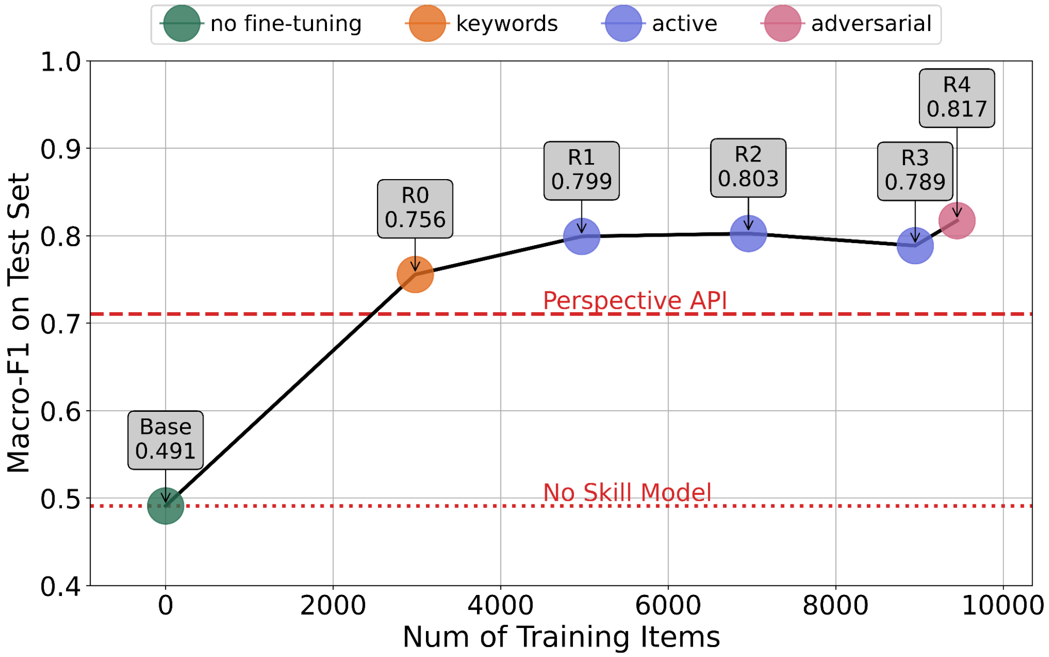

Figure 4: Macro-F1 score of models on the test set.

Alt-Text: *This figure is described in the paragraph above. It shows a concave curve between the number of training items on the x-axis and Macro-F1 on the test set on the y-axis. Model performance improves marginally for each round of training.*

We use a 0.5 cut-off for the model probability scores as standard. However, this cut-off can be adjusted, which introduces a trade-off between Precision and Recall. For instance, if only tweets which have a model score of 0.9 or more are considered Abusive then precision is likely to be very high (i.e. most of the content which it flags as Abusive actually is Abusive). However, recall is likely to be quite low because the high threshold for model score means a lot of the abuse is missed. The Precision-Recall curve in Figure 5 helps to visualise this trade-off between precision and recall for different thresholds. The goal is to maximise the area under the curve. The black dashed line shows the No Skill model, i.e. a model which always predicts the majority class (Not Abusive), and the green line shows the Base model. Both models perform poorly. The orange, purple and pink curves are our models trained at the end of R0, R3 and R4. With just a little training data, the R0 model shows huge improvements, and both precision and recall increase after each round of training as the model learns to better distinguish Abusive from Not Abusive tweets.

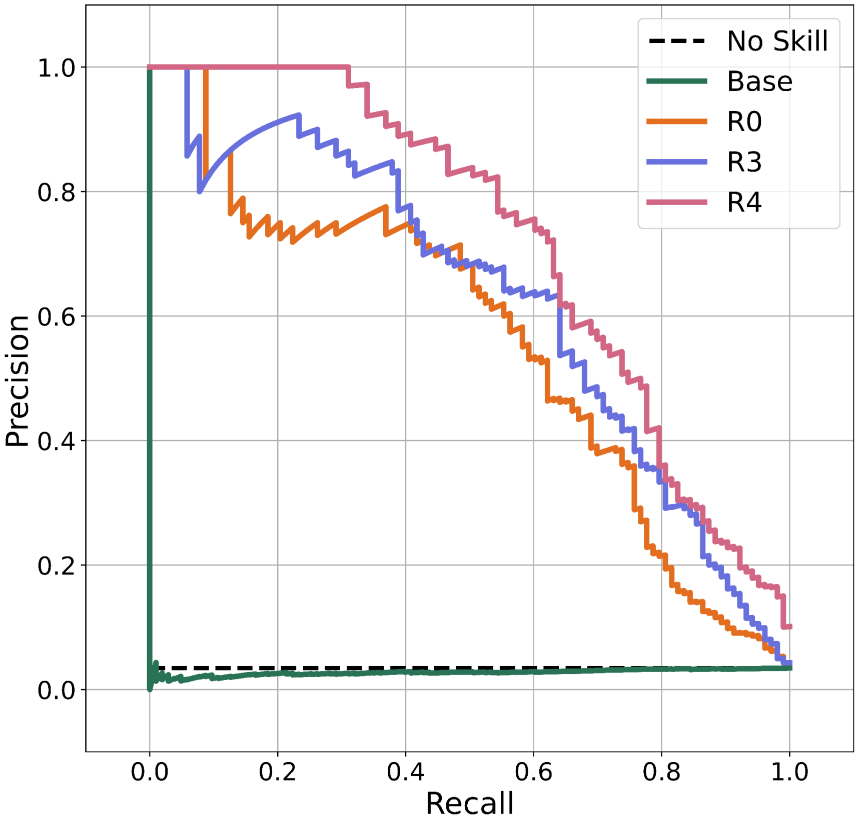

Figure 5: Precision-recall curves across rounds.

Alt-Text: *The figure is described in the paragraph above. It shows concave curves between recall on x-axis and precision on the y-axis. It shows how the precision- recall trade-off improves for each round of training.*

Assessing the Model

Our iterative process for training the AI, using active learning and adversarial data generation, shows clear benefits. The metrics indicate that model performance improves across consecutive rounds of training. We also believe that the model strengths are not fully captured by the numerical scores. Our own probing and testing of the models indicate that with each round they become far better at making nuanced distinctions between content. Two caveats apply to our model's performance, relating to (1) generalisability and (2) uncertainty.

