Journal of High School Science

Computationally investigating racial bias in NBA commentary

Mukherjee A

Submitted: July 15, 2025, Revised: version 1, August 26, 2025, version 2, September 2, 2025
Accepted: September 5, 2025

Abstract

The language that sports commentators use has a major impact on the perception of athletes. Prior social science research shows that live sports commentary often contains racial bias. However, many of these studies were predicated on manual categorizations of statements and small sets of data. This study investigated racial bias in basketball commentary by using computational techniques to analyze a large sample of data. Transcripts of NBA games from 2023-2025 were processed to create a dataset of over 20,000 player mentions tagged with the player's skin tone (lighter-skinned or darker-skinned). The data was analyzed using two approaches: sentiment analysis and machine learning classification. The results of sentiment analysis showed that mentions describing darker-skinned players and lighter-skinned players were very similar in tone. Conversely, a logistic regression model was able to predict a player's skin tone from a mention with an accuracy of 60%, signifying that there were marginal differences in the ways commentators discussed players of different races. Inspection of the model's coefficients found that terms related to athleticism (e.g., athletic, speed) were associated with darker-skinned players while words related to character and intellect (e.g., smart, mature) were linked to lighter-skinned players. This disparity in speech promotes harmful stereotypes about race, physicality and cognitive ability in sports. However, humans – regardless of ethnicity - harbor ethnocentric biases which, in this study, were detected by ML models as being marginally greater (~ 60% accuracy) than chance alone would dictate. Therefore, there is a significant risk of incorrectly attributing these universal human traits with deliberate malice and resentment against other ethnic groups or races; especially in the absence of any sentiment bias.

Keywords

Racial bias, Natural Language Processing, Machine Learning, Basketball, Sports commentary, Sentiment analysis, Logistic regression, Stereotypes, Athleticism, Intellect

Arjun Mukherjee, Los Altos High School, 201 Almond Ave, Los Altos, CA 94022, USA.
arjunmukherjee2007@gmail.com

https://doi.org/10.64336/001c.144055

1. Introduction

Watching televised sports games has become a very popular form of entertainment. Each year, televised National Basketball Association (NBA) games attract millions of viewers (1). These broadcasts feature commentators, who narrate the action and provide analysis about players. Prior research has raised concerns that the language used in this commentary is racially biased. Specifically, studies observed that non-white athletes were more likely to receive praise for their athleticism while caucasian athletes were more likely to be credited for their character and intelligence (2-4). This pattern can be damaging because it insinuates that non-white players succeed due to "god-given" physical abilities while white players achieve success due to hard work and intellect (5).

Racial bias can exist both implicitly and explicitly. Explicit racial bias occurs when someone deliberately expresses prejudiced attitudes about people based on their race. On the other hand, implicit racial bias refers to when individuals subconsciously make associations between members of a certain race and one or more attributes (6). Live sports commentary offers a valuable context for examining implicit racial bias. Due to the fast-paced nature of sports games, commentators often have little time to think before they speak. This spontaneous setting can evoke statements out of commentators that represent their subconscious beliefs and biases (7).

Racial prejudice in commentary has traditionally been analyzed through subjective categorizations of statements and small datasets. Recently, some studies have examined racial bias in sports commentary through a computational lens. To study bias in NFL commentary, Iyyer et al. used YouTube Transcripts to produce a dataset of over 250,000 player mentions (8). They then used basic Natural Language Processing techniques to analyze elements of bias within the dataset and confirm some of the conclusions of previous social science work.

This project builds on those computational methods and applies them to NBA commentary. It is important to investigate racial bias in NBA commentary because the NBA is one of the most watched sports leagues in the world. With such a large audience, the way commentators portray players can reinforce stereotypes and affect public perception on a large scale. Additionally, the player pool for the NBA is mostly African-American (9) while the commentators are predominantly Caucasian (10). This dynamic creates an interesting context to study racial bias, as the speech of Caucasian commentators may subconsciously create unfair narratives about black players.

To examine racial bias in NBA commentary, this study compiled a large dataset of broadcasts from the past 2 seasons (2023-2025) and analyzed it using computational techniques. By leveraging Natural Language Processing tools, the analysis investigated differences in the ways commentators spoke about players of different skin tones and explored linguistic patterns that contributed to such differences.

