Kicking Prejudice: Large Language Models for Racism Classification in Soccer Discourse on Social Media

SANTOS, Guto Leoni, DOS SANTOS, Vitor Gaboardi, KEARNS, Colm, SINCLAIR, Gary, BLACK, Jack <http://orcid.org/0000-0002-1595-5083>, DOIDGE, Mark, FLETCHER, Thomas, KILVINGTON, Dan, ENDO, Patricia Takako, LISTON, Katie and LYNN, Theo

Kicking Prejudice: Large Language Models for Racism Classification in Soccer Discourse on Social Media

Guto Leoni Santos[1], Vitor Gaboardi Santos[1], Colm Kearns[1], Gary Sinclair[1], Jack Black[2], Mark Doidge[3], Thomas Fletcher[4], Dan Kilvington[4], Patricia Takako Endo[5], Katie Liston[6], and Theo Lynn[1]

[1] Dublin City University, Dublin, Ireland {guto.santos,vitorgaboardidos.santos, colm.g.kearns,gary.sinclair,theo.lynn}@dcu.ie
[2] Sheffield Hallam University j.black@shu.ac.uk
[3] Loughborough University M.Doidge@lboro.ac.uk
[4] Leeds Beckett University {T.E.Fletcher,D.J.Kilvington}@leedsbeckett.ac.uk
[5] Universidade de Pernambuco patricia.endo@upe.br
[6] Ulster University k.liston@ulster.ac.uk

Abstract. In the dynamic space of Twitter, now called X, interpersonal racism surfaces when individuals from dominant racial groups engage in behaviours that diminish and harm individuals from other racial groups. It can be manifested in various forms, including pejorative name-calling, racial slurs, stereotyping, and microaggressions. The consequences of racist speech on social media are profound, perpetuating social division, reinforcing systemic inequalities, and undermining community cohesion. In the specific context of football discourse, instances of racism and hate crimes are well-documented. Regrettably, this issue has seamlessly migrated to the football discourse on social media platforms, especially Twitter. The debate on Internet freedom and social media moderation intensifies, balancing the right to freedom of expression against the imperative to protect individuals and groups from harm. In this paper, we address the challenge of detecting racism on Twitter in the context of football by using Large Language Models (LLMs). We fine-tuned different BERT-based model architectures to classify racist content in the Twitter discourse surrounding the UEFA European Football Championships. The study aims to contribute insights into the nuanced language of hate speech in soccer discussions on Twitter while underscoring the necessity for context-sensitive model training and evaluation. Additionally, Explainable Artificial Intelligence (XAI) techniques, specifically the Integrated Gradient method, are used to enhance transparency and interpretability in the decision-making processes of the LLMs, offering a comprehensive approach to mitigating racism and offensive language in online sports discourses.

Keywords: Tweet classification · Large Language Models · XAI · BERT · RoBERTa

1 Introduction

The United Nations International Convention on the Elimination of all Forms of Racial Discrimination (ICERD) defines racism as: “any distinction, exclusion, restriction or preference based on race, colour, descent, or national or ethnic origin which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life" [50]. Interpersonal racism occurs when individuals from dominant racial groups, either socially or politically, behave in ways that diminish and harm people who belong to other racial groups [6]. It manifests in many different ways including pejorative name-calling including racial slurs, stereotyping racial or ethnic minorities as less intelligent or worthy, or enacting microaggressions, amongst others [45]. Racist speech and offensive language can perpetuate social division, reinforce systemic inequalities, and undermine community cohesion [18, 39, 46]. Furthermore, studies have consistently shown that being the target of interpersonal racism can affect mental and physical health [41, 54].

Online racism occurs on Internet-based social media or direct messaging platforms and includes disparaging remarks, symbols, images, or behaviours that inflict harm [16]. Like in the “real world", not only does being the target of racism on social media affect mental health [16], research suggests mere exposure to online racism may contribute to a variety of health issues [48].

Social media has become a ubiquitous platform for the global discourse on sports and has significantly impacted the delivery and consumption of sport [22]. It offers an unprecedented space for fans to engage with teams, players, and each other [21, 35]. Unfortunately, this virtual space has also witnessed a troubling surge in the propagation of hate speech and offensive language in sports discourses [20,27]. The issue of racism and hate crime in soccer, the world’s most popular sport, are well-documented [7, 26]. It is therefore unsurprising that this issue has also migrated to the soccer discourse on social media [15, 27, 36].

The debate on Internet freedom in the context of social media moderation centres on two primary and often conflicting values: the right to freedom of expression and the need to protect individuals and groups from harm [9, 28]. With the acquisition of Twitter by Elon Musk and changes in Twitter’s moderation policies, now called X, the discussion on the role and responsibilities of social media platforms to moderate content on their platforms has once again come to the fore. Indeed, in January 2023, more than two dozen UN-appointed independent human rights experts called out for leaders of technology companies to “urgently address posts and activities that advocate hatred, and constitute incitement to discrimination, in line with international standards for freedom of expression” [19]. However, even where such platforms had the desire to moderate racist content and offensive language, such moderation is not without challenges not least due to the sheer volume of user-generated data, the nuances of language and context, and the global diversity of cultural norms, and indeed legal frameworks, regarding freedom of speech [23, 43]. This is particularly the case in

sporting contexts, and particularly soccer, where racist and offensive language are commonplace between fans offline and online.