Generalisability

Our model outperforms Perspective API on this particular task – detecting personal attacks and abuse towards English premier league footballers on Twitter. Our previous academic research demonstrates the Perspective model has some vulnerabilities in classifying complex or nuanced forms of toxic language. However, part of the performance gap can be explained by Perspective being a generalist model, designed to work for any platform and any target of abuse. In contrast, our model is a specialist. On one hand, this is a strength of the model because it is highly-adapted to the task at hand. On the other hand, this is a weakness of the model because it may be brittle to small changes in the setting or task. For example, if we applied it to other domains, such as detecting abuse directed at MPs, it may perform poorly. In future work, we plan to evaluate techniques in transfer learning and domain adaptation to improve the generalisability of our model to different platforms and targets of abuse.

Uncertainty

We have trained a single model which outputs a predicted probability of abuse for each entry (which lies between 0 and 1). To convert this predicted probability into a binary label (Abusive and Not Abusive), we use a single cut-off of 0.5. Often it is desirable to understand the *certainty* of the label, and the degree to which the label should be trusted. In principle, we could use the predicted probability as a confidence measure (i.e. scores closer to 1 are more 'confident'), but this does not capture whether the model should be so *certain* in its confidence. However, with our current setup we are unable to compute measures of certainty, such as confidence intervals, because we only receive one prediction per entity. This could be addressed by training multiple models (called an 'ensemble') and averaging the predictions from each of them. This is very computationally intensive. Alternatively, we could modify how we train and evaluate a single model to introduce some variation in its predictions. This can be achieved by using dropout layers during the prediction stage so that only part of the deep network is active at a given time. We are considering both options in our future work to help end users better understand the model outputs.

Identifying Identity Attacks

To provide more insight into the nature of abuse, we also assess whether Abusive tweets contain identity attacks.[38] We did not train a machine learning model for this task. Instead, we take the entries predicted by the model as Abusive, and conduct a keyword search to estimate which examples also contain an identity attack. The keyword list includes 114 identity terms (e.g. 'immigrants', 'Muslim', 'Black people') and slurs commonly associated with identity groups.[39] We refer to the tweets that contain terms in this keyword list as identity attacks.

[38] See the short discussion of identity attacks in the Qualitative analysis.

[39] The list of keywords is available online at: https://drive.google.com/drive/folders/1MqB5QxcQQ8y_wBdw0q6Ev81_eiG0Aai8. Please be aware that many of them are offensive, derogatory and hateful terms. You might find them offensive and they should be viewed with caution.

Quantitative analysis

Summary Statistics

2.3 million Audience Contact tweets were used in this research (see Data Description), of which 2.6% (59,871) were identified by our machine learning tool as Abusive. Of the Abusive tweets, 8.6% made reference to an identity (5,148, or 0.2% of all Audience Contact). Nearly seven out of ten Premier League players received at least one Abusive tweet, and on average 47 players received at least one Abusive tweet every day. The data is described in Table 3.

Table 3: Summary of abuse directed at players.
This table summarises the total and average daily number of tweets by summary metrics.

Metric	Total	Average (Daily)
Number of tweets	2,310,889	14,005
Number of Abusive tweets	59,871	362
Percentage of tweets that are Abusive	2.6%	2.6%
Identity attacks	5,143 (8.6%)	31 (8.6%)
Number of players who received at least one Abusive tweet	418/618	47/618
The player who received the highest number of Abusive tweets	Cristiano Ronaldo	N/A
The club who received the highest number of Abusive tweets[40]	Manchester United	N/A
Hashtag with highest TF-IDF score in Abusive tweets[41]	#oleout	N/A

When Does Abuse Peak?

During the 2021 Premier League season there were two large peaks in abuse: on 27th August 2021 and 7th November 2021. There were also three smaller peaks, when at least 1,200 Abusive tweets were sent in a single day (2nd September 2021, 24th October 2021, 21st November 2021). Figure 6 shows the total number of tweets sent each day, as well as the number of Abusive tweets.

Peak 1 (27th August 2021): Ronaldo transfer

On 27th August 2021 the largest number of Abusive tweets were sent on a single day during the whole period (n = 3,961), as well as the largest number of identity attacks (n = 301). This also was the day with the largest total number of tweets (n = 188,769), more than any other day by a factor of three. The percentage of tweets which were Abusive was 2.3%, which is marginally lower than

[40] Calculated as the club with the largest number of Abusive tweets when totalled for all players at the club.
[41] TF-IDF is an NLP method that we use to understand which hashtags appear more commonly in Abusive tweets than Not Abusive tweets. An overview is available in the Appendix.

the daily average (2.6%). The peak in activity was most likely due to the transfer of Cristiano Ronaldo, who has 98.4 million followers on Twitter, from the Italian club Juventus to Manchester United on the 27th August, and to a lesser degree by the arrest and charging of Benjamin Mendy on suspicion of rape and sexual assault.

