iJournals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 11 Issue 12 December 2023

Gender Bias in Sports Journalism: A Comparative NLP Analysis of News Article Discourse

Darsh Damani
Dhirubhai Ambani International School,
Mumbai, India
darshdamani26@gmail.com

Reetu Jain
On My Own Technology,
Mumbai, India
reetu.jain@onmyowntechnology.com

Abstract: *Sports journalism, as one important societal influencer, faces a persistent challenge : gender biases present within its coverage of athletes.The significant distinctions and biased representations of female athletes in the sports media are examined in detail in this extensive study. Careful examination of datasets created by collecting news articles from top sports news sources and Kaggle reveal inconsistencies that should be considered carefully. A significant narrative disparity is revealed by sentiment analysis, which shows a persistent pattern of lower sentiment scores (averaging at 0.0795) for articles featuring female athletes. Male pronouns predominate in both datasets used for this study, based on the pronoun frequency analysis, which supports this pattern further. Insights obtained from the word frequency analysis show that the articles about male athletes tend to highlight terms like "won," "gold," and "medal," whereas articles about female athletes tend to highlight terms like "new," "last," "body," "upper," and "looking." The utilization of LDA topic modeling shows that articles about male athletes often contain themes like "win," "victory," and "champion," while narratives about female athletes tend to focus more on "sports," "last," and "new." Sentiment distributions indicate an increased positivity bias in narratives about male athletes (94.8% positive) compared to slightly lower levels in articles about female athletes (92.8% positive), showing continuing biases towards male athletes across sports journalism. The study of language patterns, athlete names commonly used, and stereotypical usage of words contributes to further reinforce these deeply held biases. Words like "strong," "powerful," and "dominant," which regularly appear in articles about male athletes, stand in sharp contrast to words like "emotional," "weak," and "sensitive," which are more frequently linked with female athletes. This discrepancy inhibits the growth of an inclusive sports environment. These findings highlight the urgent need for significant changes in sports journalism. In order to promote a culture that honors athletic achievement irrespective of gender, efforts have to be made toward fair coverage, language change, and bias elimination. The findings of the research should be taken as an undeniable call to action for supporters, sports organizations, and media groups to come together in order to revise narratives and promote credible representations of athletes in the media.*

Keywords: **Gender Bias, Female Athletes, Underrepresentation, Biased Portrayal, Text Analysis, NLP, Sentiment Analysis, Sports Journalism, Web Scraping**

1. Introduction

Sport as a cultural phenomenon is much more than just athletic competitions. It impacts societal opinions and beliefs along with providing entertainment. The media's depictions of athletes, shape the public perceptions about athletes. But within this connection, there's a concerning trend which is the unfair portrayal and underrepresentation of female athletes. This issue gives rise to numerous challenges, including limited professional opportunities for young women in sports and a scarcity of female role models for aspiring athletes. Data starkly delineates this issue as despite constituting 40% of athletes, a meager 4% of sports media coverage

focuses on women's sports. This disparity transcends mere quantitative assessment, permeating media narratives, elite sport participation, and employment in the field. The systemic gender biases entrenched in sports journalism manifest in various forms, from unequal coverage to sexual objectification, raising critical concerns necessitating scholarly investigation. Existing research in this domain illustrates disparities in media representation of athletes based on their gender, particularly in print, televised, and online platforms. Findings consistently reveal significant under-representation, skewed coverage favoring male athletes, and the perpetuation of traditional gender stereotypes within sports reporting. This research primarily focuses on the extent of inadequate representation and the biased portrayal of female athletes in sports coverage, undermining their athletic capabilities. It endeavors to comprehensively examine contemporary sports reporting and understand how differently male and female athletes are being portrayed in news articles, uncovering gender biases present in sports coverage. The research commenced with the curation of two distinct datasets: a refined collection of 251 articles from leading sports news websites and a comprehensive assembly of 31,900 sports-related articles sourced from Kaggle. Utilizing sophisticated transformer-based ML models capable of gender classification within textual content, these datasets underwent annotation, categorizing the collected articles into news related to 'male' or 'female' athletes. The approach used for analyzing the textual content encompassed an array of NLP techniques such as stemming, lemmatization, part of speech tagging, named entity recognition, stopword removal, sentiment analysis, topic modeling, and different kinds of keyword analysis. The findings elucidated disconcerting trends: articles featuring female athletes consistently obtained lower sentiment scores compared to those spotlighting male athletes. This discrepancy extended further, with nuances in portrayal reflecting underlying biases. Beyond being an academic inquiry, this research serves as an imperative call to rectify prevalent gender biases within sports journalism. The insidious influence of biased narratives not only perpetuates stereotypes but also impedes progress toward an inclusive and equitable sports landscape. Addressing these biases is fundamental to fostering an environment that embraces and empowers athletes, irrespective of gender.

