Investigating Sports Commentator Bias within a Large Corpus of American Football Broadcasts

Jack Merullo[★♠] **Luke Yeh**[★♠] **Abram Handler**[♠]
Alvin Grissom II[♣] **Brendan O'Connor**[♠] **Mohit Iyyer**[♠]

University of Massachusetts Amherst[♠] Ursinus College[♣]
{jmerullo,lyeh,ahandler,miyyer,brenocon}@umass.edu
agrissom@ursinus.edu

Abstract

Sports broadcasters inject drama into play-by-play commentary by building team and player narratives through subjective analyses and anecdotes. Prior studies based on small datasets and manual coding show that such theatrics evince commentator bias in sports broadcasts. To examine this phenomenon, we assemble FOOTBALL, which contains 1,455 broadcast transcripts from American football games across six decades that are automatically annotated with 250K player mentions and linked with racial metadata. We identify major confounding factors for researchers examining racial bias in FOOTBALL, and perform a computational analysis that supports conclusions from prior social science studies.

1 Introduction

Sports broadcasts are major events in contemporary popular culture: televised American football (henceforth "football") games regularly draw tens of millions of viewers (Palotta, 2019). Such broadcasts feature live sports commentators who weave the game's mechanical details into a broader, more subjective narrative. Previous work suggests that this form of storytelling exhibits racial bias: nonwhite players are less frequently praised for good plays (Rainville and McCormick, 1977), while white players are more often credited with "intelligence" (Bruce, 2004; Billings, 2004). However, such prior scholarship forms conclusions from small datasets[1] and subjective manual coding of race-specific language.

We revisit this prior work using large-scale computational analysis. From YouTube, we collect broadcast football transcripts and identify mentions of players, which we link to metadata about each player's race and position. Our resulting FOOTBALL dataset contains over 1,400 games spanning six decades, automatically annotated with ∼250K player mentions (Table 1). Analysis of FOOTBALL identifies two confounding factors for research on racial bias: (1) the racial composition of many positions is very skewed (e.g., only ∼5% of running backs are white), and (2) many mentions of players describe only their actions on the field (not player attributes). We experiment with an additive log-linear model for teasing apart these confounds. We also confirm prior social science studies on racial bias in naming patterns and sentiment. Finally, we publicly release FOOTBALL,[2] the first large-scale sports commentary corpus annotated with player race, to spur further research into characterizing racial bias in mass media.

Player	Race	Mention text
Baker Mayfield	white	"Mayfield the ultimate competitor he's tough he's scrappy"
Jesse James	white	"this is a guy ...does nothing but work brings his lunch pail"
Manny Lawson	nonwhite	"good specs for that defensive end freakish athletic ability"
B.J. Daniels	nonwhite	"that otherworldly athleticism he has saw it with Michael Vick"

Table 1: Example mentions from FOOTBALL that highlight racial bias in commentator sentiment patterns.

2 Collecting the FOOTBALL dataset

We collect transcripts of 1,455 full game broadcasts from the U.S. NFL and National Collegiate Athletic Association (NCAA) recorded between 1960 and 2019. Next, we identify and link mentions of players within these transcripts to infor-

★Authors contributed equally.

[1] Rainville and McCormick (1977), for example, study only 16 games.

[2] http://github.com/jmerullo/football

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6355–6361, Hong Kong, China, November 3–7, 2019.

mation about their race (white or nonwhite) and position (e.g., quarterback). In total, FOOTBALL contains 267,778 mentions of 4,668 unique players, 65.7% of whom are nonwhite.[3] We now describe each stage of our data collection process.

2.1 Processing broadcast transcripts

We collect broadcast transcripts by downloading YouTube videos posted by nonprofessional, individual users identified by querying YouTube for football archival channels.[4] YouTube automatically captions many videos, allowing us to scrape caption transcripts from 601 NFL games and 854 NCAA games. We next identify the teams playing and game's year by searching for exact string matches in the video title and manually labeling any videos with underspecified titles.

After downloading videos, we tokenize transcripts using spaCy.[5] As part-of-speech tags predicted by spaCy are unreliable on our transcript text, we tag FOOTBALL using the ARK TweetNLP POS tagger (Owoputi et al., 2013), which is more robust to noisy and fragmented text, including TV subtitles (Jørgensen et al., 2016). Additionally, we use `phrasemachine` (Handler et al., 2016) to identify all corpus noun phrases. Finally, we identify player mentions in the transcript text using exact string matches of first, last, and full names to roster information from online archives; these rosters also contain the player's position.[6] Although we initially had concerns about the reliability of transcriptions of player names, we noticed minimal errors on more common names. Qualitatively, we noticed that even uncommon names were often correctly transcribed and capitalized. We leave a more systematic study for future work.