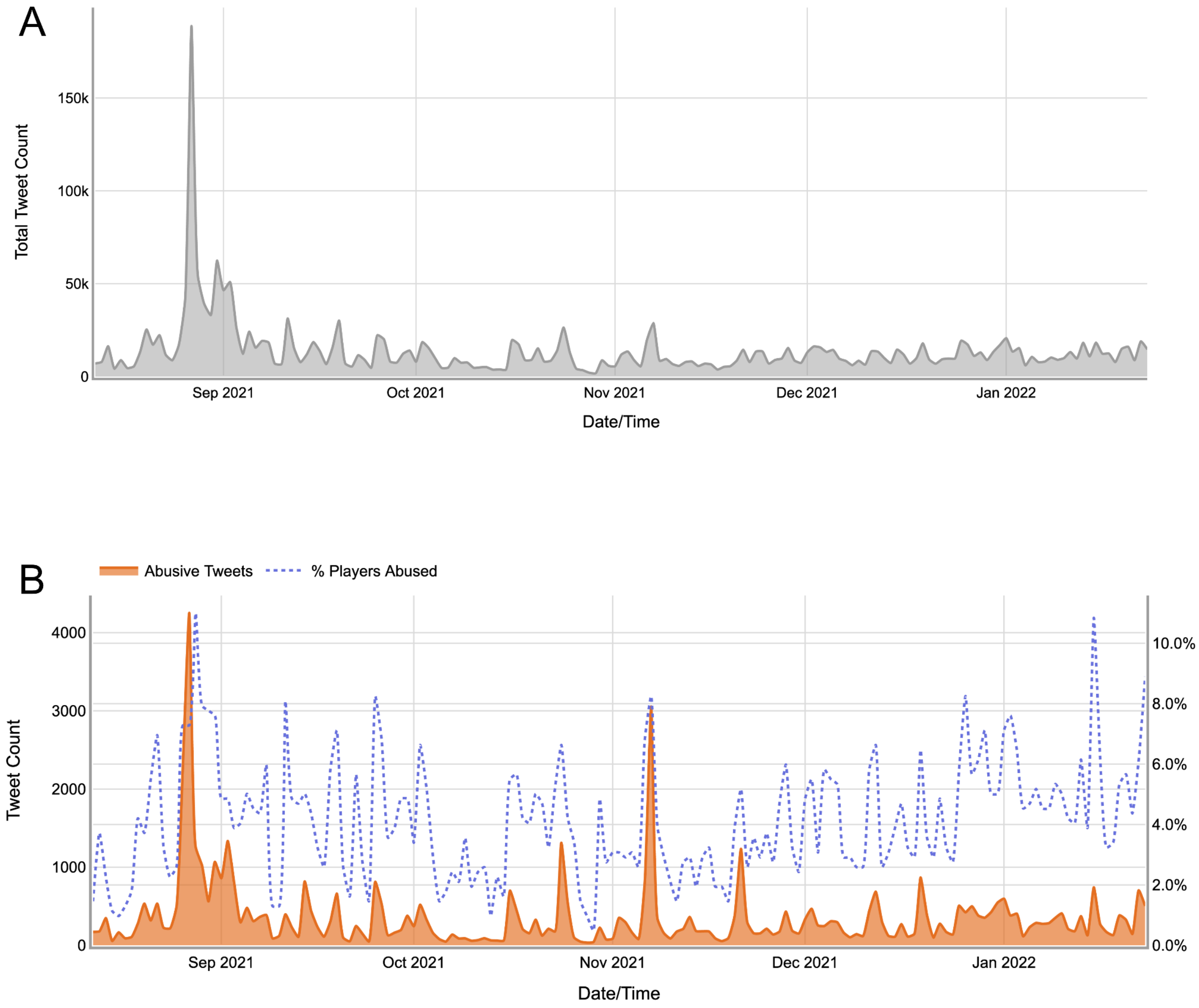

Figure 6: Panel A shows a plot of the total number of tweets sent each day. Panel B shows a plot of the number of abuse tweets sent each day, measured on the left hand y-axis, in addition to the percentage of players who received at least one abusive tweet each day, measured on the right hand y-axis.

Alt-Text: *Figure 6A shows the number of tweets sent each day in the 5 month period. There are small peaks throughout the period, with one large peak in the beginning of September. Figure 6B shows the total number of abusive tweets sent each day in the 5 month period. There is one large peak in September and one in November.*

Figure 7 shows that Cristiano Ronaldo received the most tweets (i.e. he is furthest to the right on the X axis) and the most Abusive tweets (i.e. he has the largest bubble) on the 27th August, with 170,817 total tweets and 3,825 Abusive tweets. On this day, he was mentioned in 90% of all tweets and 97% of Abusive tweets. Figure 7 also shows that most of the hashtags surfaced from the Abusive tweets using a TF-IDF score on the 27th August are references to Cristiano Ronaldo.

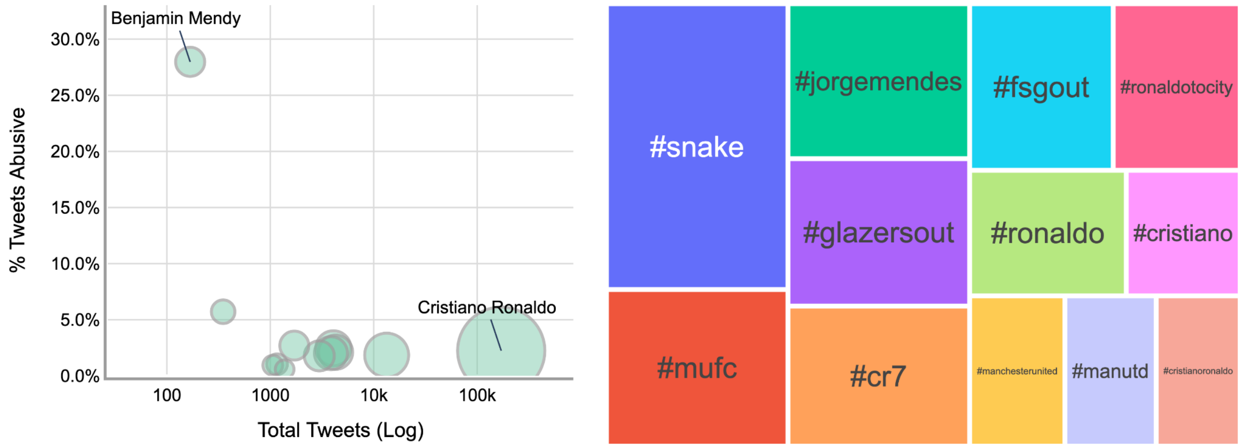

Figure 7: Scatter plot of total tweets sent to each player versus the percentage of which are abusive, with size representing the number of Abusive tweets (left), and treemap plot of hashtags with highest TF-IDF scores represented by size (right), for data from the 27th August.

Alt-Text: *This figure shows that Cristiano Ronaldo received the most tweets and the most Abusive tweets on the 27th of August. The most used hashtags on this day were #snake, #jorgemendes and #mufc.*

Peak 2 (7th November 2021): Harry Maguire's apology post

On 7th November 2021, the second largest number of Abusive tweets for a single day were sent (n = 2,903) and the percentage of tweets which were Abusive was the highest (10.6%). The total number of tweets sent did not substantially increase, which suggests that the spike was not caused by a general increase in the amount of online tweeting. The abuse was triggered by a tweet from Harry Maguire in which he apologised for Manchester United's performance, saying that they were going through "a rough period" (see Box 3). Many Twitter users reacted with insulting and demeaning language, such as telling him to "shut up" or "f*ck off". Figure 8 shows the key hashtags used in Abusive tweets, and the amount of abuse received by Harry Maguire.