2. Literature Review

The opinion of the general public about the athletes of different genders can be influenced by how the media represents the athletes in news articles. The disparities in the portrayal can also play a role in upholding the cultural norms for different genders. These reasons make it really important to study the gender bias present in the representation of the athletes by the media in news articles. In the recent studies that have been conducted on this topic, it has been observed that the media representation of athletes of different genders is significantly uneven. According to a recent research conducted to investigate gender bias in news articles, it was found that men were quoted more than three times as often as women in Canadian newspapers. Insights indicating the presence of gender bias in favor of male athletes were also obtained through an analysis of sports coverage on major media platforms. All these studies contribute to the growing body of knowledge related to presence of gender bias in media representations, particularly in sports coverage.

Implementing specialized software, **Fatemeh Torabi Asr et al. [1]** carried out a study to investigate gender bias in the media. This tracker counts the number of men and women quoted in news text by examining the daily publications of seven English-language Canadian news outlets. For enhanced data analysis, their research made use of Natural Language Processing techniques. The findings from the data obtained between October 1, 2018, and September 30, 2020 revealed that, in a variety of news outlets and time periods, men were quoted around three times more frequently than women. **Susan Tyler Eastman et al. [2]** compared sports coverage on ESPN and CNN in sportscasting, alongside The New York Times and USA Today in sports reporting. Their findings exposed a prevalent bias favoring men's sports, even during key moments in women's sports. Electronic media notably marginalized women's sports, whereas newspapers offered a more balanced portrayal. ESPN's SportsCenter showed notably higher gender bias compared to CNN's Sports Tonight, possibly due to differing audience targets. Additionally, The New York Times exhibited a concerning prevalence of gender bias, surpassing that of USA Today. **John Vincent's [3]** study examined British newspaper coverage of the 2000 Wimbledon Championships, comparing how female and male tennis players were portrayed in The Times, Daily Mail, and The Sun. Using content analysis and Connell's theory of gender power relations, the research explored coverage

amounts and recurring themes linked to gender and race. While coverage quantity showed minimal differences, qualitative analysis unveiled significant disparities. Male journalists tended to diminish the athletic achievements of female players through cultural stereotypes, trivialization, and sexual innuendo. In contrast, they consistently praised male players' athleticism, reinforcing hegemonic masculinity in their coverage. **Holger Ihle's [4]** study examined gender's influence on sports news prominence, focusing on disparities in TV reporting. Using a news-factors approach, the research analyzed seven German sports news programs to assess gender-based differences in news coverage between women's and men's sports. The findings suggested that while news factors didn't significantly differ by gender, women's sports received less prominence compared to men's. Surprisingly, this discrepancy wasn't due to journalists' perception of female athletes' performance as inferior. Instead, gender operated as its own news factor, contributing to the overall reduced visibility of women's sports in TV coverage. This inequality arose from the less prominent presentation of women's sports compared to men's. **Diane Ponterotto's [5]** paper examined the portrayal of female athleticism in press coverage, focusing on tennis player Maria Sharapova in English and Italian media. Using corpus-assisted analysis, it explored language patterns across languages in sports settings. The study uncovered a discursive frame that trivialized female athletes' bodies, stemming from two discourse strategies: one that eroticized the female body and another that portrayed the female athlete as child-like. This representation appeared to arise from sexist stereotyping aligned with male hegemonic interests.

