Sidelined: Using Natural Language Processing to Investigate Gender Bias in Basketball Sports Journalism

Nathaniel Yellin

Abstract

In this journal article, I investigate gender bias in sports journalism, focusing on ESPN.com coverage of National Collegiate Athletic Association (NCAA) basketball. With the programming language R and using natural language processing (NLP) techniques, I analyze more than 1,700 articles on ESPN.com to create an R Shiny application Sidelined. This app reveals that gender bias is prevalent in online NCAA basketball coverage at the league, team, and individual player level. At the league level, I find a significant quantitative difference in female coverage, in the number of articles, game recaps and correct assignment of articles on ESPN.com for the women's teams and players as compared to male counterparts. Being mentioned in one article is considered an outlier for female athletes. I also uncover that male players who play and score more receive more media attention, but this relationship is less clear for female players. Performing sentiment analysis on the adjectives used to describe NCAA athletes, I reveal that there is a notable qualitative difference in how female NCAA athletes are covered, especially those who play the center position. *Sidelined* is innovative and can be explored to investigate, with data visualizations and word clouds, the extent to which gender bias exists on ESPN.com at the specific team or player level. The findings of this research are significant, highlighting the need for greater attention to gender bias in online sports journalism and the potential for NLP to play a key role in identifying and addressing such biases.

Sidelined: Using Natural Language Processing to Investigate Gender Bias in Basketball Sports Journalism

Introduction

Sports journalism is an integral part of the sports industry, providing coverage and analysis of athletic competitions, teams and players. Despite the strides made towards gender equality in sports, media coverage of sports has had a long history of privileging male athletes (Messner et al., 1988; Weber & Carini, 2012). Much of the research on gender bias in sports reporting has focused on coverage in newspapers (Dunne, 2017; Hartmann-Tews, 2019), sports magazines (Bishop, 2003; Lumpkin, 2009), and broadcast media or "air time" (Cooky et al., 2015; Eastman & Billings, 2000; Higgs et al., 2003). Some studies have looked at stereotypical descriptions and framing (Angelini & Billings, 2010; Jones, 2004; Kian et al., 2009; Messner et al., 1993) or the type of questions female athletes received when interviewed (Fu et al., 2016). This favoritism towards male athletes has shown to be embedded even when major women's sporting events peaked in newsworthiness (Eastman & Billings, 2000) and even on platforms specifically designed to promote women in sports (Ancheta et al., 2019). The authors of several longitudinal studies across a variety of media platforms found that media coverage of women's sports has declined over the years despite the increased participation and athletic performance of female athletes (Cooky et al., 2015; Kane, 2013; Weber & Carini, 2012). These and other studies have challenged the assumption that the media simply provides fans with what they "want to see," finding instead that the qualitative and quantitative differences in media coverage of male and female athletes actually strengthens audiences for men's sport while marginalizing, or keeping women and the coverage of their sport play *sidelined* (Cooky et al., 2013; Fink, 2015).

Gender bias has been shown to be pronounced on ESPN, with greater marginalization in electronic media (Eastman & Billings, 2000). In one study, women's sports constituted only 5% of all televised coverage on ESPN, despite the greater use of female sports reporters and analysts on ESPN than on CNN (Tuggle, 1997). Another study found that televised coverage of women's sports in 2009 amounted to 1.6% of the time across three major networks and 1.4% for ESPN's *SportsCenter* (Messner & Cooky, 2010). Disparities in mean coverage time on ESPN's *SportsCenter* have also been shown to be dramatic (Martin et al., 2016). Male athletes have been shown to be disproportionately favored, even on espnW, a platform specifically established to expand the coverage of women's sports (Ancheta et al., 2019). While some may defer to market forces as the reason for the disparate coverage of male and female sports, this explanation is unjustified and "fails to acknowledge that sport consumption is a mediated process: what is covered, how often it is covered, and the manner in which it is covered all impact audience perceptions of value and quality" (Fink, 2015, p. 336). Media coverage of athletes is a dominant force that shapes and perpetuates public perceptions, norms and stereotypes about gender and serious attention should be given to address the coverage of females in sports, both quantitatively and qualitatively.