Box 3: Tweet sent by Harry Maguire on 7th November 2021[42]

As a group of players we are going through a tough period. We know and accept this is nowhere near good enough. We feel your frustration and disappointment, we are doing everything we can to put things right and we will put things right.

Thanks for your support ❤️🔴 *UNITED*

[42] Harry Maguire is a well-known public figure, which provides a research basis to present his tweet verbatim.

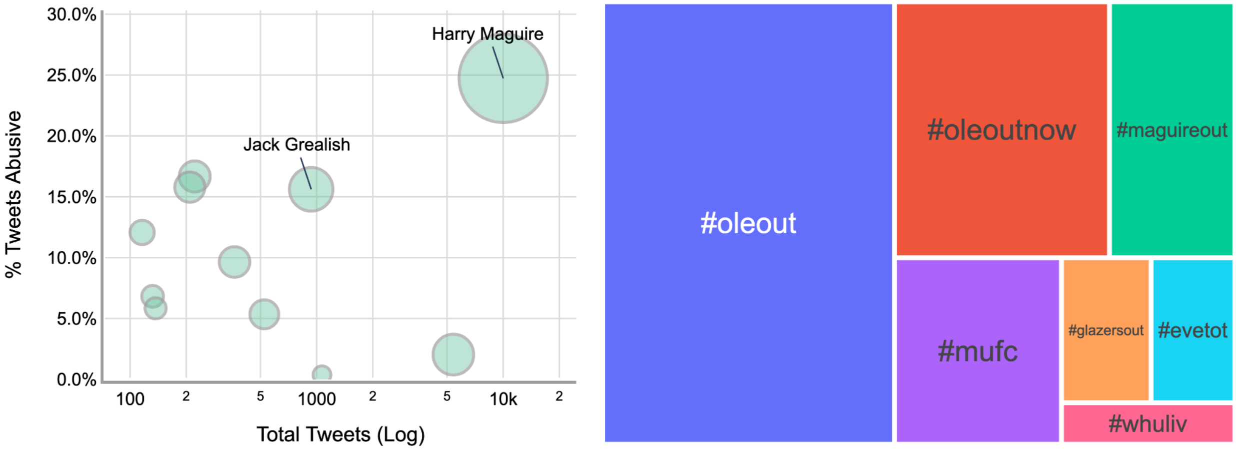

Figure 8: Scatter plot of total tweets sent to each player versus the percentage of which are abusive, with size of bubble representing the number of Abusive tweets (left), and treemap plot of hashtags with highest TF-IDF scores represented by size (right), for data from 7th November.
Alt-Text: *This figure shows that Harry Maguire received the highest proportion of abusive on November 7th. The most used hashtags on this day were #oleout, #oleoutnow and #mufc.*

Percentage of players who receive abuse

For each day, we calculated the percentage of players who received at least one Abusive tweet. As anticipated, this is positively correlated with the total amount of Abusive tweets sent each day (Pearson Correlation Coefficient = 0.55). For instance, on the 27th August 2021, when Cristiano Ronaldo transferred to Manchester United from Juventus, the most Abusive tweets were sent and the most players received abuse (11%). However, there are some interesting exceptions. On 15th January 2022, a similar number of Premier League Footballers to the 27th August received at least one Abusive tweet (10.8%, or 67 out of the 618 players), even though relatively few abusive tweets were sent (645). The total number of tweets was also relatively low, as shown in Figure 6. The dynamics between the percentage of tweets which are Abusive and the percentage of players who receive abuse can be used to identify player-specific events where the overall amount of abuse has not drastically changed, beyond one player.

Indications of coordinated behaviour

Understanding the organisation of online abuse is of increasing interest given the harm caused by coordinated attacks and "pile-ons".[43] We identify indicators which suggest there is a degree of organisation (albeit likely organic, rather than coordinated) in the abuse targeted at players. The exact text of 929 tweets is sent by different users at least twice (i.e. the tweets have been duplicated) and for 19 tweets their exact text is duplicated more than 10 times, with one tweet duplicated 100 times. The duplicated tweets are generally short and generic, such as "F*ck you @PLAYER" and "@PLAYER @CLUB Shut up", which would suggest that this is not coordinated activity. However, in some cases the duplicated tweets are sent in very short succession. For example, following Harry Maguire's tweet on 7th November 2021 (see Box 3), the tweet "@PLAYER @CLUB F*ck off" was sent to him 69 times by different users within two hours. We do

[43] See: Law Commission. 'Reforms to protect victims of online abuse and safeguard freedom of expression announced'. Accessed 21 July 2021.

not see this text tweeted on many other days than the 7th. It is possible that this duplication occurred because users saw the abusive message and decided to replicate it – indicating organic organisation rather than coordinated behaviour. However, we cannot fully confirm this without further investigation, which The Alan Turing Institute aims to undertake in future work.

Who is Targeted by Abuse?

Players vary in both how many tweets they receive overall, and in the percentage of the tweets they receive which are Abusive. Figure 9 shows the percentile of Abusive tweets received by players, against the percentile of players which is accounted for. It shows that a very small number of players receive the majority of abuse, with just 2% of players (n = 12) receiving 50% of all abuse.

