Dismantling Hate: Understanding Hate Speech Trends Against NBA Athletes

Edinam Kofi Klutse, Samuel Nuamah-Amoabeng, Hanjia Lyu(✉), and Jiebo Luo

University of Rochester, Rochester, NY 14627, USA
hlyu5@ur.rochester.edu, jluo@cs.rochester.edu

Abstract. Social media has emerged as a popular platform for sports fans to express their opinions regarding athletes' performance. The National Basketball Association (NBA) is widely recognized as one of the most popular sports leagues globally. However, an unfortunate aspect that has emerged in recent years is the presence of abusive fans within the league. Consequently, the focus of this research is to identify which NBA athletes experience abuse on Twitter and delve deeper into the underlying reasons behind such mistreatment. To address the research questions at hand, the study employs a curated set of keywords to query the Twitter API, gathering a comprehensive collection of tweets that potentially contain hate speech directed toward NBA players. A deep learning classification model is implemented, effectively identifying tweets that genuinely exhibit hate speech. We further use keyword search methods to detect the specific groups that are targeted by hate speech the most and identify topics of hate speech tweets. The findings of our research indicate that certain groups of athletes are particularly vulnerable to hate speech from fans. Notably, high-performing athletes, `Black` athletes, overweight athletes, short athletes, and athletes associated with the `LGBTQ` community are found to be highly susceptible to abusive remarks. Racism, physique shaming, play style, and anti-LGBTQ remarks are the major themes. These findings contribute to a broader understanding of the challenges faced by NBA athletes in the digital space and provide a foundation for developing strategies to combat hate speech and foster a more inclusive environment for all individuals involved in the NBA community.

Keywords: Hate speech · NBA · Social media · Natural language processing

1 Introduction

In recent years, professional athletes in the National Basketball Association (NBA) have increasingly expressed their concerns about being subjected to hatred and abuse from fans and media personnel on various social media platforms [11]. Among these platforms, Twitter has emerged as a prominent arena where fans can directly engage with players, making it a hotspot for hate speech

R. Thomson et al. (Eds.): SBP-BRiMS 2023, LNCS 14161, pp. 74–84, 2023.
https://doi.org/10.1007/978-3-031-43129-6_8

directed toward NBA athletes. Unfortunately, the prevalence of derogatory language and abusive behavior on Twitter persists despite efforts to combat it [11]. Consequently, basketball players in the NBA have become a vulnerable target group for hate speech abuse.

Within this context, it is essential to address two key questions: Who are the athletes experiencing abuse? And what are the underlying reasons behind this mistreatment? In this study, we curate a set of hate speech-related keywords to collect tweets that potentially contain hateful content against NBA players. We then employ a deep learning model to detect hate speech tweets. The keyword search methods are used to detect the specific groups of athletes that are targeted by hate speech. By analyzing the collected data, the study aims to uncover the major themes prevalent in these hate speech tweets. Next, we conduct correlation analysis on a series of players' performance statistics, their demographics, as well as their physical characteristics. Our study seeks to obtain insights into the underlying motivations behind hate speech abuse in the NBA.

2 Method

2.1 Hate Speech Detection

Detecting hateful content on Twitter is not a trivial task because users may use certain codes to avoid detection by automated systems [8,9]. Other challenges may include linguistic subtleties, varying definitions of hate speech, and limited access to data for training and testing such systems [7]. In our study, we first use keywords to collect tweets that may contain hateful content and then employ a transformer-based language model to perform the final classification.