2. Methods

2.1 Data collection

2.1.1 *Commentary extraction*

To obtain transcripts of NBA broadcasts, YouTube's automatic subtitle feature was used. Fifty one full NBA broadcasts from 2023-2025 were found on various unofficial YouTube channels. For each game, a transcript of the commentary was downloaded using YouTube's automatic transcription feature. It is important to note that YouTube's automatic transcription service has been found to be as low as 60% accurate, depending on audio quality (11). An inspection of the transcripts found that YouTube's subtitles were consistent for the general commentary and only unreliable for certain player names (i.e. "Jokic" was spelled as "Yokic"). To address this, a list of commonly misspelled player names alongside their correct spellings was maintained. Using this list, a find-and-replace procedure was applied to all transcripts to correct naming errors. Once these naming errors were resolved, an algorithm was constructed to extract a 30-word window surrounding each mention of a player's name in the transcripts. This window consisted of 15 words before the player's name and 15 words after it. All instances of the player's name in the mention were replaced with "<mentioned_player>". For windows that contained multiple player names, the algorithm was not able to determine who the statement primarily referred to. Separate mentions were created for each player with the other player's name being replaced with "teammate" or "opponent". For example, the statement "LeBron and Luka are very smart" would be separated into 2 mentions: "<mentioned_player>" and <teammate> are very smart" and "<teammate> and <mentioned_player> are very smart".

2.1.2 *Race identification*

Knowing what race each player is perceived as is crucial for analyzing bias in NBA commentary. Unfortunately, publicly available rosters or websites do not contain this data. Instead, the Fitzpatrick color scale was used to acquire information about a player's skin tone, following the approach of Curcic (2). A headshot image from NBA.com was downloaded for every player in the dataset. These images were used to measure the RGB value of each player's skin tone. Players were then assigned values on the Fitzpatrick scale, shown in Figure 1. This was performed by calculating the minimum Euclidean distance between the RGB value of a player's skin tone and the threshold RGB values associated with each category on the Fitzpatrick scale. Players who fell into categories 1-3 were classified as "lighter-skinned players" while those who fell into categories 4-6 were classified as "darker-skinned players".

Figure 1. The Fitzpatrick Skin Type Scale (12)

2.1.3 *Final dataset*

Each player mention was tagged with the player's name, skin tone, and team, and added to a structured dataset. The final dataset contained 28,756 mentions; 67% referred to darker-skinned players and 33% referred to lighter-skinned players. Figure 2 shows an example of a mention in the dataset.

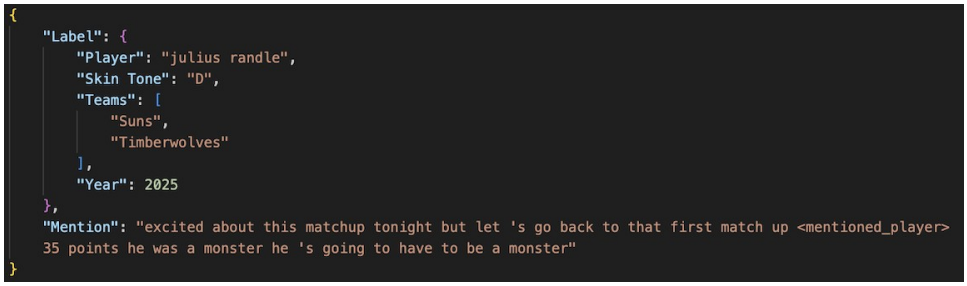

Figure 2. Example of a mention in the dataset

2.2 Text preprocessing and feature extraction

The textual information in each mention was converted into numerical data using Term Frequency - Inverse Document Frequency, implemented via the TfidfVectorizer from the scikit-learn library. This approach returns values that signify how important each word in the mention is which helps models focus on the most relevant parts of the text. These weights were generated as $TF-IDF(t,d,D)=TF(t,d)\times IDF(t,D)$ (eq1), where the TF-IDF weight for a term t in document d is the product of the term frequency (TF) of t in d and the inverse frequency of t in a collection of all the documents D. TF(t,d) is simply the number of times t appears in d divided by the total word count of d. IDF was calculated as $IDF(t,D)=\frac{\log(1+N)}{(1+DF(t))}+1$ (eq2), where N is the number of documents in D and DF is the number of documents in D containing t. The final result was a matrix of TF-IDF weights that corresponded to each word in each mention.

