












Detecting Homophobic Speech in Soccer Tweets Using Large Language Models and Explainable AI

Guto Leoni Santos¹ , Vitor Gaboardi dos Santos¹ , Colm Kearns¹ , Gary Sinclair¹ , Jack Black² , Mark Doidge³ , Thomas Fletcher⁴ , Dan Kilvington⁴ , Katie Liston⁶ , Patricia Takako Endo⁵ , and Theo Lynn¹ 

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. Homophobic speech is a form of hate speech. Social media enables hate speech to spread rapidly and widely through the internet, and unlike offline hate speech, can persist indefinitely, thereby prolonging its impact. Due to the adverse impact of hate speech, policymakers have called for greater action from online platforms to moderate and remove hate speech, including homophobic content. While homophobic hate speech is prevalent in online soccer discourses, there are few studies on this empirical context in general and specifically on the use of Large Language Models (LLMs) for detecting such speech. This study addresses this gap by proposing a homophobic speech text classification pipeline. We introduce H-DICT, a new general dictionary for identifying potential homophobic content in documents, and leverage this dictionary to curate and manually label an annotated dataset of homophobic and non-homophobic samples from the UEFA European Football Championships (the Euros) discourse on Twitter. We fine-tune and evaluate five large language models (LLMs) based on the BERT architecture - BERT, DistilBERT, RoBERTa, BERT Hate, and RoBERTa Offensive - and use Integrated Gradients, an explainable AI technique to explain each model's predictions. RoBERTa Offensive, an LLM fine-tuned specifically for detecting offensive language, presented the best performance when compared to the other LLMs.

Keywords: Soccer · Hate speech classification · Homophobic speech · Large language models · Explainable AI.

1 Introduction

The Council of Europe defines hate speech as: “all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed

personal characteristics or status such as “race”, colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation” [13]. Such expressions differ in terms of severity, the damage they inflict, and their effects on specific group members in different situations [13]. Homophobia is defined as “the irrational fear of, and aversion to, homosexuality and to lesbian, gay, and bisexual people based on prejudice” [15]. As a gendered concept, homophobic speech is therefore widely accepted as a particular type of hate speech that includes derogatory and threatening language, images, and symbols towards the lesbian, gay, and bisexual community [7]. The adverse impact of hate speech is well documented including emotional and psychological harm [28], social isolation and political radicalisation [4], and in extreme cases, suicide [32]. Similarly, studies suggest that homophobic hate speech has distinct psychological effects on those targeted by such speech, including adverse effects on psychological wellbeing [46] and contributing to depression [35], amongst others [47]. With the increase in hate crimes in general and against the LGBTQ+ community specifically [16], there is increasing pressure by policymakers on social media platforms to moderate and remove homophobic content [13]. For example, the new European Union Digital Services Act (DSA) mandates online platforms to actively monitor and address issues like hate speech, with financial penalties of up to 6% of their annual global revenue for non-compliance [41].

This paper studies the automatic classification of homophobic hate speech in the soccer discourse on Twitter (now known as X). Social media has provided unprecedented access to sports-related content and opportunities for fans to engage with teams, players, and each other [17]. However, it has also enabled the rapid and widespread propagation of hate speech in the sports discourse [23, 37]. While homophobia in soccer has been widely examined in the extant literature [12, 19, 30], there is a dearth of research on online homophobic hate speech in the soccer context. A recent scoping review by Kearns et al. [23] only identifies six papers on the topic. Similarly, while there has been a rise in the number of studies on homophobic speech detection on social media, particularly using Large Language Models (LLMs) [6, 7, 9], again there are few studies in the soccer context. We argue that given the linguistic idiosyncrasies of the soccer discourse [5, 26], it should be treated as a distinct domain for training language models for homophobic speech detection and classification.

In this paper, we propose a pipeline for automatically detecting and classifying homophobic content in the Twitter discourse in soccer. We use the UEFA European Football Championships (the Euros) as the empirical context and build a dataset of eight tournaments (four men’s and four women’s) from 2008 to 2022. After building an annotated dataset, we applied LLMs to classify the tweets and understand how terms and phrases impact performance of the models.

In summary, we make the following contributions:

- A comprehensive general dictionary (H-DICT) of terms, word stems, and phrases for detecting potential homophobic speech. H-DICT is a valuable tool for identifying potential homophobic speech in any document, across various platforms, and is not limited to the soccer domain or the Twitter

platform. It can also be used for ‘bag of words’-based research or to enhance machine learning and deep learning models.