Within the realm of collegiate sports, NCAA basketball is one of the most popular and widely followed sports in the United States. Despite the popularity of women's basketball, especially during *March Madness*, female athletes have been shown to be undervalued financially and in media terms (Zimbalist, 2019). To date, no other research has used NLP to assess both qualitatively and quantitatively the extent to which bias exists in online sports journalism, and specifically for female athletes in the NCAA. In 2016, researchers at Cornell used NLP to determine that in post tennis game press conferences, female players were asked

more questions not related to the game (Fu et al., 2016). The extensive mediatization and the powerful effect of mass media in framing modern culture and arguably assigning value to female athletes, financially and otherwise, demonstrate the relevance of research in this area. This study investigates the extent to which gender bias exists both quantitatively and qualitatively in the media coverage of female athletes in the NCAA on ESPN.com, a main source of sports news and information. I analyzed more than 1,700 articles on ESPN.com featured prior to December 29, 2022 to create an R Shiny Application, Sidelined, which demonstrates that gender bias is prominent in the quantity and quality of articles posted. On Sidelined, sentiment analysis is used to measure the overall tone of articles featured and keyword extraction methods are implemented to gather adjectives used to describe players and teams. The user of the application can interact with word clouds and data visualizations to identify whether a bias exists at the league, team or player level and evaluate whether the adjectives used in their description accentuate this bias. To narrow the gender divide in the media distribution and promotion of college sports, further research using NLP should be undertaken and serious attention given to address the limited coverage of NCAA female athletes on ESPN and other media outlets.

Methods

All sample code is available on GitHub (https://github.com/nateyellin). In this study, I investigate how female and male NCAA athletes were covered on ESPN.com in articles featured on or before December 29th, 2022. Using R and the {tidyverse*}*, {rvest*}*, {stringi*}*, and {tidytext} libraries, I wrote functions and loops to scrape four distinct pieces of data from ESPN.com: (1) team URLs, (2) player statistics by team, (3) article URLs, and (4) the texts of these articles. Team URLs refer to the ESPN web addresses for each collegiate team and article URLs refer to the web addresses for each article written. After gathering all 363 men's team

URLs and all 361 women's team URLs, I wrote functions to scrape and extract all player statistics that were nested in the *Statistics* header under each team page and an iterative loop gathered this data for all 724 teams. On ESPN, articles for each team appear under the *Home* header. Accordingly, I scraped the most recent 10 article links (or up to 10 if there were less on a team's *Home*) for each team and subsequently saved raw text into a working directory. I then applied a series of four transformations to the article texts to clean the data and prepare it for sentiment analysis. I converted all text to lowercase, removed all non-alphanumeric characters (including punctuation and special characters), deleted all leading or trailing whitespace and condensed all repeated whitespace characters (e.g., extra spaces in a row). Next, I created a list of all player's last names and parsed each team's articles in order to search for occurrences of player last names. I wrote code (available on GitHub) to remove the suffixes in the names of players (e.g., Jr. or III) when searches for occurrences of athletes' last names were performed. Once I stored all player last names, I implemented built-in functions from the R library *tokenizers* to tokenize all article texts by sentence. With these sentence-long tokens, I wrote a loop to pull out all adjectives included in sentences with at least one mention of a player's last name (representing the keyword extraction stage). I then tied the original player statistics data with the newly formed adjective list for each player and subsequently used the *afinn* library to assign a sentiment score for all sentimentally-charged adjectives, where a score of -5 represents an adjective with extremely negative sentiment and a score of +5 represents an adjective with extremely positive sentiment. The *afinn* sentiment library is a lexicon developed by Finn Årup Nielsen that has pretrained sentiment scores for over 3,300 words. I determined sentiment scores for each player's adjectives using the values in the *afinn* library and then averaged these adjectives to produce a single sentiment score for each player (even though most players did not

have adjectives associated with them, and thus had a sentiment score of 0). Subsequently, I compiled a data frame with entries for each NCAA player, including information such as their team name, their player statistics, list of adjectives (if any were used), number of adjectives, average mentions per article, sentiment score, the team's total article count, and other columns that can be viewed on the linked GitHub.