Data Collection. We use Twitter's API - Tweepy, to gather tweets containing potential hate speech targeting NBA players. To collect such tweets, we first compile a list of hate speech-related keywords. Previous research has indicated that online hate speech can stem from various motivations, including but not limited to racial discrimination, gender-based targeting, and body shaming.[1] For instance, Powell *et al.* [10] found that transgender individuals experience higher rates of digital harassment and abuse overall, and higher rates of sexual, sexuality, and gender-based harassment and abuse, as compared with heterosexual cisgender individuals. By employing this methodology, we aim to gather a dataset that captures the diverse manifestations of hate speech directed at NBA players on social media. In particular, the keyword list is composed of *nigger*, *nigga*, *bitch*, *b*tch*, *n*gg*r*, *fuck*, *bum*, *motherfucker*, *bollock*, *wanker*, *dirty*, *lame*, *bozo*, *faggot*, *pussy*, *f*ck*, *piece of shit*, *sh*t*, *bastard*, *cock*, *gay*, *lesbian*, *fucker*, *fool*, *cunt*, *asshole*, *hate*, *stupid*, *useless*, *fraud*, *cost me*, *owe me*, *lost money*, *liar*, *trash*, *ass*, *overrated*, *flop*, *flopper*, *flopping*, *coward*, *choker*, *choke artist*, *loser*,

[1] https://www.news24.com/sport/tennis/commentator-dokic-hits-out-at-fat-shaming-trolls-at-australian-open-20230123.

choking, *selfish*, *stat padder*, *ball hog*, *stat pad*, *soft*, *weak*, *retard*, *prick*, *dick*, *dickhead*. The combinations of the names of current NBA players ($n = 461$, obtained from basketballreference.com) and hate speech-related keywords are used to query tweets through Tweepy. In the end, we identify a total of 503,424 tweets of potential hate speech targeting current NBA players.

Modeling. A tweet that contains hate speech-related lexicons might be an instance of *offensive language* instead of hate speech which is defined as "language that is used to express hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" [2]. Therefore, we further leverage a transformer-based language model to detect hate speech from the collected tweets. Transformer-based models have demonstrated exceptional performance in text classification tasks across various domains [1,6,13]. In particular, we first use an open-source hate speech dataset built by Davidson *et al.* [2] to train a BERT model [3]. We then use the trained model to detect hate speech from our data corpus.

The dataset of Davidson *et al.* [2] contains 24,783 tweets of three categories - *hate speech*, *offensive language*, and *neither*. To facilitate model training and evaluation, the dataset is split, allocating 90% for the training set and the remaining 10% for the testing set. We preprocess the dataset by removing stop words using the `wordcloud` package. We then use the `bert_en_uncased_preprocess` model to convert plain text inputs into tokens that are expected by BERT. The classifier is composed of a BERT encoder and an MLP prediction head. In particular, we choose the pre-trained `BERT-Small` model as the encoder, featuring four hidden layers composed of 512 nodes each. We opt for `BERT-Small` because of its capability in achieving *adequate* classification performance, while also being *efficient* in terms of computational requirements. The MLP module consists of three components: a dense layer, a dropout layer (dropout rate = 0.2) [12], and another dense layer for predicting labels. We use ReLU activations. The model undergoes a total of 80 epochs. The learning rate is 3×10^{-5}. To optimize the training process, we employ the AdamW optimizer [5] with a weight decay set to 0.

The model achieves an overall accuracy of 91.04 on the testing set of Davidson *et al.* [2], suggesting a good performance in hate speech detection. However, it is important to note that although the dataset of Davidson *et al.* [2] provides a valuable resource, the domain of our dataset *may not perfectly align* with theirs. Consequently, any potential domain shift between the two datasets may impact the model's performance when applied to our specific dataset. As a result, we further conduct an experiment to verify the robustness of the trained model on our dataset.

Robustness Verification. We sample another validation set of 150 tweets from our dataset. Three researchers read the tweets and independently label them into three categories (*i.e.*, hate speech, offensive language, and neither). The final label is assigned with the consensus votes from three annotators. The

Table 1. Top 10 hate speech keywords related to NBA athletes.