In the last decade, transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT) [17] and Robustly optimized BERT approach (RoBERTa) [33] have emerged that can be fine-tuned to classify text for specific domains or contexts including hate speech and offensive language [11, 37, 44, 51]. It is well established that soccer should be treated as a unique context possessing its own linguistic idiosyncrasies [12,30]. It is characterised by domain-specific terminology with cultural and regional variations. Furthermore, soccer fans have idiosyncratic ways of interacting, including chants, slogans and specific metaphors and expressions [24,25,30,40]. Brown et al. [13] argue for the centrality of soccer supporters' identity to their lives: "being a supporter is a key part of their 'real' lives: a regular, structuring part of their existence that enables them to feel belonging in the relative disorder of contemporary social formations", attesting to the impact this identity has in shaping their linguistic idiosyncrasies. This is equally true in the context of hate speech and offensive language in online soccer discourse. The field of research into online hate speech and sport has grown significantly in recent years [27] and therefore should be treated as a distinct domain for training language models.

The aim of this paper is to evaluate different Large Language Models (LLMs) for classifying racist content in the Twitter discourse for the UEFA European Football Championships (the Euros). Using an approach similar to Nasir et al. [37], we first construct a dataset for training and testing LLMs performance and then fine-tune four LLMs - a basic version of BERT, a version of BERT pretrained for hate speech classification (BERT Hate Speech), a version of BERT pretrained for twitter classification (BERTweet), and a version of RoBERTa pretrained for offensive speech (RoBERTa Offensive Speech). To understand the relationship between the input data and output classification, we use an Explainable Artificial Intelligence (XAI) technique based on the Integrated Gradient method [47]. Our work contributes to in-domain and cross-domain classification of hate speech and establishes a need for fine-tuning LLMs for context-sensitive hate speech and offensive language detection in soccer discourses on Twitter. We demonstrate how XAI techniques can be used to fine-tune the models and make a labelled dataset for evaluation and benchmarking of LLMs available. Results showed that the RoBERTa Offensive speech achieved the best performance, outperforming other versions of BERT and RoBERTa.

The rest of this paper is organised as follows: Section 2 introduces LLM and XAI. Section 3 summarises related works on the detection of hate speech and offensive language on Twitter using machine learning models. Section 4 presents the data and the methodology used to evaluate the LLM models and how we use XAI to explain the models behaviour. The results for the evaluation of LLMs performance and the outcomes from the XAI analysis are presented in 5. Limitations and avenues for future research are presented in Section 6 and Section 7 concludes the paper.

2 Background

2.1 Large Language Models

Traditional language models, such as those based on Recurrent Neural Networks (RNNs), process text sequentially, which can limit their ability to effectively capture contextual nuances. In contrast, BERT marks a significant advancement in Natural Language Processing (NLP) [17]. Leveraging the power of the Transformer architecture [52], Transformers represent a paradigm shift in sequence modelling. They process input data in parallel using attention mechanisms, enabling efficient capture of long-range dependencies. BERT's utilisation of the Transformer architecture is particularly effective in tasks such as classifying racist content in different contexts on Twitter. Its bidirectional analysis comprehensively understands the context of tweets, considering both preceding and succeeding words simultaneously. BERT employs a two-step pre-training process: (i) masked language modelling (where a random subset of words in a sentence is masked, and the model predicts them), and (ii) next sentence prediction (where the model predicts if a sentence logically follows another). This pre-training equips BERT with a deep understanding of contextualised language representations. Having been pre-trained on large corpora, such as books or Wikipedia articles, BERT is fine-tuned for specific tasks like sentiment analysis, answering questions, or, in this case, classifying racist content on Twitter. Its ability to capture nuanced dependencies in language makes it adept at discerning sentiment and context, especially in dynamic domains like social media and discourses with linguistic idiosyncrasies, such as soccer discourses.

RoBERTa builds on BERT's by introducing several modifications to BERT's architecture and training methodology [33]. RoBERTa removes the Next Sentence Prediction task and trains the model on longer text sequences, thereby enhancing its contextual understanding. It also uses dynamic masking during pre-training, which helps in learning more generalizable representations.

2.2 Explainable Artificial Intelligence

XAI is crucial in understanding the decisions made by models like BERT and RoBERTa. XAI techniques improve user trust, aid in error correction, ensure compliance with regulations, and enhance collaboration between humans and artificial intelligence systems. They are particularly important in tasks like content moderation on social media, where transparency and accountability are essential. Integrated Gradients attribute model predictions to individual input features [47]. This is valuable in understanding which words or phrases are pivotal in a model's decision-making process, such as identifying racist content in discussions about the Euros on Twitter.

Integrated Gradients is a method designed to explain the predictions of machine learning models by attributing feature importance to input variables. Considering the straight line path from the baseline x' to the input x, the gradients are computed at all points along this path. In particular, integrated gradients

refer to the integral of gradients traced along a straight line path from the baseline x' to the input x. Mathematically, let $f(x)$ represent the prediction function of the model, and x denote the input vector. The attribution $IG_i(x)$ for the ith feature is computed as follows:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{1}$$

where x' denotes the baseline input, x_i is the value of the ith feature in the input vector x, and x_i' is the corresponding value in the baseline. The integral term represents the accumulated gradients along the straight path from the baseline to the input x.