- A novel manually-annotated dataset focused on homophobic speech on Twitter during the UEFA European Football Championships over a 15-year period. This dataset provides valuable samples of different types of homophobic speech in an international soccer context.
- A performance evaluation of five LLMs fine-tuned using our annotated dataset to classify homophobic speech in textual data in a soccer context. We included models trained in a general context (e.g., Bidirectional Encoder Representations from Transformers (BERT), DistilBERT, and RoBERTa) and models that were fine-tuned to classify hate and offensive speech (BERT Hate and RoBERTa Offensive).
- An analysis of the impact of different input text on the model classification using Explainable Artificial Intelligence (XAI), most specifically Integrated Gradients, to understand why models are classifying the input text as homophobic or not, and then to identify potential improvements for enhancing the models’ performance.

Our results suggest that the RoBERTa Offensive model fine-tuned on our annotated dataset achieved the best overall performance.

The rest of this paper is organized as follows: Section 2 introduces the background on LLM and XAI. Section 3 summarises related works on detecting homophobic speech on Twitter using Machine Learning (ML) and Deep Learning (DL) models. Section 4 presents the data and methodology, highlighting the dictionary development, dataset collection and labelling, the LLMs employed for homophobic detection, and the XAI method used to understand the model’s decision-making. The results for the evaluation of LLMs performance and the outcomes from the XAI analysis are presented in Section 5. Section 6 concludes the paper and presents avenues for future research.

2 Background

2.1 Large Language Models

LLMs are advanced language models with a substantial number of parameters and trained on extensive text datasets, leading to notably improved performance across a wide range of applications [10]. The foundational technology behind LLMs is the Transformer [44] architecture, which introduces a new approach to sequence modelling by processing input data concurrently through attention mechanisms and facilitating the capture of extensive contextual dependencies. BERT (Bidirectional Encoder Representations from Transformers) [14] leverages the Transformer architecture. It has proven to be particularly effective in detecting hate speech across social media platforms, including Twitter (X) [31].

BERT employs a two-step pre-training approach. First, masked language modelling is used to predict a random subset of hidden words in a sentence. Second, sentence prediction predicts whether a given sentence follows another

coherently and logically. This pre-training provides BERT with a deep understanding of contextual language representations. Pre-trained on vast corpora such as books and Wikipedia articles, BERT is fine-tuned for specific tasks, such as sentiment analysis, question answering, or, as illustrated in this paper, text classification.

Other variations of the BERT model architecture have been developed to improve performance and address processing constraints. For instance, DistilBERT [36] is a smaller and faster version of BERT with a distinct pre-training strategy but comparable language understanding capabilities and performance. Robustly optimized BERT approach (RoBERTa) [29] eliminates the next sentence prediction task and extends the model’s training to encompass longer text sequences, enriching its contextual comprehension. Additionally, RoBERTa adopts dynamic masking throughout pre-training which leads to the acquisition of more versatile representations.

2.2 Explainable Artificial Intelligence

XAI assists in understanding the decisions generated by LLMs. It is a critical tool in applications where transparency and accountability are essential, such as content moderation on social media platforms. Integrated Gradients [39] is a method used to interpret the predictions made by ML models by attributing the model prediction to its input features. It offers insights into how the input features contribute to the model’s decision-making process.

In the context of text classification, the aim is to compute the contribution of each word or token to the prediction of the Natural Language Processing (NLP) model. The Integrated Gradients method involves creating a path consisting of a series of intermediate inputs. These are created by linearly interpolating the baseline (e.g., an empty input) and the actual input. Along this path, the model generates predictions at different steps, with the number of words or tokens changing incrementally at each step. The method then approximates the integral of gradients along this path. The rationale behind this approach lies in systematically assessing the impact of individual features on the model’s prediction at each step. This allows for the evaluation of the influence of all features throughout the entire path.

Formally, given a model f with input features x , the Integrated Gradients technique calculates the feature attributions IG_i for each feature x_i as follows:

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

where x' is the baseline input, x_i is the i -th feature of the actual input x , $\frac{\partial f}{\partial x_i}$ represents the partial derivative of the model’s output with respect to the i -th feature, α denotes the interpolation coefficient between the baseline input x' and the actual input x . By integrating the gradients along this path, this technique provides a better understanding of feature contributions across the input space. This enables stakeholders to interpret the model’s decisions and identify influential features.

3 Related Work

While there is a relatively established literature on the use of LLMs for hate speech detection [2, 11, 21], the research on the use of LLMs for homophobic speech classification on social media is relatively small. Chakravarthi et al. [7] evaluated four LLMs for detecting homophobic and transphobic content on social media in India. All models evaluated were based on the BERT architecture, namely mBERT, XLM-RoBERTa, MuRIL, and Indic-BERT. To address a lack of resources for testing and training models, the authors use a data augmentation via pseudolabeling approach. The models were evaluated using metrics for precision, recall, F1-score, and accuracy. The mBERT model achieved the best performance.