2.2 Identifying player race

Racial identity in the United States is a creation of complex, fluid social and historical processes (Omi and Winant, 2014), rather than a reflection of innate differences between fixed groups. Nevertheless, popular *perceptions* of race in the United States and the prior scholarship on racial bias in sports broadcasts which informs our work (Rainville and McCormick, 1977; Rada, 1996; Billings, 2004; Rada and Wulfemeyer, 2005) typically assume hard distinctions between racial groups, which measurably affect commentary. In this work, we do not reify these racial categories; we use them as commonly understood within the context of the society in which they arise.

To conduct a large-scale re-examination of this prior work, we must identify whether each player in FOOTBALL is perceived as white or nonwhite.[7] Unfortunately, publicly available rosters or player pages do not contain this information, so we resort to crowdsourcing. We present crowd workers on the Figure Eight platform with 2,720 images of professional player headshots from the Associated Press paired with player names. We ask them to "read the player's name and examine their photo" to judge whether the player is white or nonwhite. We collect five judgements per player from crowd workers in the US, whose high inter-annotator agreement (all five workers agree on the race for 93% of players) suggests that their perceptions are very consistent. Because headshots were only available for a subset of players, the authors labeled the race of an additional 1,948 players by performing a Google Image search for the player's name[8] and manually examining the resulting images. Players whose race could not be determined from the search results were excluded from the dataset.

3 Analyzing FOOTBALL

We now demonstrate confounds in the data and revisit several established results from racial bias studies in sports broadcasting. For all experiments, we seek to analyze the statistics of contextual terms that describe or have an important association with a mentioned player. Thus, we preprocess the transcripts by collecting contextual terms in windows of five tokens around each player mention, following the approach of Ananya et al. (2019) for gendered mention analysis.[9]

We emphasize that different term extraction strategies are possible, corresponding to different

[3] See Appendix for more detailed statistics.

[4] We specifically query for `full NFL|NCAA|college football games 1960s|1970s|1980s|1990s|2000`, and the full list of channels is listed in in the Appendix.

[5] https://spacy.io/ (2.1.3), Honnibal and Montani (2017)

[6] Roster sources listed in Appendix. We tag first and last name mentions only if they can be disambiguated to a single player in the rosters from opposing teams.

[7] While we use the general term "nonwhite" in this paper, the majority of nonwhite football players are black: in 2013, 67.3% of the NFL was black and most of the remaining players (31%) were white (Lapchick, 2014).

[8] We appended "NFL" to every query to improve precision of results.

[9] If multiple player mentions fall within the same window, we exclude each term to avoid ambiguity.

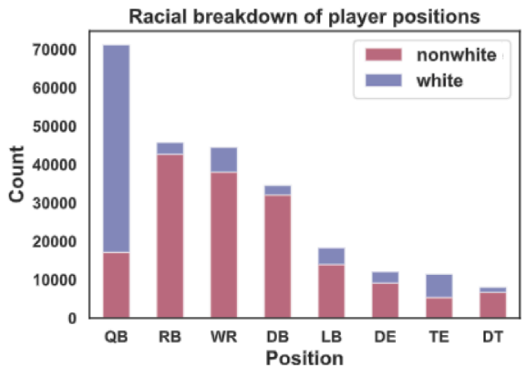

Figure 1: Almost all of the eight most frequently-mentioned positions in FOOTBALL are heavily skewed in favor of one race.

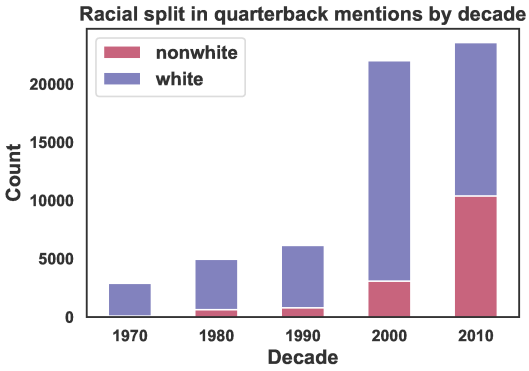

Figure 2: The percentage of nonwhite quarterbacks mentions has drastically increased over time, exemplifying the changing racial landscape in FOOTBALL across time.

precision–recall tradeoffs. For instance, instead of collecting all terms in a window (high recall) we might instead only collect terms in copular constructions with the entity mention (high precision), such as 'devoted' in "Tebow is devoted". Because mention detection strategies affect conclusions about bias in FOOTBALL, systematically defining, analyzing or even learning different possible strategies offers an exciting avenue for future work.