3 Related Works

Several research papers have been published on the use of machine learning and deep learning to detect hate speech on Twitter. Pitsilis et al. [42] applied multiple Long Short-Term Memory (LSTM) classifiers, combined with user characteristics, to classify hate speech on Twitter. Their approach combined outputs from various LSTM models using different ensemble strategies. The input considered was a combination of tweets and features related to the users' tendency towards hateful behavior, including racist, sexist, or neutral classes. Their results showed that different ensemble strategies yielded varying performance levels, with the highest F1-score for racism detection being 70.84%.

Benítez-Andrades et al [11] compared five different deep learning models for detecting racist and xenophobic content in Spanish tweets. This included two BERT-based models - Multilingual BERT and BETO [14] - and three other deep learning techniques - Convolutional Neural Network (CNN), LSTM, and a model combining CNN and LSTM. The BETO model outperformed all models evaluated, achieving 84.28% precision, 87.30% recall, and an 85.76% F1-score.

Lee et al. [31] introduced a new architecture, GCR-NN, combining Gated recurrent units (GRU), CNN, and an RNN model to predict the sentiment of racist tweets. They annotated tweets with racist content using TextBlob based on polarity, and then classified them as positive, negative, or neutral. The GCR-NN architecture outperformed other models cited in the literature.

Wang and Islam [53] proposed a CNN model, TextCNN, to classify gender and racial discrimination on Twitter. Their analysis revealed that the most negatively connotated words were related to Muslim, Islam, religion, ISIS, Mohammed, Jew, and other sensitive racial and religious terms. The model achieved an accuracy of 96.9% for gender discrimination and 98.4% for racial discrimination, however the authors considered the sentiment analysis of the tweets in order to detect racism and sexism content.

Finally, Vanetik et al. [51] explored the performance of a variety of models for classifying racism in tweets in the French language. They developed a dataset using tweets collected with a vocabulary of racist speech keywords. The authors compared various models including BERT, Random Forest, Logistic Regression,

and Extreme gradient boosting (XGBoost), using different text representation approaches like TF-IDF, N-grams, and BERT embeddings. They found that by combining BERT embeddings with logistic regression yielded the best for monolingual text representation and for cross-lingual and multilingual experiments.

Although all of these works make important contributions to the classification of racism on Twitter, none of them focus on the soccer context specifically. Furthermore, none of them presented XAI tools to help understand the behaviour of the models, i.e., the impact of input data on model prediction.

4 Data and Methods

Figure 1 presents the pipeline used to detect racism and identify impactful words in tweets featuring racist speech.

Fig. 1: Racism classification pipeline.

4.1 Dataset

For this study, we collected tweets associated with the Euros. Table 1 presents the hashtags used to define the dataset on Euro 2016 and Euro 2017. To define the tournament dataset for this study, we used hashtags and terms associated with each fixture (e.g., #ITAvIRL, #ITAIRL etc.) official championship hashtags (e.g., #euro2016 and 'euro 2016' etc.) and official Twitter accounts (e.g. @euro2016 etc.), and related variants. For each tournament, we collected tweets from one week before to one week after the tournament. In total, we generated datasets from eight tournaments, four women's tournaments and four men's tournaments from 2008 to 2022. We stored the tweets in a local database, allowing for advanced filtering through SQL queries.

To construct our datasets of hate speech samples, we first developed a dictionary of racist terms. The dictionary was initially populated with terms from the Hatebase project[7], a website created to assist organisations moderate online conversations and detect hate speech. We then expanded it with terms from extant literature on hate speech in soccer. Using this dictionary, we filtered potential racist tweets from our database with SQL queries. However, not all tweets selected necessarily contain racist content. Thus, a manual review by human coders was required. Ultimately, 1,048 racist tweets were identified.

[7] https://hatebase.org/

Table 1: Example of hashtags and terms used to collect the tweets about the Euros 2016 and 2017.