In a related work, Chakravarthi et al. [6] presented a dataset for testing and training models for classifying homophobic and transphobic content on YouTube in English, Tamil, and both English and Tamil. They also present the results of their experiment classifying content using ML and DL models and LLMs, showcasing the results from other researchers using the same dataset. They found that ensemble models worked significantly better than either ML or DL alone. A proposed model using the RoBERTa based model achieved the best results.

Nilsson et al. [34] proposed a pipeline that fine-tunes a multilingual transformer, XLM-RoBERTa, using multitask learning with SBERT as the teacher model to detect homophobic and transphobic content in YouTube comments in English, Tamil, and Malayalam. They compared single-task and multitask models and find that the former outperforms the latter.

Garcia et al. [18] addressed homophobic detection in Mexican Spanish to solve two tasks: (a) detecting whether a given tweet is homophobic and (b) identifying between different types of phobia. They conducted fine-tuning on various monolingual and multilingual LLMs, extracting sentence embeddings for each model. These embeddings were then input into a multi-input neural network. They found out that ensemble approaches lead to the best performance.

As can be seen from above, detecting homophobic speech is not a new ML/DL task, however we make a number of contributions. Firstly, we develop a general dictionary of English language words, word stems, and phrases (H-DICT) that can be used to identify potentially homophobic content, significantly expanding the available dictionaries. Furthermore, it is a general contribution in that it can be used to detect potential homophobic content in any document or platform and independent of the soccer context. This is an important contribution, as language is not static. Many new words, and particularly offensive language, make their way into use through social media. Secondly, while the LLMs we evaluate are present in the literature, such models perform differently in different contexts. We focus on a particular empirical context, soccer, which has been found to have specific linguistic idiosyncrasies that require discrete consideration [5, 26]. Consequently, we make a contribution by both using a unique annotated dataset of soccer tweets featuring homophobic language examples for testing and training ML and DL models for the homophobic speech detection task. Finally, our work applies XAI to identify individual word impact on the models' prediction. This

is critical for understanding which content features are influencing the model’s classification and, where necessary, fine-tuning training datasets and models to address common misclassifications. XAI is critical in hate speech classification, particularly on social media, to ensure that algorithms transparently balance individuals’ rights to freedom of expression in compliance with legislation such as the DSA, thereby minimising instances of unfair censorship or erroneous labelling of content as homophobic or other forms of discrimination.

4 Data and Methods

Figure 1 presents the pipeline used to create the annotated datasets, to fine-tune LLMs to classify homophobic content in the soccer discourse, and evaluate the models’ performance. The pipeline can be divided into two flows- (1) creating an annotated dataset with homophobic content, and (2) training and evaluating LLMs. In the following sections, we will describe the pipeline steps.

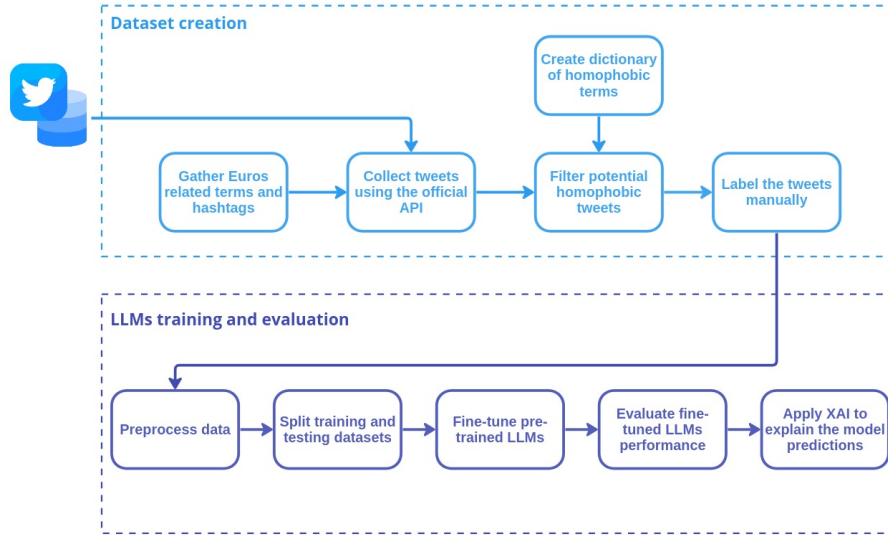


Fig. 1: Homophobic speech text classification pipeline.

4.1 Dataset

For this study, we curated tweets associated with the Euros using the Twitter Enterprise API. Spanning one week before and one week after each tournament, we collected tweets from a total of eight men’s and women’s tournaments from

2008 to 2022. Our collection criteria comprised tweets featuring the official tournament hashtag, references to the tournament name, mentions of tournament-specific usernames, UEFA, and FIFA. Furthermore, we included hashtags pertaining to all matches in the tournaments and abbreviations for team names (e.g., the match between England and Russia was tagged as #ENGvRUS and #ENGRUS). We also included the official championship hashtags (e.g., #euro2016 and ‘euro 2016’, etc.), official Twitter accounts (e.g., @euro2016, etc.). We stored all the tweets in a local database for further analysis.