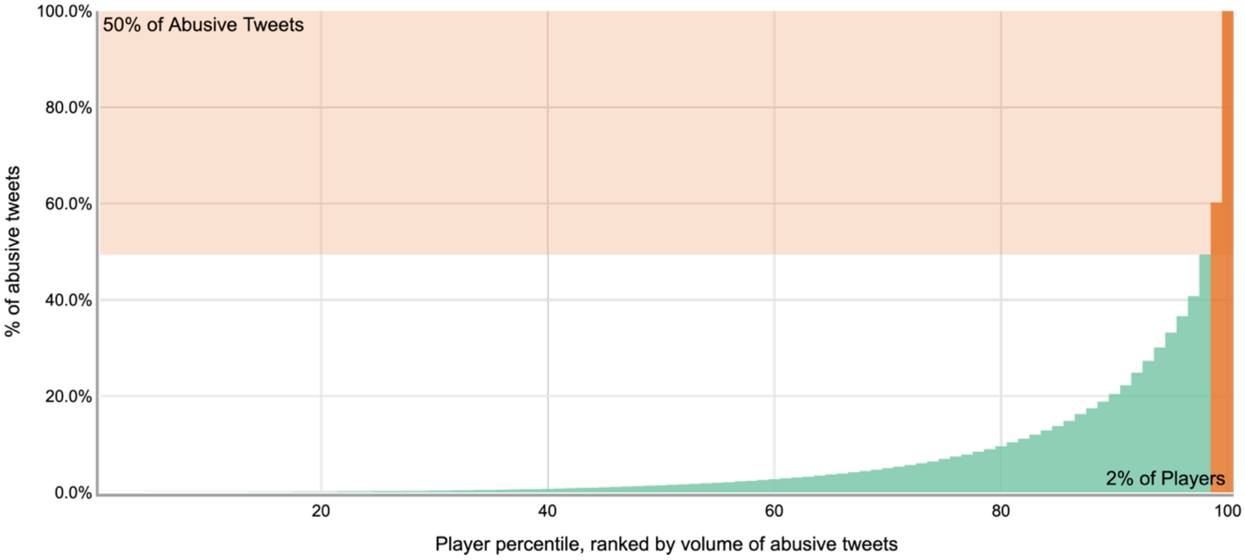

Figure 9: Cumulative histogram of the percentage of players receiving what percentage of abuse.
Alt-Text: *This figure shows that only 2% of players (n = 12) receive 50% of abuse.*

The players who receive large amounts of abuse are primarily well-known figures. Cristiano Ronaldo, Harry Maguire and Marcus Rashford received the largest number of Abusive tweets, as shown in Table 4.[44] These players also received a large number of tweets overall, as shown by the size of the dots in Figure 10 – and are well-known by the UK public. That said, important differences exist between even the most tweeted-at players, and the relationship between the total number of tweets and the number of Abusive tweets that players receive is uneven. For instance, Cristiano Ronaldo received the greatest number of tweets and the most Abusive tweets. But although Cristiano Ronaldo received eight times the total number of tweets as Harry Maguire, Cristiano Ronaldo received 40% more abuse than Harry Maguire (12,520 vs. 8,954 Abusive tweets).

Figure 10 shows that some players received large amounts of abuse, even though they receive fairly few tweets overall. Ciaran Clark, James McArthur and Benjamin Mendy[45] were the most

[44] In the Appendix, we show similar tables to Table 4 for the abuse directed at players, summed by their club and nationality.

[45] Since our tracking started Benjamin Mendy was suspended by Manchester City in August 2021 after he was charged with rape and sexual assault. He has denied the allegations against him.

targeted players (i.e. highest on the y-axis). 34%, 30% and 24% of all the tweets they received were Abusive, respectively. In all three cases, qualitative analysis of the data suggests there were specific triggers for the high volumes of Abuse that these otherwise lower-profile players received.

- **Benjamin Mendy**. Benjamin Mendy was arrested and charged on suspicion of rape and sexual assault in August 2021. 58% of the Abusive tweets aimed at him were sent on the day he was arrested/charged (26th August 2021) and 78% were sent within a week of this. The Abusive tweets mostly conveyed anger and disgust at his alleged actions.

- **Ciaran Clark**. On 30th November 2021, Ciaran Clark, a player for Newcastle, was sent off in a game against Norwich City. 78% of the Abusive tweets he received were sent on this day. Most tweets appear to be from fans of his club (Newcastle), attacking his perceived poor performance, with many suggesting he should "get out" of the club. A small number of tweets focused on his nationality. Otherwise, Ciaran Clark does not receive many tweets, compared to other players.

- **James McArthur**. On 18th October 2021, James McArthur, a player for Crystal Palace, was given a yellow card during a match against Arsenal after he stepped on the leg of Arsenal fan-favourite Bukayo Saka. Users who appear to be Arsenal fans used insults to refer to James McArthur. 54% of the Abusive tweets he received were sent on this day.