In their study on gender bias in sports media, **Cheryl Cooky et al. [6]** focused on the limited and sometimes dismissive coverage of women's sports. Their extensive research, that involves a thorough review of six weeks' worth of broadcast news coverage from ESPN's SportsCenter and local Los Angeles affiliates (KABC, KNBC, and KCBS), showed an all-time low in women's sports coverage. They argued that the difference in coverage and quality highlights media priorities, emphasizing audience participation in men's sports at the cost of women's sports. This inadequate coverage reinforces the notion that sports primarily revolve around men, overlooking the substantial number of female athletes playing at the high school, university, and professional levels. **Adrian Yip's [7]** research delved into the portrayal of female and male tennis players on the Australian Open 2015 site and ESPN. Analyzing 357 articles, the study shed light on the disparities in media coverage and the negative depiction of female athletes, perpetuating entrenched gender beliefs. The findings underscored a prevalent trend across both platforms: female players were often portrayed negatively, emphasizing their athletic shortcomings and non-athletic attributes, such as appearance and personal lives. Despite glimpses of potential for using more gender-neutral language, the overall trend reinforced traditional stereotypes about women within sports media representations. **Nathalie Koivula's [8]** study explored gender is portrayed in sports media and its impact on society. It was investigated how mass media shaped beliefs about gender-specific sports behaviors by analyzing televised sports samples in Sweden during 1995/96 (1,470 minutes) and in 1998 (528 minutes). The findings from the analysis showed a lot of gender disparities in the amount of coverage given and the nature of the coverage. In less than 10% of the sports news that was analyzed, the coverage was given to female athletes, and in less than 2% of the same, The focus was on women playing traditionally masculine sports. This highlighted how televised sports promoted gender divisions and traditional expectations of femininity and masculinity, influencing the cultural values in sports. **Amy Godoy-Pressland's [9]** paper delved into the persistent under-representation of female athletes in print media, a concern recognized by feminist media scholars. Despite recent studies suggesting progress towards gender equality, the article analyzed the portrayal of sportswomen in five British Sunday newspapers across a 24-month period from January 2008 to December 2009. The findings consistently revealed a significant under-representation of sportswomen in British print media. **Dhiman Chattopadhyay's [10]** research investigated gender bias in sports news coverage within India's media portrayal of sporting events, an area previously overlooked despite India's substantial influence as a prominent democracy and news industry. During the 2014 Incheon Asian Games, a content analysis was conducted on two major English-language newspapers in India. Echoing findings from studies in the USA and Europe, the research uncovered a consistent trend: female athletes were often depicted as secondary to their male counterparts. Their coverage tended to downplay their significance, emphasizing feminine and glamorous aspects rather than their athletic prowess.