To build our annotated datasets of homophobic speech, we developed a general dictionary of terms, words stems, and phrases. The dictionary (H-DICT) was initially populated with 68 terms sourced from the Hatebase project⁷, an online platform designed to aid organisations in moderating online discourse and identifying hate speech. We then expanded this dictionary with additional homophobic terms cited in existing literature and references on homophobic speech. It is important to note that when using such a dictionary, researchers need to also consider plurals, hyphenated forms, misspellings, and word combinations including those found within hashtags or in combination with player’s names (e.g., ‘Fabregay’ instead of ‘Fabregas’). For filtering purposes, it is also useful to include specific widely-used terms and phrases even where there is a common word stem e.g., word combinations with ‘gay’ or ‘homo’. We make H-DICT available on a GitHub repository⁸ that can be used and extended by another researchers.

Leveraging this dictionary, we employed SQL queries to search through our database for potential homophobic tweets. It is important to note that some terms and word stems may result in false positives due to overlaps of general and domain-specific vocabulary overlapping, as well as semantic ambiguity. For example, ‘trans’ is a term and word stem commonly used in homophobic speech but is also a common word stem in the language of soccer, e.g. “transferring to” and “transfer window”. Terms associated with the LGBTQ+ community, e.g., “pride”, may also be used in other domain-specific contexts, e.g., ‘English pride’ or ‘take pride in’. Similarly, some potentially homophobic terms and word stems may overlap with player names, e.g. Lyndon Dyke (Scotland) and Lars Bender (Germany). Consequently, a manual review by three human annotators specializing in hate speech was required to identify a sufficient volume and diversity of true homophobic tweets. The final decision on the label is defined where all three agree. For the purpose of this paper, our target dataset size was 1,000 tweets per class featuring homophobic content; the ultimate dataset size was 1,005.

In addition to gathering homophobic tweets, it was necessary to obtain a sample of non-homophobic tweets to allow the training of our models to distinguish between homophobic and non-homophobic content. This involved crafting queries that excluded items from H-DICT when selecting non-homophobic tweets. Human annotators evaluated these tweets to ensure the absence of homophobic connotations and to validate their pertinence to soccer-related discus-

⁷ <https://hatebase.org/>

⁸ <https://github.com/GutoL/H-DICT>

sions. To maintain dataset balance, an equal number of homophobic and non-homophobic tweets were incorporated. Our final dataset contains 2,010 tweets, evenly divided with 1,005 examples of homophobic tweets and an equivalent number of non-homophobic tweets.

It is important to note that there are pre-existing annotated datasets available both for hate speech (e.g. [3]) and homophobic speech (e.g. [8, 43]). The former may not contain specific words for homophobic speech, while the latter are in different languages. Additionally and as discussed in Section 1, the soccer discourse features linguistic idiosyncrasies [5, 26]. In both cases, the existing labelled datasets may not contain specific words and expressions that are related to the soccer context, and we decided to not include them in our study.

While some language models used in this study have the capability to process raw text, we opted to employ text preprocessing techniques to enhance comprehension by eliminating extraneous elements or noise [25]. Consequently, following data collection, we conducted preprocessing procedures involving the conversion of text to lowercase and the removal of stop words, user mentions, URLs, and emojis. Subsequently, the dataset was partitioned into training and testing sets, allocating randomly 80% of the samples for training purposes and reserving the remaining 20% for model evaluation.

4.2 Large language Models

After building the annotated datasets, we fine-tuned several pre-trained LLMs to classify homophobic content using the text in tweets. This process of fine-tuning LLMs has demonstrated efficacy in attaining state-of-the-art performance across various downstream tasks [27, 38].

Five models are used in this study - BERT, [14], DistilBERT [36], RoBERTa [29], BERT Hate [24], and RoBERTa Offensive [3]. All these LLMs are available on the Hugging Face platform. For BERT, we use the uncased version of BERT⁹, which does not make a difference between upper and lower case. Similarly, we employ an uncased version of DistilBERT¹⁰. The main goal of using DistilBERT architecture is to have a lighter version of BERT and check if this model is able to give competitive results against other BERT-based models. We utilize the base model of RoBERTa¹¹. Unlike BERT and DistilBERT, RoBERTa is case-sensitive. Therefore, we converted all texts to lower case for a fairer comparison. Since the main goal of this paper is to detect homophobic content on Twitter in the English language, we included BERT Hate¹², a monolingual model for hate speech classification of social media content in English language [24]. BERT Hate is a version of the base BERT pre-trained language model fine-tuned with a dataset of 20,227,765 texts, comprising YouTube comments and tweets in English. Finally, we also employed RoBERTa Offensive¹³, a version of

⁹ <https://huggingface.co/bert-base-uncased>

¹⁰ <https://huggingface.co/distilbert/distilbert-base-uncased>

¹¹ <https://huggingface.co/FacebookAI/roberta-base>

¹² https://huggingface.co/IMSyPP/hate_speech_en

¹³ <https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive>

the RoBERTa model fine-tuned with a dataset composed of 14,100 tweets to identify offensive speech [3].