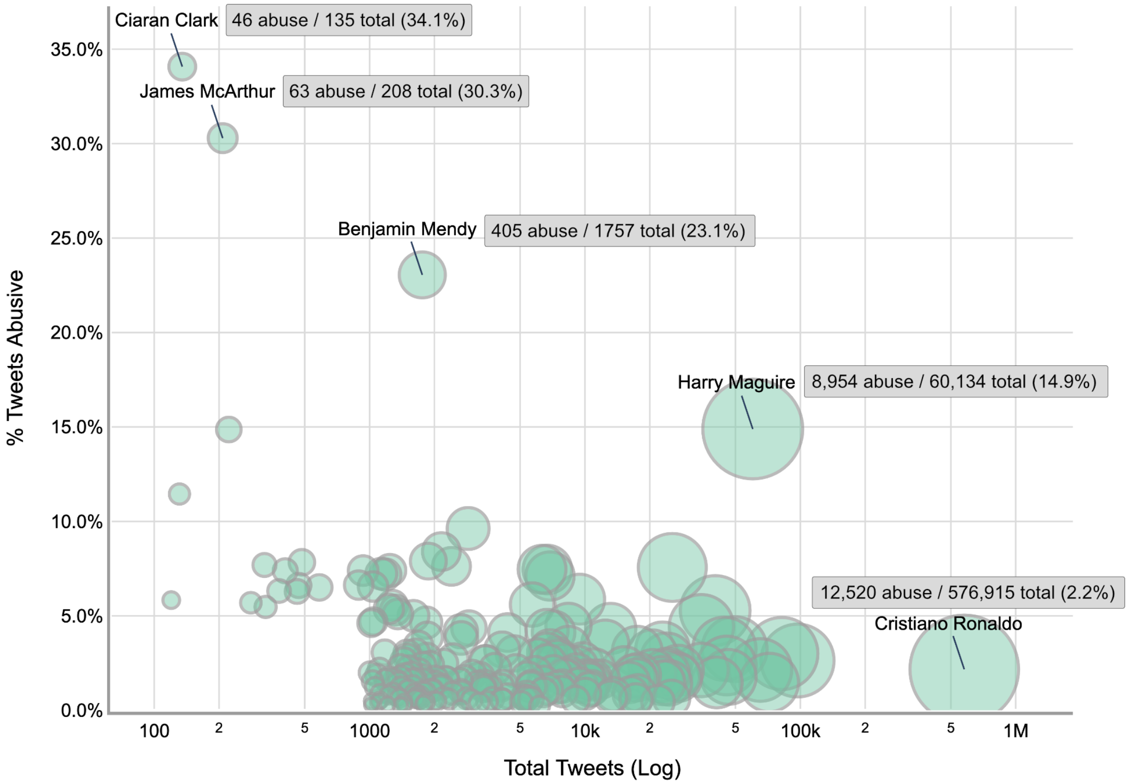

Figure 10: Total tweets sent to each player versus the percentage of which are Abusive. Each point represents one player. The size represents the number of Abusive tweets received by the player.
Alt-Text: *This figure shows that Benjamin Mendy, Ciaran Clark and James McArthur receive large amounts of abuse, even though they receive fairly few tweets overall.*

Table 4: Players who received the greatest number of Abusive tweets.
This table shows that 8 out of 10 of the most abused footballers play for Manchester United.

Player	Total number of Abusive tweets	Percentage of tweets which are Abusive	Club
Bruno Fernandes	2,464	3.00%	Manchester United
Cristiano Ronaldo	12,520	2.20%	Manchester United
David de Gea	1,394	2.10%	Manchester United
Fred (Frederico) Rodrigues Santos	1,924	7.60%	Manchester United
Harry Kane	2,127	5.30%	Tottenham Hotspur
Harry Maguire	8,954	14.90%	Manchester United
Jack Grealish	1,538	4.40%	Manchester City
Jesse Lingard	1,605	3.20%	Manchester United
Marcus Rashford	2,557	2.60%	Manchester United
Paul Pogba	1,446	3.30%	Manchester United

Who is Sending Abuse?

854,667 users tweeted at least once at a football player in our dataset. Almost 95% of them did not send anything Abusive (n = 809,760). 44,907 (5.3%) sent at least one tweet which we identified as Abusive. 82.3% of the 44,907 users sent one Abusive tweet (n = 36,959) and the other 17.7% sent one or more Abusive tweet (n = 7,948). Only 788 users sent 5 or more Abusive tweets. We had anticipated that the data would be more skewed, and the relatively uniform shape of this distribution indicates that a large proportion of the abuse is coming from users who are only rarely abusive. The distribution of all users who send Abusive tweets is shown in Figure 11.

30,452 of the 44,907 users sent both Abusive and Not Abusive tweets; and 14,455 sent only Abusive tweets. Interestingly, 93.3% of the 14,455 users who sent only Abusive tweets only sent one tweet to a football player *at all* during the period of data collection. This means that they tweeted only once at a football player and the tweet was Abusive, as shown in Figure 12. This subset of users may exhibit specific behavioural patterns which could be investigated in future work.

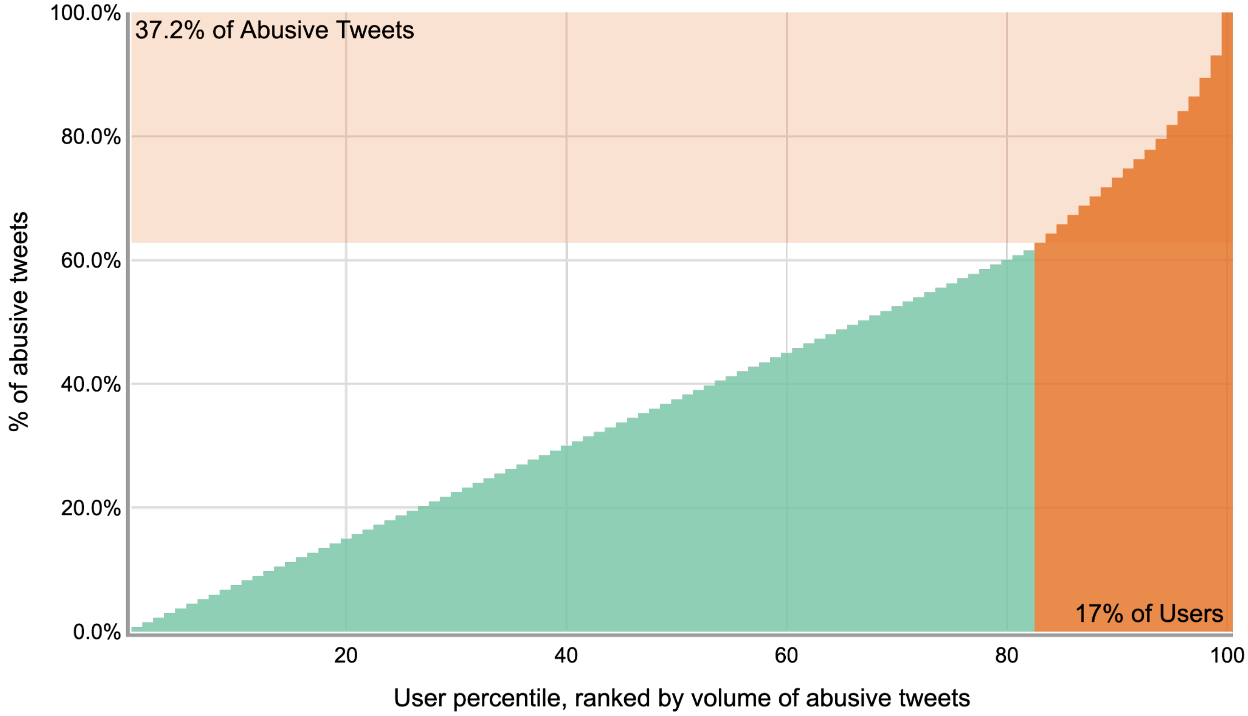

Figure 11: Cumulative histogram of the percentage of users sending what percentage of abuse.
Alt-Text: *This figure shows that a large proportion of the abuse is coming from many different users.*