2.3 Sentiment analysis

The sentiment of mentions was analyzed to determine if commentators discussed players more positively or negatively based on their skin tone. VADER (Valence Aware Dictionary and sEntiment Reasoner) was used to conduct this analysis. This tool grades a piece of text's sentiment on a scale from -1(negative) to 1(positive) by assigning individual scores to each word in the text, combining them and then adjusting for contextual factors. Each mention in the dataset was graded using this tool. The distribution of scores for mentions of lighter-skinned players and darker-skinned players was examined to assess whether commentators spoke more positively about one group versus the other.

2.4 Model training

Machine learning models were trained to predict whether a mention referred to a darker-skinned or lighter-skinned player. Accuracy close to 100% would imply the existence of significant racial bias while that close to 50% would imply that no racial bias existed. To ensure that the dataset was balanced, some mentions of darker-skinned players were randomly removed, since they appeared more frequently. This process was performed using random sampling, where each darker-skinned mention had an equal chance of being selected. Since the sampling was fully random, there was no selection bias. A total of 10,198 mentions were removed, leaving a total of

18,558 mentions in the dataset. 80% of the data was used for training and 20% was used for testing. 3 models were trained and tested – logistic regression, random forest, and BERT based models. Each model's performance was evaluated using accuracy, precision, recall and F1 score.

2.4.1 *Logistic regression*

Logistic Regression is a linear classification model that calculates the probability of a mention referring to each skin tone and outputs the one with the higher probability. It does this by applying the sigmoid function $P(y=1|x)=\frac{1}{e^{-(w \cdot x+b)}}$ (eq3), where x is the TF-IDF representation of a mention, b is the intercept and P(y=1|x) is the probability that a mention refers to a darker-skinned player. The variable w represents a vector of learned weights that is repeatedly adjusted to minimize a log loss function. To implement the model, scikit-learn's logistic regression class was used. It was trained with a method called 'saga', which allowed it to work faster and more effectively on larger datasets. L2 regularization, a technique that prevents the model from relying too much on any one signal, was also used. Furthermore, the model's coefficients were analyzed to verify whether certain kinds of words were associated with darker-skinned or lighter-skinned players. A list of 100 commonly used descriptors related to work ethic, athleticism and intelligence was compiled based on themes found in prior research (3, 7, 8). The coefficient of each word in the list as returned by the model was extracted and the results were qualitatively analyzed.

2.4.2 Random Forest

Random Forest models combine the predictions of multiple decision trees, which each examine the data from a different perspective. The RandomForestClassifier was used to implement this model using the class from scikit-learn. Two hundred decision trees were used, and each one was allowed to grow up to 20 levels deep. To ensure that trees were independent, the max_features parameter was set to log2, which forced each tree to consider a random subset of features at each split. The features input to the model were the TF-IDF representations of each mention.

2.4.3 BERT-based model

In addition to conventional machine learning models, a deep learning approach was also explored using a BERT-based model. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based language model that is already pre trained on large amounts of text data. Instead of processing text one word at a time, this model processes the entire text at once, allowing it to capture complex relationships between words. The bert-base-uncased from the Hugging Face Transformers library was used, fine-tuned on the player mention dataset. Instead of feeding the model TF-IDF vectors, the raw textual information in the mentions was tokenized using BERT's tokenizer. The model was trained in batches of 8 and had a learning warmup rate of 500 steps. Regularization with weight decay (0.01) was applied to reduce overfitting and encourage evenly-distributed weights.

3. Results

To determine whether commentators talk about players in a more positive or negative way based on their perceived race, every mention in

the dataset was assigned a sentiment score using VADER. The histogram in figure 3 shows the distributions of these scores for both darker-skinned players (blue) and lighter-skinned players (orange). The histogram shows that distributions of sentiment scores are very similar between groups – the bars on the chart are nearly the same height at each interval. This indicates that the tone of language used to describe players does not vary based on skin tone. However, sentiment is only one component of language. Even if there are almost no discrepancies in tone, bias still may be present in the content of what is said.