The fine-tuning process was performed using the following hyperparameters: 10 epochs, 5×10^{-6} learning rate, AdamW optimizer, batch size of 8 samples, and we save the model that provided the lowest loss value during the training. The training and experiments were performed using a computer with Intel(R) Core(TM) i7-12700 CPU at 2.10GHz, 32 GB RAM, and Nvidia GeForce GTX 1660 SUPER. In order to evaluate the LLMs performance, we used traditional metrics: accuracy, precision, recall, and F1-score.

4.3 Explainable AI

After fine-tuning the LLMs, we employed Integrated Gradients to verify the correlation between input data and classification output [39]. This method has gaining popularity since it can be applied to any differentiable model, has strong theoretical foundations, and it is computationally efficient compared to other models. Integrated Gradients assign a score to each input token, indicating the token’s influence on the model’s prediction. A positive score implies that the token influenced the model’s prediction, while a negative score suggests the token had an opposing influence on the model’s prediction. By applying Integrated Gradients to some tweets, we discerned the influence of specific terms on the model’s classification, assisting us to identify words closely associated with homophobia in a soccer context. The Google wordpiece tokenization mechanism [45] divides the words into tokens, which can generate sub-word units. As the Integrated Gradients method calculates a contribution score for each token, we averaged these scores to obtain a composite score for divided words.

5 Results

Table 1 presents benchmark results for the LLM models. The BERT model was the model with the lowest performance, with all metrics around 93.5%. The DistilBERT model presented slightly better results than BERT, with all metrics around 94.4%. The superiority of DistilBERT over BERT can be explained by our limited dataset (2,010 tweets). As a result, DistilBERT converged faster than the base BERT model since it has fewer parameters to adjust during training [20].

The RoBERTa base model presented better results than the base BERT model and DistilBERT, with all the metrics around 96%. This is unsurprising given RoBERTa is trained on a significantly larger corpus than the base BERT model and features additional refinements (e.g., dynamic masking) to the base BERT model. This outperformance of RoBERTa over BERT is consistent with existing literature [33].

Models already fine-tuned for hate or offensive speech showed the best results for classifying homophobic content in our experiments. The BERT Hate model presented a small improvement when compared to RoBERTa with around 96.1% for all metrics. The RoBERTa Offensive model outperformed all models

Table 1: Summary metrics for each fine-tuned LLM.

Model	Accuracy	Precision	Recall	F1-score
BERT	93.5323	93.5660	93.5323	93.5282
DistilBERT	94.4030	94.4118	94.4030	94.4037
RoBERTa	96.0199	96.0216	96.0199	96.0201
BERT Hate	96.1443	96.1698	96.1443	96.1450
RoBERTa Offensive	97.0149	97.0165	97.0149	97.0151

in the homophobia detection task, presenting all metrics at circa 97.01%. The RoBERTa Offensive model improved all metrics by 3.5%-points over the base BERT model, which performed the worst.

In order to provide a qualitative analysis, Figure 2 shows the embeddings for the BERT and RoBERTa Offensive models, the worst and best models, respectively. To compute the embeddings, we selected randomly 100 samples of homophobic and non-homophobic tweets from the testing dataset. The tweets were fed into the fine-tuned models, and the embedding representation is the output of the last layer.

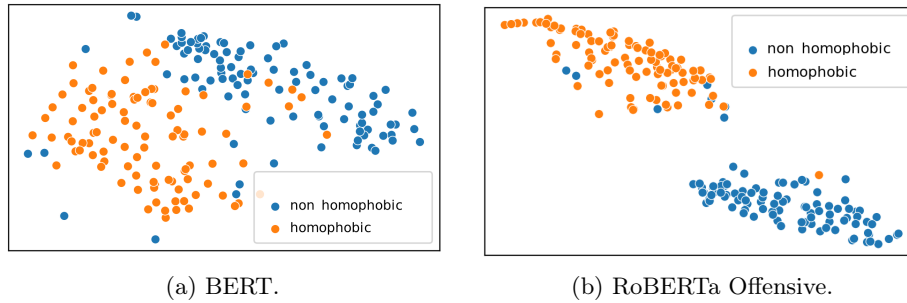


Fig. 2: Embeddings for 100 samples from the testing dataset.