Using R packages such as {shiny*}*, {shinyjs*}*, {plotly*}*, {shinythemes}, I created an R Shiny application to showcase this data with visualizations. The 755-line RShiny script is posted on the GitHub and the RShiny Sidelined application can be viewed here, https://nateyellin.shinyapps.io/Sidelined, which is best viewed on a laptop or desktop.

Results

The results of this research and accompanying data visualizations can be viewed on the interactive R Shiny app Sidelined (https://nateyellin.shinyapps.io/Sidelined/). In summary, there are three main categories of data that I investigated for gender bias in sports reporting at the league level:

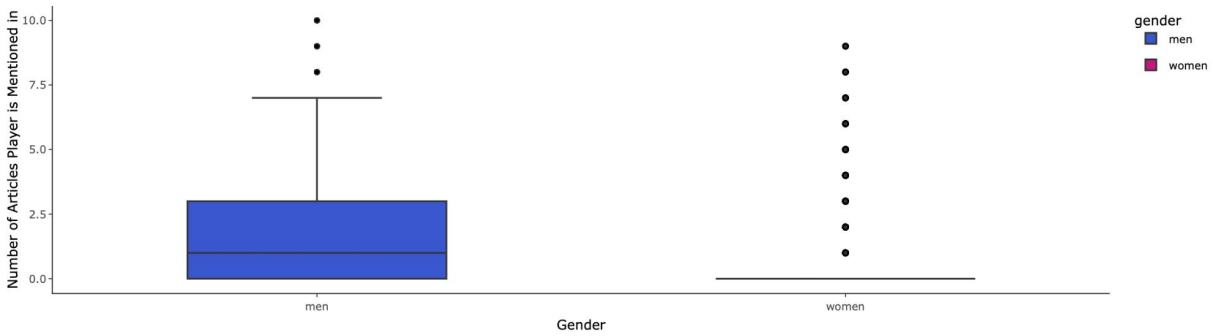
1. Quantity of Articles on ESPN
2. Sentiment of Adjectives used to describe both genders
3. Media Coverage based on Player Performance

The box and whisker plot in Figure 1 looks at the disparity in article mentions between men and women college basketball players on ESPN.com. Different quartile values can be seen for both genders when the R Shiny app Sidelined is accessed and the user hovers over the box plots. The median National Collegiate Athletic Association men's basketball (NCAAM) player is mentioned in one article, while the median National Collegiate Athletic Association women's basketball (NCAAW) player is mentioned in zero articles. In fact, the 75th quartile for men is

three article features, while this same quartile value is zero article features for women. Accordingly, being mentioned in even one article is considered an outlier for female athletes.