Rank	Word	Frequency
1	ass	1,786
2	hate	1,693
3	gay	801
4	stupid	781
5	people	627
6	white	617
7	man	570
8	nigger	541
9	dirty	463
10	racist	436

Fleiss' Kappa score of the three annotators is 0.35, indicating fair agreement. Subsequently, we evaluate the performance of our classifier using this manually labeled dataset. This three-class classifier achieves an accuracy of 79.33. Moreover, it exhibits a weighted F1 score of 79.59, a precision of 80.96, and a recall rate of 79.33. These results collectively demonstrate a commendable performance for a three-class classification problem. Finally, we apply our model to the entire collected tweets.

3 Results

From the dataset comprising 503,424 collected tweets, we find 3.33% ($n = 16{,}784$) of the tweets are classified as hate speech, and 60.11% ($n = 302{,}605$) are offensive language. The remaining 36.56% ($n = 184{,}033$) of the tweets are neither hate speech nor offensive language. We remove stopwords and apply lemmatization and tokenization to hate speech tweets. Table 1 shows the top 10 words that appear most frequently in hate speech on NBA athletes.

To identify the NBA athletes who were targeted by hate speech the most, we use the keyword search method. In particular, by leveraging the extracted player names and Twitter handles, we discover the top 50 NBA athletes who are subjected to the highest levels of hateful content. Table 2 shows the top 10 NBA athletes with the most associated hate speech tweets. Notably, the list of the 50 most hated athletes includes popular names such as Lebron James, Kevin Durant, Ja Morant, Steph Curry, Devin Booker, Anthony Davis, *etc.* Two primary reasons can contribute to the observed phenomenon. Firstly, popular players often attract more attention and discussions, thereby increasing the likelihood of encountering hateful content. The prominence of these players within the NBA creates a higher probability of hate speech directed toward them. Secondly, high-profile players and notable Twitter accounts tend to become targets for hate speech due to the potential for amplified online visibility [4].

To further characterize the targets of hate speech on NBA athletes, we use different sets of keywords to search for relevant tweets. The targeted groups mined are `Black`, `White`, `Jews`, `dirty players`, `LGBTQ`, `chokers`, `selfish players`, `fat players`, `racists`, and `short players`. Table 3 summarizes the keywords used for each group.

Table 2. Top 10 NBA athletes with the most associated hate speech tweets.

Rank	Player	# Tweets
1	Anthony Davis	3,211
2	Ja Morant	2,469
3	Anthony Edwards	2,173
4	Mckinley Wright IV	1,199
5	Lonnie Walker IV	1,199
6	Alex Len	892
7	LeBron James	784
8	Russell Westbrook	596
9	Chris Paul	562
10	Kevin Durant	539

Table 3. Keywords of the targets of hate speech on NBA athletes.

Group	Keywords
`Black`	nigger, nigga, n*gg*r, black, niggers
`White`	white
`Jews`	jews
`Dirty player`	dirty, flop, flopper, flopping
`LGBTQ`	faggot, gay, lesbian
`Choker`	choker, choke artist
`Selfish player`	selfish, stat padder, ball hog, stat pad
`Fat player`	fat
`Racist`	racist
`Short player`	short, little, small

The group that experiences the highest degree of targeting is the `Black` community, with a significant count of 4,124 tweets specifically directed toward them. It is worth noting that out of the top 50 NBA athletes that are associated with the most hate speech tweets, 48 are of African descent, while 2 are

of Caucasian descent. However, in 2022, approximately 71.8% of NBA players were African American.[2] These statistics raise important questions about the potential influence of racial bias in the criticism directed toward athletes.

Following closely, the `LGBTQ` community faces a substantial number of 2,938 tweets aimed at their community. The count of tweets targeting the `White` individuals ranks third, totaling 1,035 tweets. In fourth place, there are 698 tweets directed toward `dirty players`. Additionally, 470 tweets specifically target `selfish players`, while 468 tweets aim at individuals characterized as `racists`. Moreover, there are 212 tweets targeting `fat players` and 199 tweets focusing on `short players`. The `Jewish` community is the subject of 130 tweets, and 64 tweets are directed at individuals referred to as `chokers`. Figure 1 shows the tweet distribution of targeted groups.