Figure 2a illustrates that, even though the BERT model was able to separate the two groups of tweets, there are some points of homophobic tweets that are within the group of non-homophobic tweets. Similarly, in the lower left corner of Figure 2a, there are some non-homophobic tweets that are within the homophobic group. These findings may indicate misclassifications that adversely affected the model's performance.

On the other hand, the RoBERTa Offensive model created more disjoint groups, as shown in Figure 2b. The group of non-homophobic tweets is located in the lower right corner, while the homophobic tweets group is located in the upper left corner. There are few non-homophobic tweets that are in the homophobic group, which can be considered false positives. In the context of classifying homophobic content, false positives are less harmful than false negatives, as a

tweet that is not homophobic could be classified as homophobic. However, a tweet that is homophobic being classified as non-homophobic is more harmful, since harmful content would go unnoticed by the model. As shown in Figure 2b, only one tweet that is homophobic is in the non-homophobic group, shown that the RoBERTa Offensive model was able to identify false negatives.

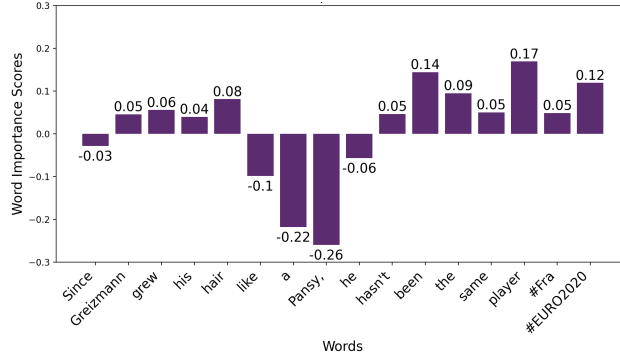
Table 2 shows three examples of homophobic tweets and the predictions made by the BERT and RoBERTa Offensive models. The use of the term “gayest” in Example 1 implies that Mexican Waves, a common stadium activity during sports events, are somehow frivolous, silly, or lacking in masculinity. Using “gay” as a pejorative term in this way is disrespectful and reinforces harmful stereotypes about LGBTQ+ individuals. Both BERT and RoBERTa Offensive models were able to identify the homophobic content in this text.

Table 2: Classification of homophobic tweets with BERT and RoBERTa Offensive models.

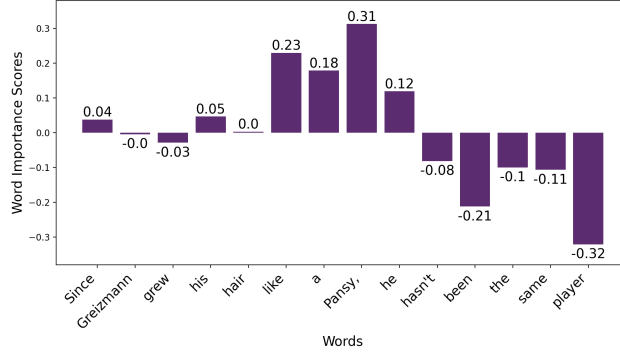
#	Tweet Text	BERT Prediction	RoBERTa Prediction
1	Mexican Waves are the gayest thing ever. #Wimbledon #Euro2012	homophobic	homophobic
2	Football now a game for pansies! #SWISPA #EURO2020	non-homophobic	homophobic
3	Since Greizmann grew his hair like a Pansy, he hasn’t been the same player #Fra #EURO2020	non-homophobic	non-homophobic

In Example 2, the user uses the term “pansies” which is often used as a derogatory slang term to insult someone’s masculinity or imply that they are weak or effeminate, which can perpetuate harmful stereotypes about gender and sexuality. Considering the context of the tweet, it suggests that soccer is becoming less masculine or less tough, implicitly implying that toughness or masculinity is a necessary or desirable quality in sports which in turn can contribute to a culture of toxic masculinity. The BERT model was not able to detect the homophobic content in this tweet, while RoBERTa Offensive model was able to identify that the word “pansies” was used in an offensive way. Example 3 shows an example of homophobic content where both models were not able to classify the text correctly. Similar to Example 2, the term “pansy” is used in an offensive way to insult Griezmann’s hairstyle. Again, this language perpetuates harmful stereotypes about masculinity and implies that adopting a certain hairstyle or appearance is undesirable or weak. Therefore, this speech is disrespectful not only to Griezmann, but also to individuals who may choose to express themselves through their appearance in various ways. For this example, both models were not able to classify it correctly, i.e., even the RoBERTa model, which was the model with the best performance, was not able to classify it correctly.

To better understand the misclassification of the RoBERTa Offensive model in Example 3, we used the Integrated Gradients technique. Figure 3 shows the contribution score of each token on the model’s prediction. In this case, we are analysing the predictions that say that a homophobic tweet was not homophobic (a false negative example). Figure 3a indicates that the word “Pansy” has the highest negative value, suggesting that it influences the model prediction towards a negative result, in this case it indicates the content is likely homophobic. Despite the expected negative score for this word due to its homophobic connotation, the final model classification was non-homophobic. We believe this misclassification is likely related to the positive weighting RoBERTa gave to the remaining tokens relative to the non-homophobic class. For instance, the token “#EURO2020” had a relatively high score towards the classification, while it should be more neutral.