Figure 1

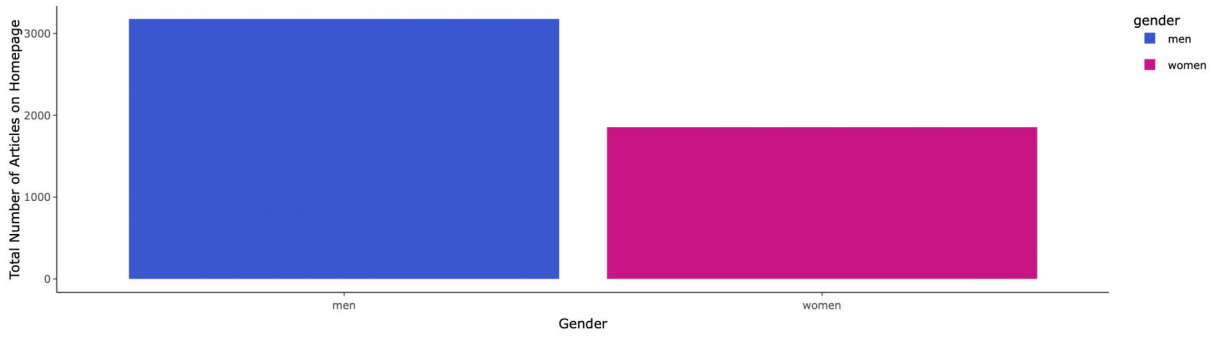
Quantity of Articles



In Figure 2, I use bar graphs to compare the total number of articles featured on the homepages for NCAAM teams with the total number of articles featured on homepages for NCAAW teams. When you visit ESPN.com and select a team, a number of recent articles are displayed. This graph compares the quantity of articles available on the homepages of teams for the NCAAM and NCAAW. Based on an ESPN.com web scrape on December 29th, 2022, there are a total of 3,178 articles available on NCAAM team home pages as compared to only 1,855 articles available on the NCAAW team home pages.

Figure 2

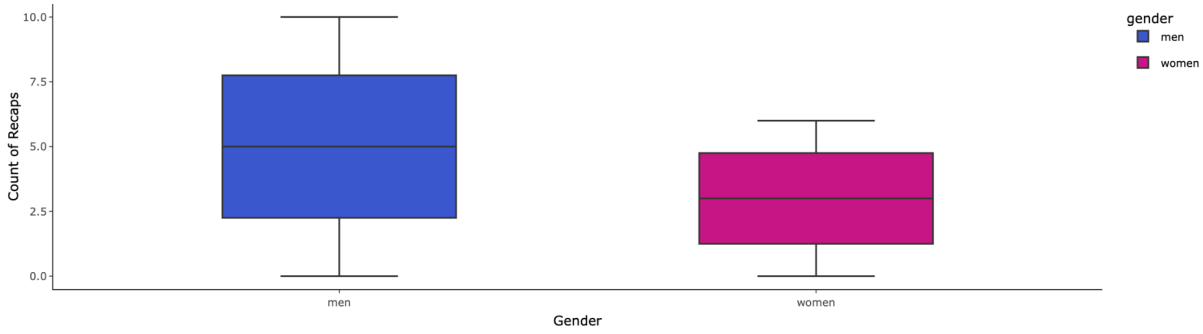
Total Homepage Articles for NCAAM Teams and NCAAW Teams



In Figure 3, I examine game recaps, short, computer-written summaries of recent games. Game recaps include basic information like the final score, leading scorers and game statistics for a game. Game recaps are important to update readers on the particulars of every game. Based on this analysis, there is a disparity between game recaps for the NCAA men's teams and for the women's teams. The median NCAAM team has five game recaps featured on their homepage, while the median NCAAW team has only three. Hovering on the graph on the app can further elucidate on different quartiles and values for this disparity.