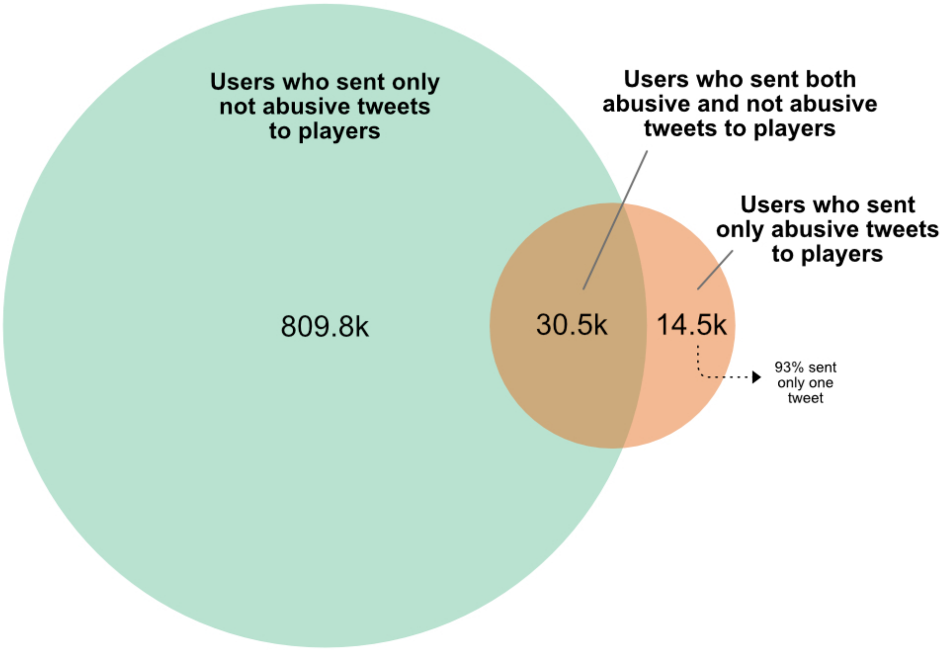
Users who sent only not abusive tweets to players
Users who sent both abusive and not abusive tweets to players
Users who sent only abusive tweets to players
809.8k
30.5k
14.5k
93% sent only one tweet

Figure 12: Venn diagram (circles are not to scale) showing the number of users who send tweets to footballers according to whether they send Abusive tweets or Not Abusive tweets.
Alt-Text: *This figure shows that the majority of users sending abusive tweets only sent one tweet.*

Conclusion

This report presents new insights into abuse directed against Premier League football players on Twitter, and showcases an innovative real-time methodology for assessing online content and activity that creates a risk of harm, such as abuse directed against individuals. This Ofcom commissioned report is one output of a larger Turing project utilising the new Online Harms Observatory, which can be used for ongoing assessment of online threats.

Through writing this report, we identified some advantages of The Observatory compared with a standalone report, specifically the fact that the time required to produce a report is often months – during which the social media landscape may have changed. For instance, once analysis had been completed for this report, we identified several important events which drove large amounts of abuse in February and March 2022, such as the allegations relating to Kurt Zouma's treatment of an animal. This will be considered in future Alan Turing Institute outputs.

Lessons Learnt

Delivering this project has identified several lessons learnt for similar endeavours, which The Alan Turing Institute shares with the research and policymaking community.

1. **Data processing:** Our preliminary qualitative analysis showed that tweets which @ mention the players vary considerably in terms of who they are really targeted at. As such, we focused primarily on Audience Contact data for this report. It is likely that our AI tool would underperform on other tweets (such as long chains of replies). We have found few studies which adopt a similar analytical approach as most simply take a stream of all tweets. This could undermine the integrity of their analyses. Analytically-informed data processing is essential for ensuring insights are meaningful and robust.

2. **Data labelling:** High quality data labelling is essential for ensuring the AI is trustworthy and reliable. We faced numerous challenges with crowd-sourced annotation and a large proportion of the data was reviewed by our experts. Expert-driven annotation is essential when working with subjective and complex categories, such as abuse.

3. **AI training:** The iterative approach that we took to training the AI resulted in a high performing and robust model. However, it also produced numerous logistical challenges given that the process is path-dependent. In practice, this means that data and model quality must be checked at every stage and cannot be revisited post-hoc – which is very different to most machine learning projects. Further, the large number of stages involved in each round presents challenges for the research team, who must be available at specific times to complete their part of the process. Effective coordination across the team is essential when the process for training the AI is complex and path-dependent.

Next Steps

The Alan Turing Institute proposes several extensions to further develop the findings in this report through The Observatory:

1. **Expand the AI:** We could use the more granular categories from our qualitative analysis to create machine learning tools which automatically detect other types of content, particularly (a) Criticisms and (b) Positive language. Otherwise, we will continue to optimise model performance for the binary task (Abusive or Not Abusive), including keeping it up-to-date as the nature of abuse on Twitter changes.

2. **Expand analyses:** We aim to provide more analytical insights into users by investigating the role of bots and anonymous accounts, coordinated behaviour (using network analyses) and account takedowns; events (using time series analysis), how abusive content is responded to, and the content of tweets.

3. **Expand coverage:** This project has only focused on men's football players. We aim to expand coverage to women's football players in the near-future, as well as players in different leagues and country football systems. We will also monitor abuse directed against groups of other prominent individuals, such as MPs.

Appendix

Data Statement

To document the creation and provenance of our final, labelled dataset of 13,418 tweets, we present a data statement.[46]

I. Curation Rationale

In order to study the prevalence and trends of abuse directed towards footballers, we collected tweets from the Twitter API across 618 English Premier League footballers in the 2021/22 Season. In total, we collected 2.3 million tweets which contained 'audience contact' (see Data Description). Of these, 13,500 were labelled by crowd-sourced annotators, whose labels were then validated by expert annotators to ensure quality and consistency. After quality control checks in each round of annotation, the final dataset contained 13,418 tweets (see Table 2 for detail). The purpose of our labelled dataset is to train a model for the prediction task of Abusive or Not Abusive, but we also collected secondary labels within Not Abusive tweets (Positive, Neutral, Critical), alongside an indicator of whether the tweet contains a personal or an identity attack.