Euro Year	Gender	Hashtags and Terms
2016	Men	#EURO2016 "Euro 2016" #euro16 #euros #euros2016 @euro2016 @uefa @fifa #FRAvROU #FRAROU #ALBvSUI #ALBSUI #WALvSVK #WALSK #ENGvRUS #ENGRUS #TURvCRO #TURCRO #POLvNIR #POLNIR #GERvUKR #GERUKR #ESPvCZE #ESPCZE #IRLvSWE #IRLSWE #BELvITA #BELITA #AUTvHUN #AUTHUN #PvISL #PISL #RUSvSVK #RUSSK #ROUvSUI #ROUSUI #FRAvALB #FRAALB #ENGvWAL #ENGWAL #UKRvNIR #UKRNIR #GERvPOL #GERPOL #ITAvSWE #ITASWE #CZEvCRO #CZECRO #ESPvTUR #ESPTUR #BELvIRL #BELIRL #ISLvHUN #ISLHUN #PvAUT #PAUT #SUIvFRA #SUIFRA #ROUvALB #ROUALB #SVKvENG #SKENG #RUSvWAL #RUSWAL #NIRvGER #NIRGER #UKRvPOL #UKRPOL #CROvESP #CROESP #CZEvTUR #CZETUR #HUNvP #HUNP #ISLvAUT #ISLAUT #ITAvIRL #ITAIRL #SWEvBEL #SWEBEL #SUIvPOL #SUIPOL
2017	Women	#WEURO2017 #euro17 #euros #euros2017 #weuros #WEUROS2017 #weuro @UEFAWomensEuro @uefa @fifa #NEDvN #NEDN #DENvBEL #DENBEL #GERvSWE #GERSWE #ITAvRUS #ITARUS #ESPvP #ESPP #ENGvSCO #ENGSCO #AUTvSUI #AUTSUI #FRAvISL #FRAISL #NvBEL #NBEL #NEDvDEN #NEDDEN #SWEvRUS #SWERUS #GERvITA #GERITA #SCOvP #SCOP #ENGvESP #ENGESP #SUIvFRA #SUIFRA #BELvNED #BELNED #NvDEN #NDEN #SWEvGER #SWEGER #RUSvITA #RUSITA #PvENG #PENG #SCOvESP #SCOESP #FRAvAUT #FRAAUT #ISLvSUI #ISLSUI #BELvN #BELN #NEDvDEN #NEDDEN #GERvDEN #GERDEN #SWEvNED #SWENED #ENGvFRA #ENGFRA #NEDvENG #NEDENG #DENvAUT #DENAUT #NEDvDEN #NEDDEN

In addition to racist tweets, we needed a sample of non-racist tweets to fine-tune our models to distinguish between racist and non-racist content. We selected these tweets using queries that excluded terms from our racism dictionary. Human coders also reviewed these tweets to ensure that they were not related to racism but rather to the soccer context. An equal number of racist and non-racist tweets were included to create a balanced dataset. Therefore, our final dataset is composed of 2,096 tweets (1,048 racist tweets and 1,048 non-racist tweets).

Some language models in our study can handle raw text, but we decided to apply text preprocessing techniques to increase text comprehension by removing useless parts of the text or noise [29]. Therefore, after data collection, we apply preprocessing which consists of converting the text to lowercase and removing stop words, user mentions, URLs, and emojis. The dataset was split into 80% for training and 20% for model evaluation.

4.2 Large language Models

After preparing the training and testing datasets, we selected various pre-trained LLMs for fine-tuning, a process proven effective for achieving state-of-the-art performance in downstream tasks [32].

All LLMs we considered in this work are available on the hugging face platform[8]. The first model is the traditional BERT model [17]. We used the uncased

[8] `https://huggingface.co/docs/hub/index`

version of this model[9], since we converted all text to lower case during the preprocessing phase.

Given our focus on racism, a type of hate speech, we also included BERT Hate Speech, a model fine-tuned on diverse hate speech categories using 16 datasets [3]. We selected the version trained in English language data. We also considered a widely-used variation of the RoBERTa model, BERTweet[10], optimized for the unique characteristics of tweets, including short length, informal grammar, and irregular vocabulary [38]. This model was trained on large datasets of English tweets, including a dataset related to COVID-19. Special tokens were used for user mentions and URLs. Finally, we used RoBERTa Offensive, an LLM trained for various text-related tasks including offensive and hate speech detection [10].

4.3 Explainable AI

Following the fine-tuning of LLMs, we applied XAI techniques, specifically Integrated Gradients, to elucidate the relationship between input data and output classification [47]. This method assigns a score to each word in the input text, indicating its impact on the model's classification. We applied Integrated Gradients to each tweet individually to understand the impact of each word on the model classification. Additionally, we used it to identify words more directly related to racism in a soccer context. By running the model on a dataset of tweets and calculating word scores, we identified and ranked words based on their frequency and impact on the model's classification.

It's important to note that due to BERT's wordpiece tokenization [55], some words are divided into sub-word units. Integrated Gradients assigns scores to each sub-word unit, so we averaged these scores to obtain a composite score for words split into multiple units.

5 Results

Table 2 presents benchmark results for the LLM models evaluated in this study. The BERT Hate Speech model showed the lowest performance, with all metrics falling below 90%. Its recall of 80.21% indicates a significant limitation in accurately identifying racist tweets. Consequently, its F1-score, at only 87.25% is also lower compared to other models.

The basic BERT model outperformed the model specifically trained to identify hate speech. This could be due to its training on diverse datasets and multilingual models, which may have exposed it to a broader range of hate speech vocabulary than that found in our specific soccer context. The basic BERT model achieved an accuracy of 94.75% and a recall of 92.19%, marking improvements of 7.27% and 14.92%, respectively, over the BERT Hate Speech model. The substantial improvement in recall directly contributed to a 7.91% increase in the F1-score.

[9] `https://huggingface.co/bert-base-uncased`
[10] `https://huggingface.co/vinai/bertweet-base`

Table 2: Comparison of Large Language Models.