Figure 3

Number of Games Featured on Team Homepages

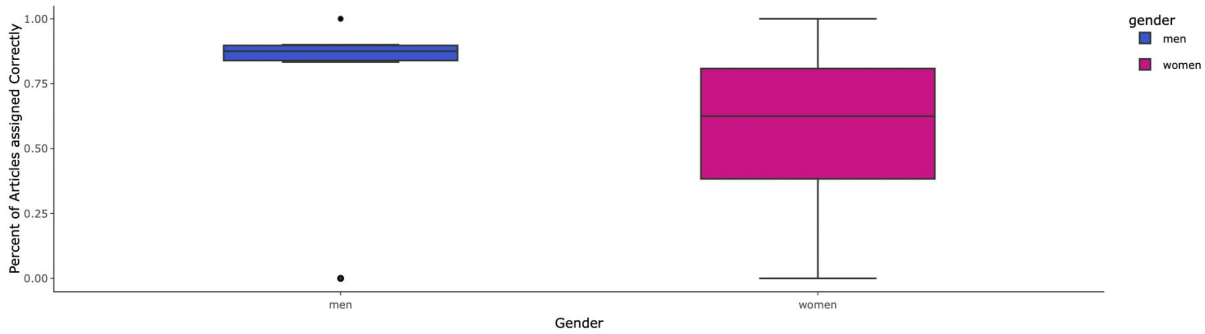


In Figure 4, I describe what percentage of homepage articles are assigned correctly. Every ESPN.com article is categorized by sport in their URL. For example, the URL for an article about professional baseball would include */mlb*. When focusing specifically on college basketball, the study evaluates how many articles scraped from the teams' homepages are correctly assigned in their URL. In other words, what percentage of articles found on the NCAAM team's homepage included men's college basketball and what percentage of articles found on a NCAAW team's homepage included women's college basketball. The results of the box and whisker plot show that the median NCAA men's team has 88% of its articles correctly assigned, while the median NCAA women's team has only 63% of its articles correctly assigned.

This underscores that if an ESPN user visits their favorite women's NCAA basketball team home page, there is only a 63% chance that the article featured on their homepage is related to the correct team or sport.

Figure 4

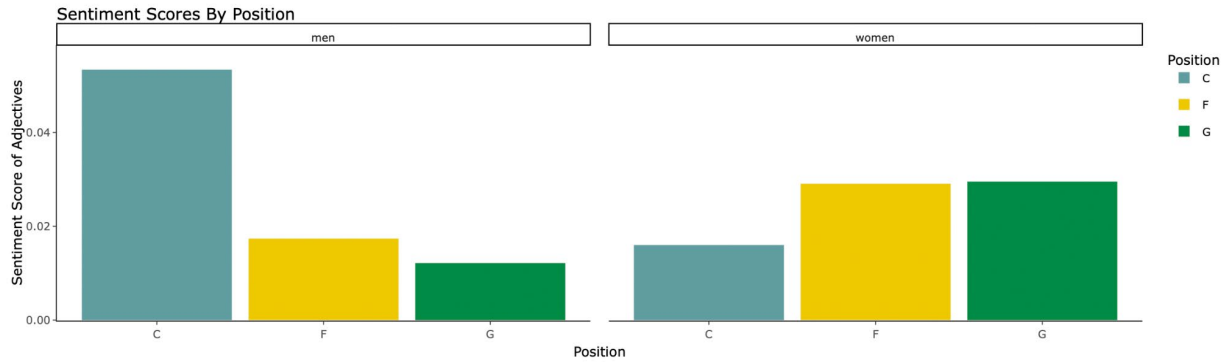
Percentage of Homepage Articles Assigned Correctly



I use the graph in Figure 5 to examine the sentiment scores of the adjectives used to describe players by their listed positions on ESPN.com. This graph below highlights a notable difference in sentiment scores of adjectives used to describe NCAA male and female athletes based on position. On the Sidelined application, a user can further hover over the graph to view the sentiment values for different positions played by basketball athletes. The data reveals that the adjectives used to describe centers in men's basketball are the most positive across both leagues and all position groups when position is analyzed. Interestingly, while centers in men's basketball have overwhelmingly positive sentiment, adjectives used to describe centers in women's basketball have much less positive sentiment.

Figure 5

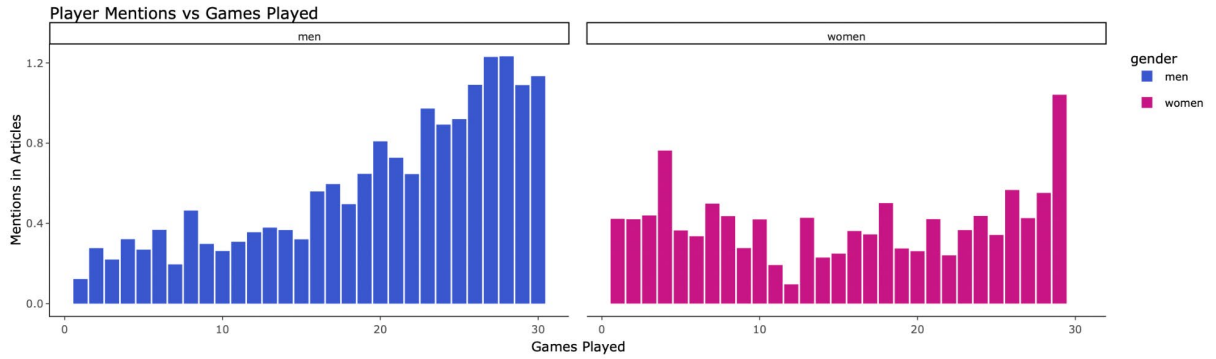
Sentiment of Adjectives Used to Describe Both Genders