3.1 Statistical and linguistic confounds

Identifying racial bias in football broadcasts presents both statistical and linguistic modeling challenges. Many descriptions of players in broadcasts describe temporary player states (e.g., "Smith deep in the backfield") or discrete player actions ("Ogden with a huge block"), rather than possibly-biased descriptions of players themselves ("Cooper is one scrappy receiver"). Moreover, many players' actions ("passes the ball downfield") depend on the position they play, which is often skewed by race (Figure 1). Furthermore, the racial composition of mentions across different decades can differ dramatically—Figure 2 shows these changes for quarterback mentions—which makes the problem even more complex. Modeling biased descriptions of players thus requires disentangling attributes describing shifting, position-dependent player actions on field (e.g., "Paulsen the tight end with a *fast* catch") from attributes referring to intrinsic characteristics of individual players ("Paulsen is just so, so *fast*").

To demonstrate this challenge, we distinguish between per-position effects and racial effects using an additive, log-linear model which represents the log probability that a word or noun phrase w will describe a player entity e as the sum of two learned coefficients, corresponding to two observed covariates. One observed covariate records a player's race and the other a player's position, which allows us to use learned coefficients to represent how much a player's race or position contributes to the chance of observing an (w, e) pair.

Formally, we model such effects using a sparse MAP estimation variant of SAGE (Eisenstein et al., 2011). We define the binary vector $y_e \in \{0, 1\}^J$ to represent the observed player covariates of race (white or nonwhite) and position. For example, component $y_{e,k}$ will be set to 1 if player e is a quarterback and the component k indexes the quarterback covariate; y_e is a concatenation of two one-hot vectors. We then model $p(w \mid e) \propto \exp\left(\beta_w + (\gamma y_e)_w\right)$, with $\beta_w \in \mathbb{R}^{|\mathcal{V}|}$ as a background distribution over the vocabulary $\mathcal{V}$, set to empirical corpus-wide word and phrase log-probabilities, and $\gamma \in \mathbb{R}^{J \times |\mathcal{V}|}$ as a matrix of feature effects on those log probabilities. $\gamma_{j,w}$ denotes the *difference* in log-probability of w for the j^{th} player feature being on versus off. For example, if j indexes the quarterback covariate and w indexes the word "tough", then $\gamma_{j,w}$ represents how much more likely the word "tough" is to be applied to quarterbacks over the base distribution. We impose a uniform Laplace prior on all elements of γ to induce sparsity, and learn a MAP estimate with the LibLBFGS implementation of OWL-QN, an L1-capable quasi-Newtonian convex optimizer (Andrew and Gao, 2007; Okazaki, 2010).

Table 2 shows several highest-valued $\gamma_{j,w}$ for a subset of the J covariates. The adjective "quick"

White	*long way, long time, valuable*
DB	*strong safety, free safety, state university*
RB	*second effort,single setback,ground game*
QB	*freshman quarterback, arm strength, easier*
WR	*quick slant, end zone touchdown, punt returner*

Table 2: Top terms for the white, defensive back (DB), running back (RB), quarterback (QB), and wide receiver (WR) covariates for the log linear model.

is predictive of wide receivers, but refers to an action the players take on the field ("quick slant"), not an attribute of the receivers themselves. We also find that since "strong safety" is a kind of defensive back, the adjective "strong" is often associated with defensive backs, who are often nonwhite. In this case, "strong" does not reflect racial bias. Preliminary experiments with per-position mention-level race classifiers, as per Ananya et al. (2019), were also unable to disentangle race and position.

These results suggest that a more sophisticated approach may be necessary to isolate race effects from the confounds; it also raises sharp conceptual questions about the meaning of race-conditional statistical effects in social scientific inquiry, since race is a multifaceted construct (a "bundle of sticks," as Sen and Wasow (2016) argue). For future work, it may be useful to think of comparisons between otherwise similar players: how do broadcasters differ in their discussions of two players who are both quarterbacks, and who have similar in-game performance, but differ by race?

We now describe two experiments that sidestep some of these confounds, each motivated by prior work in social science: the first examines player naming patterns, which are less tied to action on field than player attributes. The other uses words with known sentiment polarity to identify positive and negative attributes, regardless of player position or game mechanics.