Model	Accuracy	Precision	Recall	F1-score
BERT	94.75	96.20	92.19	94.15
BERT Hate Speech	89.26	95.65	80.21	87.25
BERTweet	95.47	96.76	93.23	94.96
RoBERTa Offensive Speech	**96.18**	**97.31**	**94.27**	**95.77**

The RoBERTa models surpassed both BERT models in our racism classification task. BERTweet demonstrated an accuracy of 95.47%, a precision of 96.76%, and a recall of 93.23%. Its relatively high and similar precision and recall led to a robust F1-score of 94.96%.

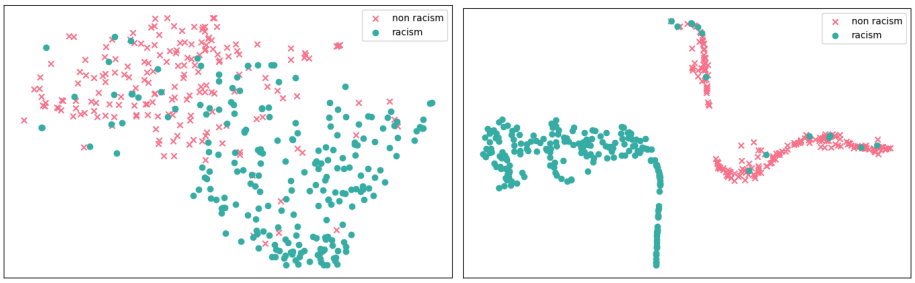
The RoBERTa Offensive Speech exhibited the highest performance among the evaluated models. It achieved an accuracy of 96.18%, which is 6.92% higher than that of the lowest-performing model (BERT Hate Speech). Its precision and recall were also superior, with the recall being 14.06% higher than the lowest observed. This resulted in the highest F1-score of 95.77%. The RoBERTa Offensive Speech model's training on a dataset of offensive tweets [56] likely contributed to its proficiency in recognising vocabulary relevant to soccer-related tweets.

Considering the tweets classified as racist by all models, most were indeed racist, as indicated by the high precision metrics. However, only the BERT, BERTweet, and RoBERTa Offensive Speech models were effective in correctly identifying racist content, as reflected in their high recall values. These results are echoed in the F1-score, which is the harmonic mean of precision and recall. Models with a high F1-score, such as the RoBERTa Offensive Speech (95.77%), can effectively minimise both false positives and false negatives in classifying racist content in tweets.

Figure 2 illustrates the embeddings for the last hidden layer considering the RoBERTa Offensive Speech before and after fine-tuning with the data presented in this paper, since it was the model the presented best results. Each point of the embedding is a vector with dimension 768, so we use the t-SNE technique to reduce the high dimensionality of the vectors to two dimensions [34]. The green dots are the racist tweets and the red crosses are the non-racist tweets. Figure 2a shows the embeddings before the fine-tuning, i.e., the embeddings with the knowledge of the RoBERTa model that was trained only to detect offensive speech. It is possible to note two different groups of tweets, since the model was trained to detect offensive speech and racist tweets tend to be offensive, showing that the model already has a good performance to represent racist and non-racist tweets. However, there is a large overlap between racist and non-racist tweets, meaning the model is not able to clearly differentiate between the two types of tweets, making classification difficult.

After fine-tuning (Figure 2b), there is a clear difference between the two categories of tweets, showing that there are two groups of tweets. Although it is possible to see the difference between the racist and no-racist tweets, there

are few racist tweets inside the non-racist tweet cluster, which compromised the performance of the model, resulting in the performance presented in Table 2.



(a) Before fine-tuning. (b) After fine-tuning.

Fig. 2: Embeddings calculated using the RoBERTa Offensive Speech.

Figures 3 and 4 summarise Integrated Gradient results for racist and non-racist tweets, respectively. It is important to note that the words shown on the horizontal axis underwent preprocessing, explaining the absence of some original tweet words.

The non-racist tweet "*Im in tears right now, we are in the final #EURO2020 #ENGDEN #ENG*" highlights *final* and *#ENG* as impactful words, directly relating to the soccer context. Conversely, the word *tears* is the word that had a minor negative impact on the model's prediction. The racist tweet "*If it weren't for the "niggers" England wouldn't of got out of group stages. You lot are shite, be grateful #eng #Euro2020Final #euro2020*", shows*niggers* as having the most significant impact, a clear racist slur. Also, the word *shite* also positively influenced the prediction, likely due to the training of the RoBERTa Offensive Speech model on offensive language. Interestingly, the word "grateful" negatively affected racism prediction, as it's not typically linked to racist contexts.

Table 3 shows the more frequent words with a high impact on the model prediction of tweets that were classified as racist. Words that are usually used to discriminate black people appear as the most frequent words (e.g. *black*, *monkey*, *niggers*, and *negro*). Two terms related to England (*English* and *England*), and the justification is that England played the final of Euro 2020 and three black players missed the penalties, which resulted in the defeat of the English team, resulting in racist reactions against them on Twitter [8].

The term *French* is also notable, likely conflating the target of the racism and is potentially linked to racism against Kylian Mbappé after he missed a crucial penalty against Switzerland, leading to France's elimination in the round of 16. Similarly, *Muslims* is identified as a significant word in the context of racism, reflecting biases and discrimination in the dataset's specific context. Historically speaking, anti-Muslim speech is not a new phenomenon [2], and it is not surprising that words related to it in datasets about racism and hate speech.