In Figure 6, I show that male players who play more are mentioned more often, but there is no clear relationship in article mentions for female players based on number of games played. In fact, female athletes that play in 25 games receive as many mentions as those that play in zero games.

Figure 6

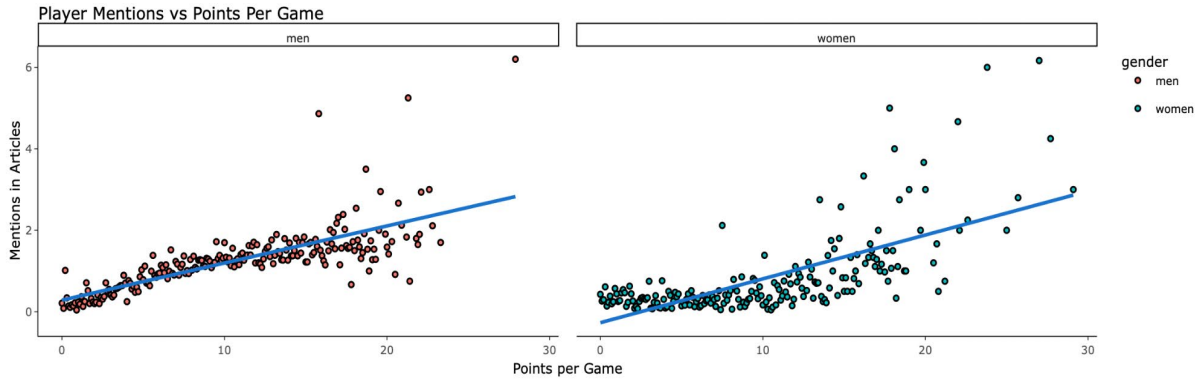
Media Coverage Based on Player Performance



In Figure 7, I reveal that in general, players that average more points per game receive, on average, receive more media mentions. However, this relationship is less clear for female players: the R^2 for female players is 0.46 while the R^2 for male players is 0.56. For the men, 56% of the variation in mentions can be explained by points per game while for the women, only 46% of the variation in mentions can be explained by points per game.

Figure 7

Player Mentions versus Points Per Game



Results at the Team and Individual Player Level

I also found gender bias at the team and individual player level. On the R Shiny application Sidelined, users can interact with the dropdown menus to select a specific NCAA team or player, view statistics and engage with interactive visualizations and a word cloud of adjectives used to describe that team or player. This interactive application enables coaches, athletic departments, fans and journalists to compare male and female teams and players across quantity of articles, average mentions per article, sentiment scores of adjectives used, and through a word cloud that details the adjectives used to describe such players and teams.

Discussion

Consumers of sports media assume that coverage is objective, accurate, responsible and fair. When selecting to cover a specific event, a number of decisions are made to determine how the story is "framed": what facts to include, how much coverage that story should be given, and what kind of words and pictures should be used to tell that story. All of these decisions are what sociologists and sports management researchers refer to as "agenda setting," because "media coverage (potentially) shapes the perception of the audience and that the amount of quality of coverage devoted to a specific sport, issue, or athlete can affect the perceived importance of the

sport, issue, or athlete (Hartmann 2019, p. 269). Producers, commentators, and journalists have referred to market forces to explain the lack of attention to women's sports, contending that coverage is a response to viewer and reader demand. These professionals reduce the issue too simplistically and ignore the circular nature of the media coverage-reception relationship. Empirical studies have shown that the kind of media attention given to athletes has a profound effect on creating or hindering consumer interest and demand and bears enormous implications for the lives of women and men within sport and beyond. Limiting the coverage of women's sports legitimates stereotypical gender roles for men and women in sports and ultimately, has the power to perpetuate social, political and economic differences between the two genders.