(a) Example of tweet text including hashtags misclassified as non-homophobic.



(b) Example of tweet text excluding hashtags classified correctly as homophobic.

Fig. 3: Integrated Gradients results for Example 3 in Table 2 misclassified by the RoBERTa Offensive model.

As a result, we removed the hashtags words (“#EURO2020” and “#FRA”) to evaluate the model performance for this specific example, a common pre-processing procedure when dealing the tweets [31]. As shown in Figure 3b, the model now correctly classified the tweet as homophobic and the sequence of words “like a Pansy” had the highest positives scores, indicating the impact of these tokens on the prediction. These findings are important as it highlights the benefit of using XAI for understanding the relative impact of unexpected content features on the classification task. Using XAI merely for homophobic content would nearly be pointless given the use of the H-DICT dictionary. However, XAI provides a unique insight on the impact of non-homophobic terms, in this case the tournament hashtags. Once corrected, the model’s performance is further improved. We hypothesise that this misclassification occurred due to a higher incidence of terms like “#EURO2020” in the non-homophobic samples during the training phase, which lead to a slightly higher influence of these words into the non-homophobic class.

6 Conclusion

This research explored the critical issue of hate speech and derogatory language on social media platforms, with a specific focus on homophobic content in the online soccer discourse on Twitter. We developed H-DICT, a general dictionary of terms, word stems, and phrases that can be used to identify potentially homophobic content in documents. Using H-DICT, we created an annotated dataset for training and testing models for classifying homophobic speech specifically in soccer. Using this annotated dataset, we then fine-tuned five variations of BERT-based models to classify homophobic speech in English language tweets from the Euros discourse on Twitter. Our performance evaluation of these LLMs, suggested that the RoBERTa Offensive model performed best, with all the metrics circa 97.01%. Furthermore, we used Integrated Gradients to explain how specific tokens contribute to the model prediction. This analysis helped us understand the reason why some tweets were misclassified as non-homophobic, despite the models’ ability to identify the key tokens that rendered the tweets homophobic.

Homophobia is not confined to any single language or culture; it manifests in various forms worldwide. Expanding the detection of homophobic speech beyond English is crucial for fostering inclusivity and combatting discrimination on a global scale. In future research, we plan to expand H-DICT with neologisms from sources such as Urban Dictionary and for international languages. Furthermore, we plan to use H-DICT to create datasets from different online platforms (e.g. Facebook, Instagram, and TikTok) and expand for different sports, in order to create a larger and more diverse dataset. This will enable more generalised use, but also greater insights when operationalised on full datasets. As shown Figure 3, removing the hashtags resulted in an improvement in the model classification. Therefore, we plan to use XAI to identify terms, which can be considered noise, that can be removed during the preprocessing step. In this paper, we used BERT-based models exclusively. In the future, we will evaluate emerging models

including Mistral [22], LLaMA [42], and API-based LLM services such as OpenAI GPT models [1] and Google Gemini [40]. Through fine-tuning, these LLMs may be more effective in classifying homophobic hate speech and in particular identifying more nuanced linguistic patterns indicative of homophobic speech.

Acknowledgment

The research in this paper was partially funded by 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. Anjum, Katarya, R.: Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security* pp. 1–32 (2023)
3. 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)
4. Bilewicz, M., Soral, W.: Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology* **41**, 3–33 (2020)
5. Billings, A.C.: *Defining sport communication*. Taylor & Francis (2016)
6. Chakravarthi, B.R.: Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics* pp. 1–20 (2023)
7. Chakravarthi, B.R., Hande, A., Ponnusamy, R., Kumaresan, P.K., Priyadharshini, R.: How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights* **2**(2), 100119 (2022)
8. Chakravarthi, B.R., Priyadharshini, R., Ponnusamy, R., Kumaresan, P.K., Sampath, K., Thenmozhi, D., Thangasamy, S., Nallathambi, R., McCrae, J.P.: Dataset for identification of homophobia and transphobia in multilingual youtube comments. arXiv preprint arXiv:2109.00227 (2021)
9. Chanda, S., Mishra, A., Pal, S.: Sentiment analysis and homophobia detection of code-mixed dravidian languages leveraging pre-trained model and word-level language tag. In: *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid)*. CEUR (2022)
10. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* (2023)
11. Chiu, K.L., Collins, A., Alexander, R.: Detecting hate speech with gpt-3. arXiv preprint arXiv:2103.12407 (2021)
12. Cleland, J., MacDonald, C.: Social media, digital technology, and masculinity in sport. In: *Sport, Social Media, and Digital Technology: Sociological Approaches*, pp. 49–66. Emerald Publishing Limited (2022)