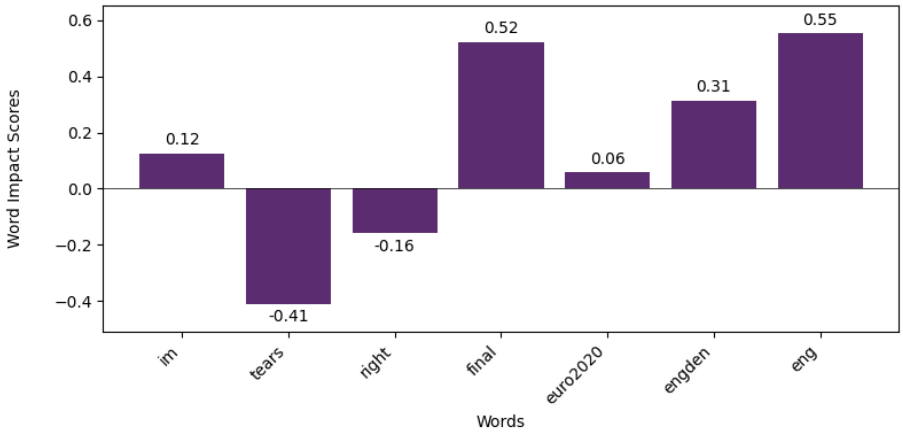


Fig. 3: Impact of words for the sentence: I'm in tears right now, we are in the final #EURO2020 #ENGDEN #ENG.

Table 3: Words that have high impact on the tweets that were classified as racist.

Word	**Frequency**
black	56
monkey	53
niggers	35
fucking	18
English	10
people	7
French	6
negro	6
England	6
Muslims	6

6 Limitations and Avenues for Future Research

In this study, we focused on four models based on BERT. Although these models have been applied successfully in different text classification tasks, including hate speech classification on Twitter, new LLMs have emerged in recent years. Meta's LLaMA, OpenAI's GPT LLMs, and PaLM 2, amongst others were proposed and are able to interpret natural language instructions and producing responses across a vast array of subjects. The last model released by OpenAI, GPT-4 [1], exhibits human-level performance on different benchmarks. Llama 2 [49] is a family of pretrained and fine-tuned open source LLMs proposed by Meta that outperformed others LLMs in the literature. PaLM 2 [5] is the model proposed by Google that was designed to deal with different languages and domains.LLMs are trained on a vast amount of public data, including tweets, which enables them

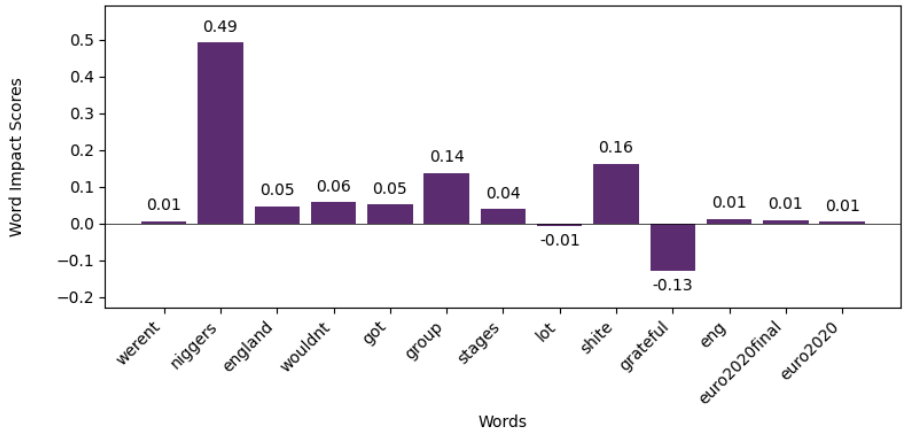


Fig. 4: Impact of words for the sentence: If it weren't for the "niggers" England wouldn't of got out of group stages. You lot are shite, be grateful #eng #Euro2020Final #euro2020.

to develop a deep understanding of the nuances of language, including those that may convey racist sentiments.

As well as evaluating other LLMs for racism detection, there is a significant opportunity for research on hybrid and ensemble models that combine various architectures like CNN, LSTM, BERT, and RoBERTa. These combined models may be more adept at capturing the subtle nuances of hate speech across different contexts and languages, thereby enhancing detection accuracy. The usage of these models can be also combined with different preprocessing techniques, from keep more information in the text (e.g. emojis) to try to change the text representation (e.g. stemming) [4]. This is critical given the negative consequences of labelling an individual incorrectly as a racist. Furthermore, it is vital to address potential biases in AI models and to uphold ethical standards in the detection and classification of hate speech. Future research must focus on developing fair and unbiased models that respect ethical guidelines and considerations.