The research done in this study illuminates that women are still being *sidelined* in sports journalism, both in quantity of media coverage and qualitatively, in the sentiment of adjectives or words that are used to describe these athletes. At the league level, I found that female players were found to have received quantitatively less coverage overall, in game recaps and even when they play more. Being mentioned in one article is considered to be an outlier for a female NCAA basketball athlete. Through sentiment analysis, male players who play *center* were found to be described in articles with positive sentiment, while articles about female counterparts were not. Athletes who play the center position often rebound, box out and block shots—all requiring physical strength and dominance, traits not usually attributed to women. Further research should be performed to evaluate whether the media is perpetuating harmful stereotypes that females lack the traits associated with playing the center position or that these traits are "unwomanly" and thus, this position is negatively perceived when played by women. Further investigation should be also conducted to better understand why female athletes are still receiving less coverage and why those who play and score more do not receive as much media attention as their male

counterparts. Finally, subsequent studies should be conducted to evaluate why the NCAAW team's women's home pages are less likely to have an article correctly assigned and answer the question of why ESPN is posting articles of other sports on the homepages of NCAAW teams. It would be important to confirm that the NCAAW homepages are not being filled with irrelevant information to obscure the fact that there is insufficient coverage of its female players.

Conclusion

In this study, I highlight the significance of NLP as a valuable tool for investigating gender bias in sports journalism and shedding light on the pervasive issue of gender bias in NCAA sports coverage. NLP techniques offer a systematic, objective and efficient approach to analyzing vast amounts of text data and identifying patterns and trends that may not be immediately apparent to human readers. NLP was used in this study to quantify the presence and nature of gender bias in sports journalism. By analyzing language patterns and word choices in sports articles, I used sentiment analysis to measure the overall tone of articles and determine whether it is more positive or negative towards male or female athletes. Keyword analysis also invites users to be investigators and identify which words or phrases are most commonly used to describe male and female teams and specific players and whether these descriptions differ in terms of their connotations.

The findings of this research have implications for media professionals, sports organizations, and policymakers in promoting gender equality in sports journalism and fostering a more inclusive and equitable sports culture. NLP can be used to help identify the underlying causes and effects of gender bias in sports journalism. For example, researchers can use named entity recognition to identify the entities being referenced in the text, and determine whether female athletes are more likely to be referenced in relation to their physical appearance or

personal relationships rather than their athletic abilities. NLP can be further explored for developing and evaluating interventions aimed at mitigating gender bias in sports journalism. For example, NLP can be used to analyze the effectiveness of media training programs that aim to educate sports journalists on gender bias and encourage more equitable coverage of male and female athletes. Additionally, NLP can be used to evaluate the impact of policy interventions, such as the implementation of quotas or guidelines for sports media coverage. As elucidated in this study, the use of NLP has the potential to provide powerful insight into the ways in which gender bias may be perpetuated in sports journalism and the potential impact it can have on female athletes' media coverage, sponsorship opportunities, and public perception. Further research is warranted to explore the findings found in this study as well as the broader use of NLP in evaluating potential gender bias in the online coverage of other sports, at professional levels and in other mediums of social communication or mediatization. Using NLP techniques in this study has proven that it can be used as a powerful tool for investigating gender bias in NCAA basketball and more broadly, uncovering patterns and trends that warrant further attention in promoting more equitable and inclusive coverage in sports journalism.

References

Ancheta, A. S., Peet, J., Abuyen, A., & Shifflett, B. (2019). *Gender and Sport Journal of Kinesiology and Wellness, Student Edition*, *9*(2), 13–20.

Angelini, J. R., & Billings, A. C. (2010). An agenda that sets the frames: Gender, language, and NBC's Americanized Olympic telecast. *Journal of Language and Social Psychology*, *29*(3), 363–386.