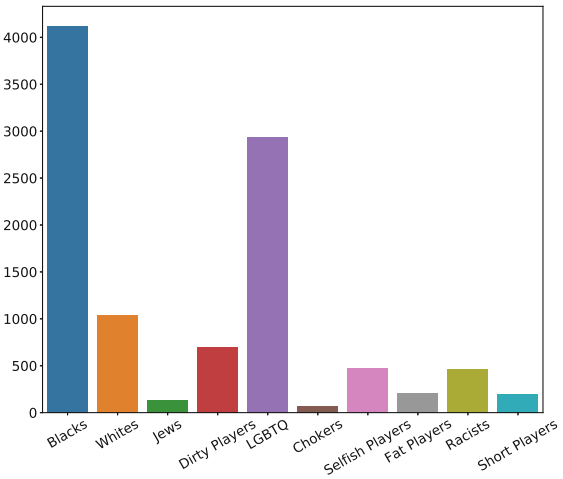

Fig. 1. Distribution of tweets related to the targeted groups of hate speech against NBA athletes.

Upon identifying the targeted groups within the hate tweets, these categories are subsequently organized into distinct topics, namely racism, physique shaming, play style, and anti-LGBTQ sentiments. More specifically, tweets about `Black`, `White`, and `Jews` are grouped into the racism topic. Tweets about `fat players` and `short players` are included in the physique shaming topic. The play style topic contains tweets about `selfish players` and `chokers`. Tweets about `LGBTQ` are included in the anti-LGBTQ topic. This classification enables a more comprehensive understanding of the underlying themes present within hate speech. The topic that emerges as the most prevalent is racism, with a support count of 5,289 instances. Following closely, the topic of anti-LGBTQ exhibits a

[2] https://43530132-36e9-4f52-811a-182c7a91933b.filesusr.com/ugd/403016_901e54ed015c44fb83df939d2070dc17.pdf.

support count of 2,940. Play style, on the other hand, garners a support count of 534, while physique shaming records a support count of 411. These figures highlight the relative prominence and occurrence of each topic within the analyzed hate speech tweets. Figure 2 shows the distribution of these topics.

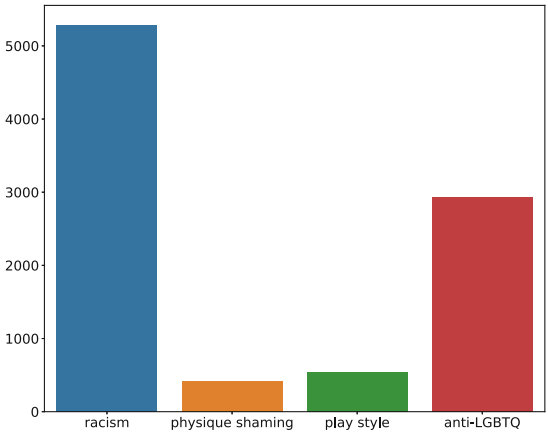

Fig. 2. Topic distributions of hate speech tweets related to NBA athletes.

To understand the potential correlation between hate speech tweets and players' performance, we compute the correlation coefficients of the number of hate speech tweets and a series of players' performance statistics, player demographics as well as their physical characteristics. The performance statistics of the NBA athletes are collected from basketballreference.com. Variables include:

- **Age**
- **G: Games Played.** The number of games in which a player has participated.
- **GS: Games Started.** The number of games in which a player was listed as a starter in the team's lineup.
- **MP: Minutes Played.** The number of minutes a player has been on the court during games.
- **TOV: Turnovers.** The number of times a player loses possession of the ball to the opposing team through errors such as bad passes, mishandling the ball, or offensive fouls.
- **Impact:** A player's influence or effect on the game. It encompasses various aspects of a player's performance that contribute to their team's success. The impact of a player can be evaluated through a combination of statistics, observations, and contextual analysis.
- **TS%: True Shooting Percentage.** It measures a player's shooting efficiency by taking into account their field goals, three-pointers, and free throws.
- **Usage:** It is a metric that quantifies the percentage of team plays or possessions that a player uses while they are on the court. Usage rate helps evaluate

the level of involvement and offensive responsibility a player has within their team's offensive system.
- **BMI: Body Mass Index.** It is a measure used to assess body composition and provide an indication of whether a person's weight is within a healthy range relative to their height.

The results revealed that the number of hate tweets demonstrated positive correlations with GS (Games Started), MP (Minutes Played), TOV (Turnovers), Impact, TS% (True Shooting Percentage), usage, and BMI (Body Mass Index). Conversely, hate tweet frequency showed negative correlations with age and G (Games Played) (Fig. 3).

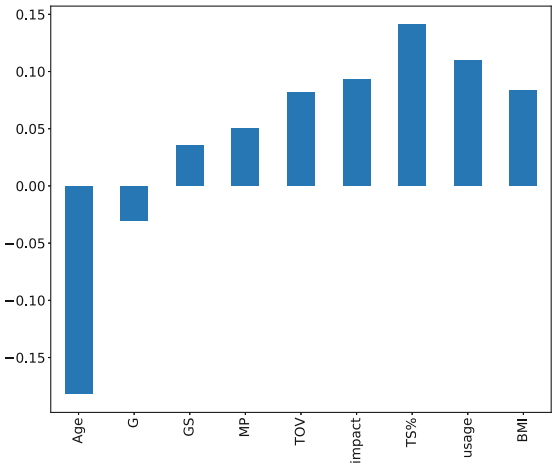

Fig. 3. Correlation coefficients between the number of hate speech tweets and variables including performance statistics, demographics, and physical characteristics of the top 50 most hated NBA athletes.

However, it is worth noting that MP (Minutes Played), TOV (Turnovers), Impact, GS (Games Started), TS% (True Shooting Percentage), and usage exhibit strong correlations with each other (Fig. 4). This suggests that their correlations with the number of hate tweets may be attributed to the fact that they are all performance metrics. Our analysis reveals that players who excel in their performance often become targets of hate speech, likely stemming from rival fans and individuals who may have financial stakes in outcomes, such as bettors.

Regarding the positive correlation observed between BMI and the number of hate speech tweets, we discover that a significant portion of the top 50 most hated NBA athletes consists of individuals categorized as overweight. Specifically, among these athletes, 17 individuals have a BMI exceeding 25. This suggests that their weight status might make them susceptible targets for fat shaming or height shaming through hate speech on social media platforms.

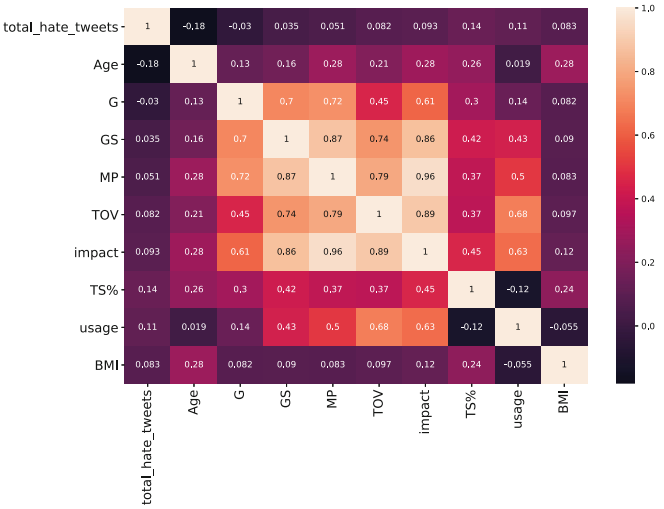

Fig. 4. Heatmap of correlations between the attributes of top 50 hated NBA athletes.