3.2 Exploring naming patterns

Naming patterns in sports broadcasting—how commentators refer to players by name (e.g., first or last name)—are influenced by player attributes, as shown by prior small-scale studies. For example, Koivula (1999) find that women are more frequently referred to by their first names than men in a variety of sports. Bruce (2004) discover a similar trend for race in basketball games: white players are more frequently referred to by their last names than nonwhite players, often because

Position	Race	First	Last	Full
QB	white	8.3%	20.0%	71.7%
QB	nonwhite	18.1%	7.5%	74.5%
WR	white	6.9%	36.5%	56.5%
WR	nonwhite	11.3%	24.1%	64.6%
RB	white	10.5%	41.2%	48.4%
RB	nonwhite	8.5%	35.4%	56.1%
TE	white	16.6%	18.7%	64.7%
TE	nonwhite	13.8%	16.6%	69.7%

Table 3: White players at the four major offensive positions are referred to by last name more often than nonwhite players at the same positions, a discrepancy that may reflect unconscious racial boundary-marking.

commentators believe their first names sound too "normal". Bruce (2004) further points out that the "practice of having fun or playing with the names of people from non-dominant racial groups" contributes to racial "othering". A per-position analysis of player mentions in FOOTBALL corroborates these findings for all offensive positions (Table 3).

3.3 Sentiment patterns

Prior studies examine the relationship between commentator sentiment and player race: Rainville and McCormick (1977) conclude that white players receive more positive coverage than black players, and Rada (1996) shows that nonwhite players are praised more for physical attributes and less for cognitive attributes than white ones.

To examine sentiment patterns within FOOTBALL, we assign a binary sentiment label to contextualized terms (i.e., a window of words around a player mention) by searching for words that match those in domain-specific sentiment lexicons from Hamilton et al. (2016).[10] This method identifies 49,787 windows containing sentiment-laden words, only 12.8% of which are of negative polarity, similar to the 8.3% figure reported by Rada (1996).[11] We compute a list of the most positive words for each race ranked by ratio of relative frequencies (Monroe et al., 2008).[12] A qualitative inspection of these lists

[10] We use a filtered intersection of lexicons from the NFL, CFB, and sports subreddits, yielding 121 positive and 125 negative words.

[11] Preliminary experiments with a state-of-the-art sentiment model trained on the Stanford Sentiment Treebank (Peters et al., 2018) produced qualitatively unreliable predictions due to the noise in FOOTBALL.

[12] We follow Monroe et al. (2008) in removing infrequent words before ranking; specifically, a word must occur at least ten times for each race to be considered.

Race	Most positive words
white (all)	*enjoying, favorite, calm, appreciate, loving, miracle, spectacular, perfect, cool, smart*
nonwhite (all)	*speed, gift, versatile, gifted, playmaker, natural, monster, wow, beast, athletic*
white (QBs)	*cool, smart, favorite, safe, spectacular, excellent, class, fantastic, good, interesting*
nonwhite (QBs)	*ability, athletic, brilliant, awareness, quiet, highest, speed, wow, excited, wonderful*

Table 4: Positive comments for nonwhite players (top two rows: all player mentions; bottom two rows: only quarterback mentions) focus on their athleticism, while white players are praised for personality and intelligence.

(Table 4) confirms that nonwhite players are much more frequently praised for physical ability than white players, who are praised for personality and intelligence (see Table 1 for more examples).

Limitations: The small lexicon results in the detection of relatively few sentiment-laden windows; furthermore, some of those are false positives (e.g., "beast mode" is the nickname of former NFL running back Marshawn Lynch). The former issue precludes a per-position analysis for all non-QB positions, as we are unable to detect enough sentiment terms to draw meaningful conclusions. The top two rows of Table 4, which were derived from all mentions regardless of position, are thus tainted by the positional confound discussed in Section 3.1. The bottom two rows of Table 4 are derived from the same analysis applied to just quarterback windows; qualitatively, the results appear similar to those in the top two rows. That said, we hope that future work on contextualized term extraction and sentiment detection in noisy domains can shed more light on the relationship between race and commentator sentiment patterns.