13. Council of Europe: Combating hate speech. Council of Europe (2022)
14. 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)
15. European Union Agency for Fundamental Rights: Homophobia and discrimination on grounds of sexual orientation and gender identity in the EU member states: Part II-The social situation. European Union Agency for Fundamental Rights (2009)
16. FBI: FBI releases 2022 crime in the nation statistics. FBI (2023)
17. 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)
18. García-Díaz, J.A., Jiménez-Zafra, S.M., Valencia-García, R.: Umuteam at homomex 2023: Fine-tuning large language models integration for solving hate-speech detection in mexican spanish (2023)
19. 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)
20. Gupta, P., Gandhi, S., Chakravarthi, B.R.: Leveraging transfer learning techniques—bert, roberta, albert and distilbert for fake review detection. In: *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*. pp. 75–82 (2021)
21. Jahan, M.S., Oussalah, M.: A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing* p. 126232 (2023)
22. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. *arXiv preprint arXiv:2310.06825* (2023)
23. 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)
24. Kralj Novak, P., Scantamburlo, T., Pelicon, A., Cinelli, M., Mozetič, I., Zollo, F.: Handling disagreement in hate speech modelling. In: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. pp. 681–695. Springer (2022)
25. 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)
26. Lavric, E., Pisek, G., Skinner, A., Stadler, W.: *The linguistics of football*, vol. 38. Narr Francke Attempto Verlag (2008)
27. Lee, J.S., Hsiang, J.: Patent classification by fine-tuning bert language model. *World Patent Information* **61**, 101965 (2020)
28. Leets, L., Giles, H.: Words as weapons—when do they wound? investigations of harmful speech. *Human Communication Research* **24**(2), 260–301 (1997)
29. 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)
30. Magrath, R.: ‘to try and gain an advantage for my team’: Homophobic and homo-sexually themed chanting among english football fans. *Sociology* **52**(4), 709–726 (2018)
31. Mozafari, M., Farahbakhsh, R., Crespi, N.: A bert-based transfer learning approach for hate speech detection in online social media. In: *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019* 8. pp. 928–940. Springer (2020)

32. Mullen, B., Smyth, J.M.: Immigrant suicide rates as a function of ethnophaulisms: Hate speech predicts death. *Psychosomatic Medicine* **66**(3), 343–348 (2004)
33. Murarka, A., Radhakrishnan, B., Ravichandran, S.: Detection and classification of mental illnesses on social media using roberta. *arXiv preprint arXiv:2011.11226* (2020)
34. Nilsson, F., Al-Azzawi, S.S.S., Kovács, G.: Leveraging sentiment data for the detection of homophobic/transphobic content in a multi-task, multi-lingual setting using transformers. In: 14th Forum for Information Retrieval Evaluation, FIRE 2022, December 9-13, 2022, Kolkata, India. vol. 3395, pp. 196–207. CEUR-WS (2022)
35. Polders, L.A.: Factors affecting vulnerability to depression among gay men and lesbian women. Ph.D. thesis, University of South Africa (2006)
36. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019)
37. Santos, G.L., dos Santos, V.G., Kearns, C., Sinclair, G., Black, J., Doidge, M., Fletcher, T., Kilvington, D., Endo, P.T., Liston, K., et al.: Kicking prejudice: Large language models for racism classification in soccer discourse on social media. In: International Conference on Advanced Information Systems Engineering. pp. 547–562. Springer (2024)
38. dos Santos, V.G., Santos, G.L., Lynn, T., Benatallah, B.: Identifying citizen-related issues from social media using llm-based data augmentation. In: International Conference on Advanced Information Systems Engineering. pp. 531–546. Springer (2024)
39. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
40. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023)
41. Tourkochoriti, I.: The digital services act and the eu as the global regulator of the internet. *Chi. J. Int'l L.* **24**, 129 (2023)
42. 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)
43. Vásquez, J., Andersen, S., Bel-Enguix, G., Gómez-Adorno, H., Ojeda-Trueba, S.L.: Homo-mex: A mexican spanish annotated corpus for lgbt+ phobia detection on twitter. In: The 7th Workshop on Online Abuse and Harms (WOAH). pp. 202–214 (2023)
44. 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)
45. 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)
46. Zochniak, K., Lewicka, O., Wybrańska, Z., Bilewicz, M.: Homophobic hate speech affects well-being of highly identified lgbt people. *Journal of Language and Social Psychology* p. 0261927X231174569 (2023)
47. Ștefăniță, O., Buf, D.M.: Hate speech in social media and its effects on the lgbt community: A review of the current research. *Romanian Journal of Communication and Public Relations* **23**(1), 47–55 (2021)