We focus on one type of hate speech - racism. Future research should broaden its scope to encompass various forms of hate speech such as homophobic speech, ableist language, xenophobia, and religious-based hate speech, amongst others. Each of these areas presents unique linguistic characteristics and challenges, necessitating specialised attention in the development and training of models. Classifiers that distinguish between hate speech that meet the legal threshold for criminal or civil action as opposed to merely offensive language be of significant value for law enforcement, online platforms, and researchers. Furthermore, investigating both paradigmatic (stereotypical), non-paradigmatic (non-stereotypical), and appropriated slurs is another important research direction. Future studies should aim to classify these slurs effectively while also determining the targets and perpetrators of hate speech. This approach will provide a

more detailed understanding of the dynamics and patterns of online hate speech and wider online abuse.

Given that our study was primarily focused on a dataset based on one international soccer championship limited to the European continent, it is crucial to recognise that these findings might have limited applicability in other different although related contexts, such as domestic league and cup competitions, championships on other continents e.g. AFCON and the World Cup, as well as different genders. Future studies should consider fine-tuning their models to these specific scenarios to ensure both relevance and accuracy. Moreover, our study was limited to the English language and one time period, extending research to include multi-lingual datasets is essential, considering that hate speech is a pervasive issue transcending language barriers. This expansion will allow for a more comprehensive and inclusive approach to understanding and combating hate speech globally. Longitudinal research across multiple tournaments may allow for new insights on the evolution of hate speech/offensive language in a soccer context and the effectiveness of platform moderation over time.

The creation of effective tools for monitoring social media and strategies for intervention is imperative. This includes developing systems capable of real-time detection and response to mitigate the proliferation of hate speech, especially during major sporting events, but also includes deriving insights on the evolution of hate speech, triggers, perpetrators, targets, and effective responses.

As social media's role in shaping global sports discourse continues to grow, addressing the rise in hate speech and offensive language becomes increasingly urgent. This study's insights emphasise the potential of LLMs in detecting and analysing such speech in the context of soccer. Moving forward, the challenge lies in expanding research to cover a wider array of hate speech types, employing more sophisticated AI models, and adapting these models to various contexts and languages. Through these endeavours, we can better understand and confront the escalating issue of hate speech in online sports conversations.

7 Conclusions

This study delved into the critical issue of hate speech and offensive language on social media, with a particular focus on the discourse surrounding soccer. By employing LLMs to detect instances of racism on Twitter, specifically related to discussions about the Euros, we were able to gain significant insights. Among the four pre-trained models that we evaluated, the RoBERTa model, which is tailored for detecting offensive speech, emerged as the most effective. Analysis using the Integrated Gradients algorithm highlighted that derogatory terms targeting black individuals and references to French and English nationalities had the most significant impact on the models' ability to identify racist tweets, reflecting the real-world incidents of racism that occurred during the Euro Cup 2020.

Acknowledgment

This work was supported by funding from the UK Arts and Humanities Research Council and the Irish Research Council (grant number AH/W001624/1), and the Federation Internationale de l'Automobile.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Acim, R.: Islamophobia, racism and the vilification of the muslim diaspora. Islamophobia Studies Journal (2019)
3. Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A.: Deep learning models for multilingual hate speech detection. arXiv preprint arXiv:2004.06465 (2020)
4. Anandarajan, M., Hill, C., Nolan, T., Anandarajan, M., Hill, C., Nolan, T.: Text preprocessing. Practical text analytics: Maximizing the value of text data pp. 45–59 (2019)
5. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023)
6. Association, A.P., et al.: Apa resolution on harnessing psychology to combat racism: Adopting a uniform definition and understanding (2021)
7. Back, L., Crabbe, T., Solomos, J.: The changing face of football: Racism, identity and multiculture in the English game. Berg (2001)
8. Back, L., Mills, K.: 'when you score you're english, when you miss you're black': Euro 2020 and the racial politics of a penalty shoot-out. Soundings **79**(79), 110–121 (2021)
9. Balkin, J.M.: Free speech is a triangle. Colum. L. Rev. **118**, 2011 (2018)
10. Barbieri, F., Camacho-Collados, J., Neves, L., Espinosa-Anke, L.T.: Unified benchmark and comparative evaluation for tweet classification. arXiv preprint arXiv:2020.12421 (2020)
11. Benítez-Andrades, J.A., González-Jiménez, Á., López-Brea, Á., Aveleira-Mata, J., Alija-Pérez, J.M., García-Ordás, M.T.: Detecting racism and xenophobia using deep learning models on twitter data: Cnn, lstm and bert. PeerJ Computer Science **8**, e906 (2022)
12. Billings, A.C.: Defining sport communication. Taylor & Francis (2016)
13. Brown, A., Crabbe, T., Mellor, G.: Introduction: Football and community–practical and theoretical considerations. In: Football and community in the global context, pp. 1–10. Routledge (2013)
14. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. arXiv preprint arXiv:2308.02976 (2023)
15. Cullen, A., Williams, M.: Online hate speech targeting the england and wales men's football teams during the 2022 fifa world cup (2023)
16. Del Toro, J., Wang, M.T.: Online racism and mental health among black american adolescents in 2020. Journal of the American Academy of Child & Adolescent Psychiatry **62**(1), 25–36 (2023)
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