II. Language Variety

The data was collected via the Twitter API from August 2021 to January 2022. All tweets are in English, but we did not filter geographically. To select only English tweets, we first filtered out any tweets not in English using the Twitter API attribute then additionally asked annotators to mark when a tweet was not in English. Finally, experts removed any non-English tweets that still remained. This choice was motivated by (1) our focus on the English Premier League season, (2) the study authors' expertise and (3) the greater availability of abusive keyword lists for English language. Focusing only on English tweets limits the applicability of our dataset or our model to non-English language parts of football twitter, but our methods could be replicated to analyse these groups.

III. Speaker Demographics

The speakers are users on Twitter who author tweets. For privacy reasons, user demographics were not collected for this report. Our dataset of 13,418 tweets contains tweets authored by 12,291 unique accounts. Given general statistics on Twitter users, we expect user demographics to be skewed towards younger (25-35), urban, and male users.[47]

IV. Annotator Demographics

Crowd annotators were recruited using Appen, a crowd-sourced annotation platform that hosts annotators from a variety of countries and backgrounds. Each entry received between 3-5 annotations. Explicit annotator demographics were not analysed for this study. Each entry was also validated by an expert annotator. Expert annotators were authors of this study – i.e. English-speaking researchers with extensive subject matter expertise in online harms.

[46] As advised by Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

[47] Pew Research Center. (2019) 'Sizing Up Twitter Users'.

V. Speech Situation

All tweets were posted between August 2021 and January 2022. Tweets are restricted to 280 characters in length and represent short-form written-language documents, often containing spontaneous communications and personal responses or opinions.

VI. Text Characteristics

The genre of text is tweets directed at or mentioning footballer players within the English Premier League season. The test and validation sets are randomly sampled from the pool, each containing approximately 3.5% Abusive tweets. The training data is selected via an Iterative Training Process, and contains 24% Abusive tweets. The dataset of 13,418 tweets contains tweets directed towards 452 unique players from 12,291 unique user accounts.

Model Fitting Details

In each round, we used an uncased DeBERTa v3 base model with a sequence classification head. All model training and evaluation was implemented in Python using the transformers library from HuggingFace.[48] We finetuned the model from scratch after the end of each round because incremental training has been shown to introduce stochasticity into model performance. During the rounds of data collection, we trained a model for 3 epochs with a weighted Adam optimizer, and early-stopping on the validation set loss. Other parameters were set to HuggingFace defaults. After all our data was collected (at the end of R4), we tuned hyperparameters of the final model using a grid-search over learning rates of [5e-6,5e-4,1e-3], weight decays of [0.01, 0], warm-up steps of [0,100,500], and epochs of [2,3,4]. We used the Macro-F1 score on the validation set to select the best model, which used a learning rate of 5e-6, a weight decay of 0.01, warm-up steps of 0 and 2 epochs. Training each transformer model took approximately 10 minutes on NC12s GPU-enabled virtual machine. For active data acquisition, we implemented fast search of dense embeddings using the FAISS python library.[49] Active data acquisition took approximately 50 minutes on the same GPU machine.

[48] https://github.com/huggingface/transformers
[49] https://github.com/facebookresearch/faiss

Football-Specific Terms and Phrases

Table A1: Football-specific terms and phrases.

Term	Meaning
MOTM	Man of the Match
SIU	Ronaldo's celebration
POTM	Player of the Match
FPL	Fantasy Premier League
YNWA	You'll Never Walk Alone (positive) / You'll Never Walk Again (abusive)
COYBIG	Come On You Boys In Green
ARSWAT	Arsenal vs. Watford game
CR7	Cristiano Ronaldo
Slabhead	Nickname for Harry Maguire
YJB	A popular phrase used for individuals that support Swansea City Football Club (You Jack Bastard).

TF-IDF Scoring Method

TF-IDF stands for 'Term Frequency – Inverse Document Frequency'. It is a natural language processing (NLP) method for discerning how relevant a word or sentence is to a document or class in some corpus of documents. In our use case, we want to know how relevant a specific hashtag is to a tweet being Abusive. Our term frequency is the number of times the hashtag appears in Abusive tweets, and our inverse document frequency is the logarithm of the number of total tweets in our corpus over the total number of tweets that the hashtag is used in.

For example, imagine that we have 100 total tweets, of which 10 are Abusive, and the hashtag '#mufc' appeared in 50 tweets in total, 5 of which are Abusive. The hashtag '#losers' appears in 20 tweets in total, 5 of which are Abusive. The TF-IDF score for '#mufc' would be 1.5 (5 x log(100/50)), and the TF-IDF score for '#losers' would be 3.5 (= 5 x log(100/20)). We would consider '#losers' to be more relevant to Abusive tweets than '#mufc'. Even though the two hashtags were used in the same number of Abusive tweets, '#losers' appears in fewer tweets overall, so is a more informative term.

Abuse Received by Clubs and Nationalities, Summed by Players

The following tables show the total number of Abusive tweets and the percentage of tweets which are Abusive, summed over all players for their club. The tables show the clubs ranked by both the absolute number of tweets and the percentage of tweets which are Abusive. Note that these tables do not indicate that players were abused *because* of their club membership, rather that their club membership is correlated with different experiences of online abuse.

Table A2: Clubs with the most Abusive tweets directed at players.
This table shows Manchester United, Manchester City and Chelsea are the top three.

Club	Total number of Abusive tweets	Percentage of tweets which are Abusive
Manchester United	37,892	3.2%
Manchester City	5,213	2.4%
Chelsea	4,908	1.5%
Arsenal	3,830	2.1%
Liverpool	3,743	1.7%
Tottenham Hotspur	3,059	3.7%
Everton	1,330	3.1%
Aston Villa	1,305	2.2%
Leicester City	991	2.8%
West Ham United	755	1.6%

Table A3: Clubs with the most abuse directed at players as percentage of all tweets players receive.
This table shows Tottenham Hotspur, Manchester United and Everton are the top three.

Club	Total number of Abusive tweets	Percentage of tweets which are Abusive
Tottenham Hotspur	3,059	3.7%
Manchester United	37,892	3.2%
Everton	1,330	3.1%
Crystal Palace	504	3.0%
Leicester City	991	2.8%
Manchester City	5,213	2.4%
Aston Villa	1,305	2.2%
Arsenal	3,830	2.1%
Newcastle United	706	2.0%
Liverpool	3,743	1.7%