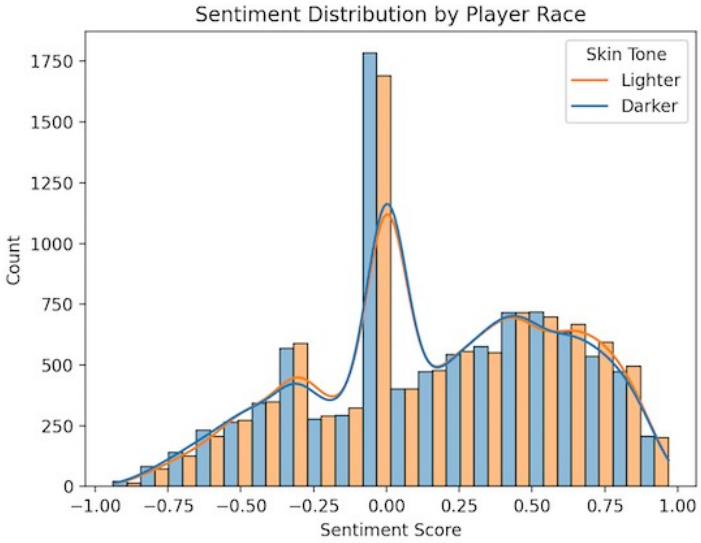

Figure 3. Distribution of sentiment scores for mentions of darker-skinned players and lighter-skinned players

The graph also provides insights about NBA commentary in general. There is a clear peak at 0 for both groups which makes sense since commentators often state factual statements about the game (i.e. "gets the rebound" or "takes the shot"). Additionally, the distributions tail off on both edges, suggesting that commentators tend to avoid using highly positive and negative statements.

Along with sentiment analysis, models were trained to verify whether a player's race could be predicted solely based on a snippet of the commentator describing them. Table 1 shows the accuracy, precision, recall and F1 scores of the three models described above.

Table 1. Test accuracy, precision, recall and F1 Score of machine learning models

Model	Test Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.600	0.600	0.605	0.602
Random Forest	0.557	0.553	0.597	0.574
BERT-based model	0.500	0.500	0.313	0.385

The BERT-based model, despite its ability to understand advanced contextual relationships, was only able to obtain an accuracy of 50%, which is no better than random guessing. Additionally, the model had a recall of 0.313, meaning that it was only able to correctly identify only 31.3% of all actual biased comments. Overall, the model performed the worst in all 4 metrics and struggled to make reliable predictions. One possible explanation for this is that commentary is often fragmented and very different from the well-structured materials BERT is pre-trained on, which could confuse its contextual understanding.

The Random Forest model performed slightly better with an accuracy of 55.7% and an F1 score of 0.574. The model was marginally more accurate than random guessing, suggesting that it was not able to detect differences in the ways commentators described players of different races. The model's confusion matrix, shown in figure 4, reveals these insights about its performance.

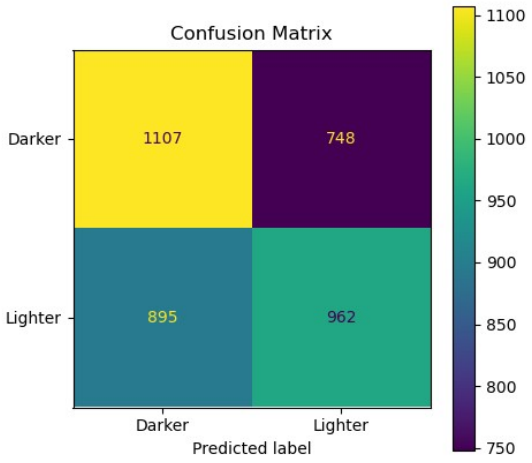

Figure 4. Confusion matrix of the random forest model

Of the evenly split test dataset of 3,712 mentions, the model guessed that 2,002 referred to darker-skinned players (Bottom Right + Top Right) and 1,710 referred to lighter-skinned players (Bottom Left + Top Left). This imbalance indicates that the model's decision making was slightly skewed towards predicting that mentions referred to darker-skinned players, showing that it had difficulty differentiating between commentary of lighter-skinned and darker-skinned players.

The logistic regression model performed the best in all 4 metrics. It was able to achieve an accuracy, precision, recall and F1 score of 0.6. This means that a player's race could be predicted from a short window of commentary at an accuracy of 60%, which is 20% greater than chance alone. The model's marginally

greater-than-chance accuracy indicates that there are identifiable differences in the ways commentators talk about players of different skin tones. Figure 5 shows the model's confusion matrix. Out of the 3,712 mentions in the test set, the model correctly classified 1,107 mentions of lighter-skinned players and 1,123 of darker-skinned players. Moreover, the number of lighter-skinned players the model classifies as darker-skinned (750) was about the same as the number of darker-skinned players the model classifies as lighter-skinned (732). These nearly symmetrical classification patterns suggest that the model is not able to identify discernible differences in the way the commentators talk about light-skinned versus dark-skinned players.