Bishop, R. (2003). Missing in action: Feature coverage of women's sports in Sports Illustrated. *Journal of Sport and Social Issues*, *27*(2), 184–194.

Cooky, C., Messner, M. A., & Hextrum, R. H. (2013). Women play sport, but not on TV: A longitudinal study of televised news media. *Communication & Sport*, *1*(3), 203–230.

Cooky, C., Messner, M. A., & Musto, M. (2015). 'It's dude time!' A quarter century of excluding women's sports in televised news and highlight shows. *Communication and Sport*, *3*(3), 261–287.

Dunne, C. (2017). An examination of the photographic coverage of sportswomen in the Irish print media: A study of an Irish broadsheet newspaper. *Sport in Society*, *20*(11), 1780–1798.

Eastman, S. T., & Billings, A. C. (2000). Sportscasting and sports reporting: The power of gender bias. *Journal of Sport and Social Issues*, *24*(2), 192–213.

Fink, J. S. (2015). Female athletes, women's sport, and the sport media commercial complex: Have we really "come a long way, baby"?. *Sport Management Review*, *18*(3), 331–342. https://doi.org/10.1016/j.smr.2014.05.0

Fu, L., Danescu-Niculescu-Mizil, C., & Lee, L. (2016) Tie-breaker: Using language models to quantify gender bias in sports journalism. *arVix Computation and Language.* https://doi.org/10.48550/arXiv.1607.03895

Hartmann-Tews, I. (2019). Sports, the Media, and Gender. In J. Maguire, M. Falcous, & K. Liston (Eds.), *The Business and Culture of Sports: Society, Politics, Economy, Environment: Vol. 2: Sociocultural Perspectives* (pp. 267–280). Macmillan Reference USA.

Higgs, C. T., Weiller, K. H., & Martin, S. B. (2003). Gender bias in the 1996 Olympic games: A comparative analysis. *Journal of Sport & Social Issues*, *27*(1), 52–64.

Jones, D. (2004). Half the story? Olympic women on ABC news online. *Media International Australia*, *110*(1), 132–146.

Kane, M. J. (2013). The better sportswomen get, the more the media ignore them. *Communication and Sport*, *1*(3), 231–236.

Kian, E. T. M., Mondello, M., & Vincent, J. (2009). ESPN: The women's sports network? A content analysis of internet coverage of March madness. *Journal of Broadcasting & Electronic Media*, *53*(3), 477–495.

Lumpkin, A. (2009). Female Representation in Feature Articles Published by Sports Illustrated in the 1990s. *Women in Sport and Physical Activity Journal*, *18*(2), 38–51.

Martin, T. G., Suh, Y. I., Williams, A. S., Locey, J., Ramirez, J., & Alea, M. (2016). Comparative analysis of female and male coverage on ESPN's SportsCenter. *Global Sport Business Journal*, *4*(1), 14–22.

Messner, M. (1988). Sport and male domination: The female athlete as contested ideological terrain. *Sociology of Sport Journal*, *5*, 197–211. https://doi.org/10.1123/ssj.5.3.197

Messner, M. A., Duncan, M. C., & Jensen, K. (1993). Separating the men from the girls: The gendered language of televised sports. *Gender & Society*, *7*(1), 121–137.

Messner, M., & Cooky, C. (2010). Gender in televised sports. *Center for Feminist Research*, *39*, 437–453.

Tuggle, C. A. (1997). *Differences in television sports reporting of men's and women's athletics: ESPN SportsCenter and CNN Sports Tonight*. *Journal of Broadcasting & Electronic Media*, *41*, 14–24.

Weber, J. D., & Carini, R. M. (2012). Where are the female athletes in Sports Illustrated? A content analysis of covers (2000–2011). *International Review for the Sociology of Sport*, *48*(2), 196–203. https://doi.org/10.1177/1012690211434 230

Zimbalist, A. (2019). Female athletes are undervalued, in both money and media terms. *Forbes*. https://www.forbes.com/sites/andrewzimbalist/2019/04/10/female-athletes-are-undervalued-in-both-money-and-media-terms/?sh=1abb786813ed