A research by **Bethany Shifflett et al. [11]** looked at how professional athletes are portrayed in ESPN Sport Science web videos. The whole ESPN Sport Science web video library was analyzed by the researchers, who investigated the disparities in content and amount of time dedicated to studying male and female players. 88% of the total video time was devoted to the examination of male athletes, indicating a bias in favor of male athletes. This disparity is consistent with other research that demonstrates just how little is known about women's sports. However, in contrast to traditional media coverage, these videos regardless of the athlete's gender emphasized athleticism. **Sainz-de-Baranda et al. [12]** examined Spain's top four Twitter sports accounts to explore gender biases in sports reporting. Analyzing 6544 tweets over five months, they found 96.19% focused on male athletes, with only 3.81% featuring women. Football dominated coverage (72.11%) for both genders, followed by basketball (6.63%). Despite Spanish women athletes' achievements, they remained significantly underrepresented. Female athletes received more coverage in "gender-appropriate" sports, but not in line with their accomplishments. The study revealed Twitter mirrors traditional media, perpetuating and sometimes amplifying gender biases in sports coverage. **Steph MacKay's [13]** study at the University of Ottawa examined varsity sports coverage in student newspapers from 2004 to 2007. Contrary to previous disparities, their analysis surprisingly showed minimal differences in the quantity and length of articles and photos between male and female athletes. Female athletes received more coverage overall, yet men's sports often featured on the front page. Notably, the textual analysis revealed that female athletes were rarely sexualized or trivialized, portrayed primarily as athletes rather than as gendered subjects in the student-run newspapers. **Christopher King's study [14]** scrutinized British national newspaper coverage of male and female athletes at the Olympic Games from 1948 onwards. Through content analysis in The Times and the Daily Mail, it revealed a marked increase in track and field coverage since 1948. Until the 2004 Athens Games, female athletes were consistently underrepresented compared to men. While recent years show improved equality in coverage for female track and field athletes, the dominance of men in sports journalism persists, as discussed in the study. In the study conducted by **Jamell Dacon et al. [15]** they analyzed gender bias in news abstracts highlighting the predominance of men in news coverage. They emphasized how females were often superficially represented or seen as inferior, which leads to their underrepresentation in news categories compared to males. They employed three text-analysis fairness metrics while analyzing 296,965 news abstracts. The findings of their study emphasized the widespread marginalization and socially constructed biases against women in the media. In an effort to detect both implicit and explicit gender biases in news articles at scale, the researcher devised a methodology employing natural language processing (NLP) techniques.

3. Methodology

A. Data Collection :

i. ***Web Scraping using newspaper3k library :*** To prepare the initial dataset of sports news articles for investigating gender bias, articles from top sports news websites such as ESPN, Fox Sports, CBS Sports, etc., were collected via web scraping using Python's newspaper3k library. The gathered articles encompassed various sports categories like cricket, football, tennis, etc. Functions from Python's 'os' module facilitated file and path handling. A list of dictionaries containing information like URLs, website names, and sports categories was formulated. This list underwent iteration, extracting URLs from each dictionary. Leveraging the newspaper3k library's build function, a source object for the web page linked with each URL was created. Subsequently, all articles from the webpage were downloaded, parsed, and various article details, including headline, content, and publish date, were extracted utilizing available attributes. Collating details of articles from diverse websites into a list, structured data was stored in text files. These files, housing articles from various websites, were generated and stored within a directory designated for the dataset using Python's os module. After text files holding articles from each website on the list had been prepared, they were converted into CSV files using Python's pandas module in order to enhance the structure of the dataset. Relevant details including the title, article content, publication date, website source, URL, and sports category were all included in these CSV files. The CSV files containing articles from all the selected websites were prepared, and then they were merged into a single, consolidated dataset. After additional processing, irrelevant rows were eliminated, leaving a final dataset with 214 articles from various sports categories and websites.

ii. ***Web Scraping using BeautifulSoup & Requests :*** Using Python's BeautifulSoup and Requests modules, an extensive range of sports articles were picked by scraping content from popular news websites including Times Of India, NDTV, and BBC. URLs of these websites' sports-related webpages were collected. Through an inspection of the HTML structure of the webpages, particular div tags containing anchor tags linked to pages with headlines and complete articles were identified. Within these webpages, div tags holding heading and paragraph tags containing the headlines and sport articles content were identified. The data extracted from these tags, including headlines and complete articles, were then stored into a CSV file under respective columns. This entire process was performed for various web pages containing sports articles and CSV files containing the headlines and articles from different websites were created. These CSV files were merged to create a single dataset containing the articles and headlines from all websites. This dataset was then combined with another dataset containing sports articles and the details associated with them scraped using newspaper3k library. Through text analysis on this combined dataset, potential gender biases present in sports journalism were explored.

iii. Custom Dataset Compilation from Kaggle Resources : In order to do a comprehensive analysis that produces informative results, an additional dataset from Kaggle was obtained. This Kaggle dataset contained news articles from a variety of topics. There were two subsets in the Kaggle dataset: a train set with 120,001 articles and a test set with 7,601 articles. By merging these sets, an extensive dataset containing 127,602 articles was created. 31,901 sports articles were obtained as a subset by extracting all articles labeled with class index 2 from this dataset, as the class index 2 represented the sports category and we are investigating gender bias in sports journalism. This refined dataset along with the dataset containing scraped articles served as the basis for investigating gender bias in sports journalism.