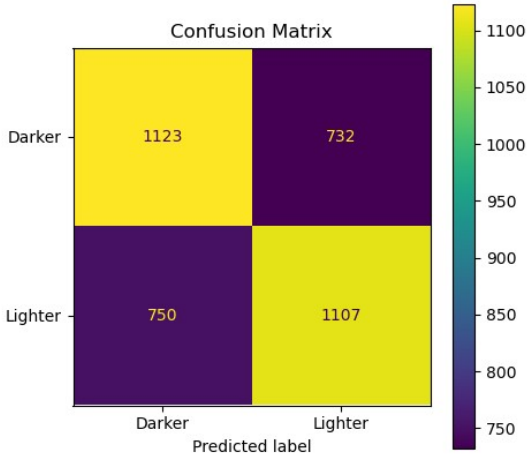

Figure 5. Confusion matrix of the logistic regression model

Since logistic regression is a highly interpretable model, its coefficients were inspected to understand what was driving its predictions. Tables 2 and 3 show the highest weighted positive and negative descriptors, respectively. In this context, coefficients represent how a word's presence changes the log-odds of a mention being predicted as referring to a darker-skinned player. Thus, positive coefficients indicate that the model associates the word with darker-skinned players while negative coefficients suggest that the model associates the word with lighter-skinned players. These coefficients aren't assigned randomly; the model learns them by observing patterns in the training data. Many descriptors related to athleticism had high positive coefficients, such as "athletic" (2.77), "fast" (1.89) and "explosiveness" (0.7). On the other hand, many of the words with the largest negative coefficients were related to cognitive ability and work ethic, like "smart" (-1.55), "scrappy" (-0.626) and "disciplined" (-0.515).

Table 2. Most positively weighted descriptors from logistic regression model

Word	Coefficient
athletic	2.7697
fast	1.8931
machine	1.7722
trouble	1.3808
long	1.1974
spectacular	1.0165
wild	0.8880
bully	0.7837
unstoppable	0.7426
explosiveness	0.6997

Table 3. Most negatively weighted descriptors from logistic regression model

Word	Coefficient
hustling	-1.8102
smart	-1.5516
unfortunate	-0.9710
steady	-0.8364
intense	-0.7352
aware	-0.6332
scrappy	-0.6259
powerful	-0.6249
mature	-0.5564
disciplined	-0.5153

To verify whether these words are associated with skin tone, the proportion of mentions that contained each descriptor was calculated for both groups and used to conduct a two-tailed z-test. This test determines if the difference in the frequency of a word appearing in mentions of one skin tone compared to the other is statistically significant. The results, shown in table 4, reveal that several words appear disproportionately in mentions of a certain skin tone. Many of the descriptors that were associated with darker-skinned players emphasize physicality, such as “bully” ($z = 2.83$, $p = 0.005$) and “athletic” ($z = 2.14$, $p = 0.032$) On the other hand, words describing cognitive ability or effort like “hustling” ($z = -3.20$, $p = 0.001$) and “intelligence” ($z = -2.06$, $p = 0.0393$) were linked to lighter-skinned players. No words related to physical ability are associated with lighter-skinned players and no words related to intellect or character are associated with darker-skinned players. These findings suggest that commentators overly focus on the physical attributes of darker-skinned players and the intelligence or character of lighter-skinned players, which aligns with the findings of previous social science work (3, 7, 8).

Table 4. Words significantly associated with player skin tone based on proportions Z-test

Word	Z-Statistic	P-Value	Association
hustling	-3.20	0.0014	Lighter-skinned
machine	3.07	0.0021	Darker-skinned
bully	2.83	0.0046	Darker-skinned
intense	-2.52	0.0116	Lighter-skinned
beauty	-2.38	0.0171	Lighter-skinned
trouble	2.32	0.0206	Darker-skinned
entertaining	2.21	0.0271	Darker-skinned
athletic	2.14	0.0323	Darker-skinned
unfortunate	-2.14	0.0322	Lighter-skinned
unstoppable	2.11	0.0352	Darker-skinned
scrappy	-2.06	0.0393	Lighter-skinned
intelligence	-2.06	0.0393	Lighter-skinned
spectacular	2.01	0.0441	Darker-skinned
unbelievable	1.98	0.0481	Darker-skinned

While physicality and cognitive ability are both very useful characteristics in sports, the emphasis of these traits for different racial groups can be harmful. By emphasizing the athletic abilities of darker-skinned players and the intellect and work ethic of lighter-skinned players, commentators reinforce the racial stereotype that darker-skinned players derive their success from natural physical gifts while lighter-skinned players obtain success through hard work and intelligence. This narrative makes it seem like lighter-skinned players have earned their success more. It also downplays the cognitive skills of darker-skinned players, which may make it more difficult for them to obtain sports related employment after they retire.