4 Discussions and Conclusions

In this study, we compile a list of hate speech-related and NBA athletes-related keywords to collect tweets that potentially contain hateful content toward NBA athletes. We then fine-tune a BERT model to classify collected tweets into hate speech, offensive language, and neither on an open hate speech dataset [2]. After examining the classifier performance on a manually labeled subset of our collected tweets, we find that out of the 503,424 tweets, 3.33% ($n = 16{,}784$) are classified as hate speech, and 60.11% ($n = 302{,}605$) are offensive language. Our model achieves an overall accuracy of 79.33, and a weighted F1 score of 79.59. These results demonstrate the effectiveness of our classification approach in discerning hate speech and offensive language within the collected dataset.

To gain a deeper understanding of the specific groups that are more susceptible to hate speech, we use the keyword search method. Through this process, we uncover notable patterns indicating that athletes belonging to the `Black` community and the `LGBTQ` community are disproportionately targeted with hate speech. Additionally, players who possess a distinct play style, as well as those who are shorter or overweight, emerge as prominent targets for such abuse. These findings shed light on the specific demographics and characteristics of athletes who are most likely to face hate speech within the NBA community. Racism, physical shaming, play styles, and anti-LGBTQ remarks are the major themes found in our collected dataset.

In conclusion, this study provides valuable insights into the prevalence of hate speech directed toward NBA athletes on social media platforms. By employing a combination of keyword searches and machine learning techniques, we have

identified the targeted groups and major themes of hate speech within the NBA community.

Moving forward, further research can explore the impact of hate speech on the mental well-being of the targeted athletes and evaluate potential interventions to mitigate this issue. Additionally, analyzing the role of social media platforms and their policies in addressing hate speech toward athletes could contribute to fostering a safer online environment. It is essential to continue monitoring and addressing this ongoing problem to promote respect, inclusivity, and support for athletes across all platforms. By understanding hate speech in sports and taking proactive measures, we can work toward creating a positive and supportive environment for athletes to thrive both on and off the court.

Acknowledgments. This research was supported in part by the Goergen Institute for Data Science.

References

1. Chen, L., Lyu, H., Yang, T., Wang, Yu., Luo, J.: Fine-grained analysis of the use of neutral and controversial terms for COVID-19 on social media. In: Thomson, R., Hussain, M.N., Dancy, C., Pyke, A. (eds.) SBP-BRiMS 2021. LNCS, vol. 12720, pp. 57–67. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80387-2_6
2. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 512–515 (2017)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
4. ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., Belding, E.: Peer to peer hate: hate speech instigators and their targets. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
5. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
6. Lyu, H., et al.: Social media study of public opinions on potential COVID-19 vaccines: informing dissent, disparities, and dissemination. Intell. Med. **2**(01), 1–12 (2022)
7. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: challenges and solutions. PLoS ONE **14**(8), e0221152 (2019)
8. Magu, R., Joshi, K., Luo, J.: Detecting the hate code on social media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 608–611 (2017)
9. Magu, R., Luo, J.: Determining code words in euphemistic hate speech using word embedding networks. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 93–100 (2018)

10. Powell, A., Scott, A.J., Henry, N.: Digital harassment and abuse: experiences of sexuality and gender minority adults. Eur. J. Criminol. **17**(2), 199–223 (2020)
11. Reynolds, T.: NBA enacting zero-tolerance rules for abusive, hateful fan behavior (2019)
12. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
13. Zhang, Y., Lyu, H., Liu, Y., Zhang, X., Wang, Y., Luo, J.: Monitoring depression trends on twitter during the COVID-19 pandemic: observational study. JMIR Infodemiol. **1**(1), e26769 (2021)