4 Related Work

Our work revisits specific findings from social science (§3) on racial bias in sports broadcasts. Such non-computational studies typically examine a small number of games drawn from a single season and rely on manual coding to identify differences in announcer speech (Rainville and McCormick, 1977; Billings, 2004; Rada and Wulfemeyer, 2005). For example, Rada (1996) perform a fine-grained analysis of five games from the 1992 season, coding for aspects such as players' cognitive or physical attributes. Our computational approach allows us to revisit this type of work (§3) using FOOTBALL, without relying on subjective human coding.

Within NLP, researchers have studied gender bias in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017), racial bias in police stops (Voigt et al., 2017) and on Twitter (Hasanuzzaman et al., 2017), and biases in NLP tools like sentiment analysis systems (Kiritchenko and Mohammad, 2018). Especially related to our work is that of Ananya et al. (2019), who analyze mention-level gender bias, and Fu et al. (2019), who examine gender bias in tennis broadcasts. Other datasets in the sports domain include the event-annotated baseball commentaries of Keshet et al. (2011) and the WNBA and NBA basketball commentaries of Aull and Brown (2013), but we emphasize that FOOTBALL is the first large-scale sports commentary corpus annotated for race.

5 Conclusion

We collect and release FOOTBALL to support large-scale, longitudinal analysis of racial bias in sports commentary, a major category of mass media. Our analysis confirms the results of prior smaller-scale social science studies on commentator sentiment and naming patterns. However, we find that baseline NLP methods for quantifying mention-level genderedness (Ananya et al., 2019) and modeling covariate effects (Eisenstein et al., 2011) cannot overcome the statistical and linguistic confounds in this dataset. We hope that presenting such a technically-challenging resource, along with an analysis showing the limitations of current bias-detection techniques, will contribute to the emerging literature on bias in language. Important future directions include examining the temporal aspect of bias as well as developing more precise mention identification techniques.

Acknowledgments

We thank the anonymous reviewers for their insightful comments, and Emma Strubell, Patrick Verga, David Fisher, Katie Keith, Su Lin Blodgett, and other members of the UMass NLP group for help with earlier drafts of the paper. This work was partially supported by NSF IIS-1814955.

References

Ananya Ananya, Nitya Parthasarthi, and Sameer Singh. 2019. Genderquant: Quantifying mention-level genderedness. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

G. Andrew and J. Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the International Conference of Machine Learning*.

Laura L Aull and David West Brown. 2013. Fighting words: a corpus analysis of gender representations in sports reportage. *Corpora*, 8(1).

Andrew C Billings. 2004. Depicting the quarterback in black and white: A content analysis of college and professional football broadcast commentary. *Howard Journal of Communications*, 15(4).

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of Advances in Neural Information Processing Systems*.

Toni Bruce. 2004. Marking the boundaries of the 'normal'in televised sports: The play-by-play of race. *Media, Culture & Society*, 26(6).

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334).

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the International Conference of Machine Learning*.

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2019. Tie-breaker: Using language models to quantify gender bias in sports journalism. In *International Joint Conference on Artificial Intelligence*.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of Empirical Methods in Natural Language Processing*.

Abram Handler, Matthew Denny, Hanna M. Wallach, and Brendan T. O'Connor. 2016. Bag of what? simple noun phrase extraction for text analysis. In *NLP+CSS@EMNLP*.

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. 2017. Demographic word embeddings for racism detection on twitter. In *International Joint Conference on Natural Language Processing*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Ezra Keshet, Terrence Szymanski, and Stephen Tyndall. 2011. BALLGAME: A corpus for computational semantics. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*.

Nathalie Koivula. 1999. Gender stereotyping in televised media sport coverage. *Sex roles*, 41(7-8).

Richard Lapchick. 2014. The 2014 racial and gender report card: National football league.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4).

Naoaki Okazaki. LibLBFGS: a Library of Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [online]. 2010.

Michael Omi and Howard Winant. 2014. *Racial formation in the United States*. Routledge.

Olutobi Owoputi, Brendan T. O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Frank Palotta. 2019. NFL ratings rebound after two seasons of declining viewership. *CNN Business*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

James A Rada. 1996. Color blind-sided: Racial bias in network television's coverage of professional football games. *Howard Journal of Communications*, 7(3).

James A Rada and K Tim Wulfemeyer. 2005. Color coded: Racial descriptors in television coverage of intercollegiate sports. *Journal of Broadcasting & Electronic Media*, 49(1).

Raymond E Rainville and Edward McCormick. 1977. Extent of covert racial prejudice in pro football announcers' speech. *Journalism Quarterly*, 54(1).

Maya Sen and Omar Wasow. 2016. Race as a 'bundle of sticks': Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522.

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25).