4 Conclusion

The voices of NBA commentators reach millions of people each year, which means that potential racial bias in their speech can have severe consequences. This study used machine learning techniques to investigate whether NBA commentators discuss players differently based on race. Sentiment analysis showed that there were no significant differences in the tone of the language that commentators used to discuss players of different races. However, a logistic regression model was able to predict a player's race from a short snippet of commentary with 60% accuracy, marginally greater than that which would be expected to occur due to pure chance; which could be interpreted as indicative of slight differences in which commentators described light-skinned versus black-skinned players. The coefficients of this model found that words associated with athleticism were more likely to be linked to darker-skinned players, while terms related to work ethic and intellect were associated with lighter-skinned players. This disparity in language use suggests that commentators may unconsciously reinforce stereotypes based on race. However, the fact that accuracy, precision and recall were ~ 0.6 for any model, combined with the fact that sentiment analysis found no racial bias, cautions against blithely attributing

what could arguably be normal ethnocentric bias (normally present in individuals belonging to any ethnicity or race), to concerted and deliberate contemptuous malignity toward other ethnic groups.

Github Repository
https://github.com/ArjunMukherjee1/NBA-Racial-Bias/

5 References

1. Adgate B. NBA Ratings Declined By 2% During The 2024-25 Regular Season. Forbes, 2025. https://www.forbes.com/sites/bradadgate/2025/04/17/for-the-2024-25-nba-regular-season-ratings-declined-by2/

2. Curcic D. Racial Bias in NBA Commentary (Analysis). RunRepeat, 2024. https://runrepeat.com/racial-bias-in-nba-commentary

3. Rada JA. Color blind-sided: Racial bias in network television's coverage of professional football games. Howard Journal of Communications, 7, 231–239, 1996. https://doi.org/10.1080/10646179609361727

4. Foy SL, Ray R. Skin in the Game: Colorism and the Subtle Operation of Stereotypes in Men's College Basketball. American Journal of Sociology, 125, 730-785, 2019. https://doi.org/10.1086/707243

5. Fitzgerald C, Martin A, Berner D, Hurst S. Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. BMC Psychology, 7, 29, 2019. https://doi.org/10.1186/s40359-019-0299-7

6. Rada JA, Wulfemeyer KT. Color Coded: Racial Descriptors in Television Coverage of Intercollegiate Sports. Journal of Broadcasting & Electronic Media, 49, 65–85, 2005. https://doi.org/10.1207/s15506878jobem4901_5

7. Merullo J, Yeh L, Handler A, Grissom A, O'Connor B, Iyyer M. Investigating sports commentator bias within a large corpus of American football broadcasts. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, pp 6355-6361, 2019. https://doi.org/10.18653/v1/D19-1666

8. Eastman T, Billings AC. Biased voices of sports: Racial and gender stereotyping in college basketball announcing. Howard J Commun, 12,183–201, 2001. https://doi.org/10.1080/106461701753287714

9. Gaetano A, Shivdasani K, Chen A, Garbis N, Salazar D. Racial Concordance Between NBA and MLB Players and Their Team Physicians. Orthopaedic Journal of Sports Medicine, 13, 2025. https://doi.org/10.1177/23259671241310472

10. Halberstam JB. Fighting the battle of minority broadcasters; Some 80% play in league; Only 10% do play-by-play. Sports Broadcast Journal, 2025. https://www.sportsbroadcastjournal.com/fighting-the-battle-of-minority-broadcasters-some-80-play-in-league-only-10-do-play-by-play/

11. Correcting YouTube Auto-Captions. University of Minnesota Duluth. https://itss.d.umn.edu/service-catalog/mediahub/request-media-accessibility-services/about-accessibility/about-0

12. Understanding the Fitzpatrick Scale in PMU. Nuva Colors, 2023. https://nuvacolors.com/blogs/news/the-pmu-industry-in-a-nutshell