18. Dovidio, J.F., Gaertner, S.L.: On the nature of contemporary prejudice: The causes, consequences, and challenges of aversive racism. (1998)
19. Experts, U.: Freedom of speech is not freedom to spread racial hatred on social media. United Nations (2023)
20. Farrington, N., Hall, L., Kilvington, D., Price, J., Saeed, A.: Sport, racism and social media. Routledge (2017)
21. Fenton, A., Keegan, B.J., Parry, K.D.: Understanding sporting social media brand communities, place and social capital: A netnography of football fans. Communication & Sport **11**(2), 313–333 (2023)
22. Filo, K., Lock, D., Karg, A.: Sport and social media research: A review. Sport management review **18**(2), 166–181 (2015)
23. Gillespie, T.: Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press (2018)
24. Glynn, E., Brown, D.H.: Discrimination on football twitter: the role of humour in the othering of minorities. Sport in Society **26**(8), 1432–1454 (2023)
25. Hoffmann, T.: Cognitive sociolinguistic aspects of football chants: The role of social and physical context in usage-based construction grammar. Zeitschrift für Anglistik und Amerikanistik **63**(3), 273–294 (2015)
26. Kassimeris, C., Lawrence, S., Pipini, M.: Racism in football. Soccer & Society **23**(8), 824–833 (2022)
27. Kearns, C., Sinclair, G., Black, J., Doidge, M., Fletcher, T., Kilvington, D., Liston, K., Lynn, T., Rosati, P.: A scoping review of research on online hate and sport. Communication & Sport **11**(2), 402–430 (2023)
28. Klonick, K.: The new governors: The people, rules, and processes governing online speech. Harv. L. Rev. **131**, 1598 (2017)
29. Kurniasih, A., Manik, L.P.: On the role of text preprocessing in bert embedding-based dnns for classifying informal texts. Neuron **1024**(512), 927–34 (2022)
30. Lavric, E., Pisek, G., Skinner, A., Stadler, W.: The linguistics of football, vol. 38. Narr Francke Attempto Verlag (2008)
31. Lee, E., Rustam, F., Washington, P.B., El Barakaz, F., Aljedaani, W., Ashraf, I.: Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcr-nn model. IEEE Access **10**, 9717–9728 (2022)
32. Lee, J.S., Hsiang, J.: Patent classification by fine-tuning bert language model. World Patent Information **61**, 101965 (2020)
33. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
34. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
35. McDonald, H., Biscaia, R., Yoshida, M., Conduit, J., Doyle, J.P.: Customer engagement in sport: An updated review and research agenda. Journal of Sport Management **36**(3), 289–304 (2022)
36. Miranda, S., Gouveia, C., Di Fátima, B., Antunes, A.C.: Hate speech on social media: behaviour of portuguese football fans on facebook. Soccer & Society pp. 1–16 (2023)
37. Nasir, A., Sharma, A., Jaidka, K.: Llms and finetuning: Benchmarking cross-domain performance for hate speech detection. arXiv preprint arXiv:2310.18964 (2023)
38. Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for english tweets. arXiv preprint arXiv:2005.10200 (2020)

39. Pager, D., Shepherd, H.: The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. Annu. Rev. Sociol **34**, 181–209 (2008)
40. Papadima, A., Photiadis, T.: Communication in social media: Football clubs, language, and ideology. Journal of Modern Greek Studies **37**(1), 127–147 (2019)
41. Paradies, Y., Ben, J., Denson, N., Elias, A., Priest, N., Pieterse, A., Gupta, A., Kelaher, M., Gee, G.: Racism as a determinant of health: a systematic review and meta-analysis. PloS one **10**(9), e0138511 (2015)
42. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in twitter data using recurrent neural networks. Applied Intelligence **48**, 4730–4742 (2018)
43. Roberts, S.T.: Behind the screen. Yale University Press (2019)
44. Sarkar, D., Zampieri, M., Ranasinghe, T., Ororbia, A.: fbert: A neural transformer for identifying offensive content. arXiv preprint arXiv:2109.05074 (2021)
45. Staff, A.: Race and ethnicity guidelines in psychology: Promoting responsiveness and equity12
46. Sue, D.W., Capodilupo, C.M., Torino, G.C., Bucceri, J.M., Holder, A., Nadal, K.L., Esquilin, M.: Racial microaggressions in everyday life: implications for clinical practice. American psychologist **62**(4), 271 (2007)
47. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
48. Tao, X., Fisher, C.B.: Exposure to social media racial discrimination and mental health among adolescents of color. Journal of youth and adolescence pp. 1–15 (2022)
49. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
50. UNIES, N.: International convention on the elimination of all forms of racial discrimination. UN General Assembly (UNGA) (2006)
51. Vanetik, N., Mimoun, E.: Detection of racist language in french tweets. Information **13**(7), 318 (2022)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
53. Wang, L., Islam, T.: Automatic detection of cyberbullying: Racism and sexism on twitter. In: Cybersecurity in the Age of Smart Societies: Proceedings of the 14th International Conference on Global Security, Safety and Sustainability, London, September 2022. pp. 105–122. Springer (2023)
54. Williams, D.R., Mohammed, S.A.: Discrimination and racial disparities in health: evidence and needed research. Journal of behavioral medicine **32**, 20–47 (2009)
55. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
56. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666 (2019)