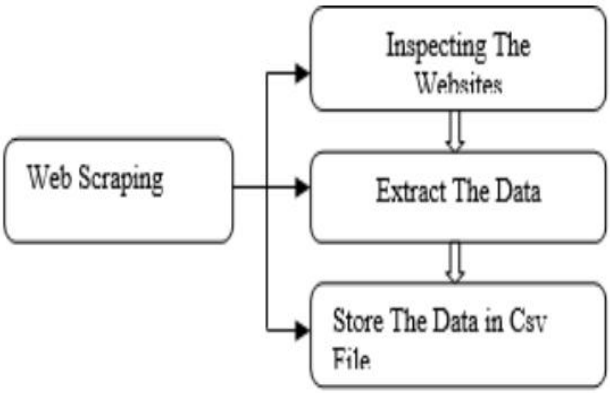

Fig 1. Web Scraping and Dataset Preparation

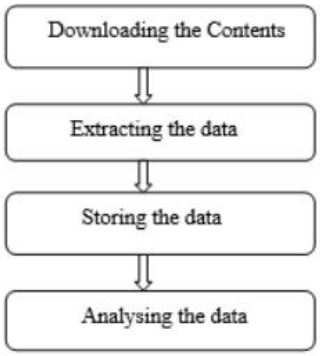

Fig 2. Major steps involved in this study

B. Data Preprocessing :

i. Scraped Articles Dataset Processing : Python's pandas package was used to explore the dataset containing all the scraped articles, and the results indicated that there are 251 rows and 8 columns. No missing values were found in any of the rows or columns of the dataset. 'Published_Date,' 'Website,' 'Source_URL,' and 'Category' columns were eliminated as they were not required for the investigation of presence of gender bias in the dataset. The names of a few columns were also changed; "Content" became "Complete Article" and "Title" became "Headline." Punctuation marks were removed from headlines and articles using Python's string module. The 'Complete Article' column was emptied of any rows that included irrelevant or incomplete articles. In order to further filter the dataset, URLs, numerals, and special characters from headlines and articles were removed using custom regular expressions. For consistency, all headlines and articles were also changed to lowercase.

ii. Custom Kaggle Dataset Processing : Python's pandas library was used to explore the dataset containing 31,900 rows and 3 columns. No missing values were detected in any row or column within the dataset. Custom regular

expressions were employed to remove punctuation marks, URLs, numbers, and special characters from the content of the dataset. The class index value for all the rows was 2 which represented the sports category for news articles, as all the articles present in the dataset were sports articles only, the class index column was redundant so it was removed. Additional preprocessing was done to convert the content of the dataset to lowercase. The column titles were also changed by renaming 'Title' to 'Headline' and 'Description' to 'Article.'

C. Annotation :

i. Obtaining a ML Model for Gender Labeling : It was crucial to identify these stories as pertaining to either male or female athletes in order to obtain insightful information on how male and female athletes were portrayed in sports news items across both the created datasets. Labeling every article in both databases manually would have taken a lot of time. For this reason, a machine learning model that has been trained to recognize gender based on textual content was employed. This model, sourced from the hugging face model hub, an online platform hosting various machine learning models for diverse tasks enabled gender identification of athletes featured in sports articles within both the datasets. Specifically, a text classification model named "name_to_gender" from this hub was used via the pipeline object of the transformers library. A custom function which incorporated this classifier, was used for labeling of genders for all articles across both the datasets.

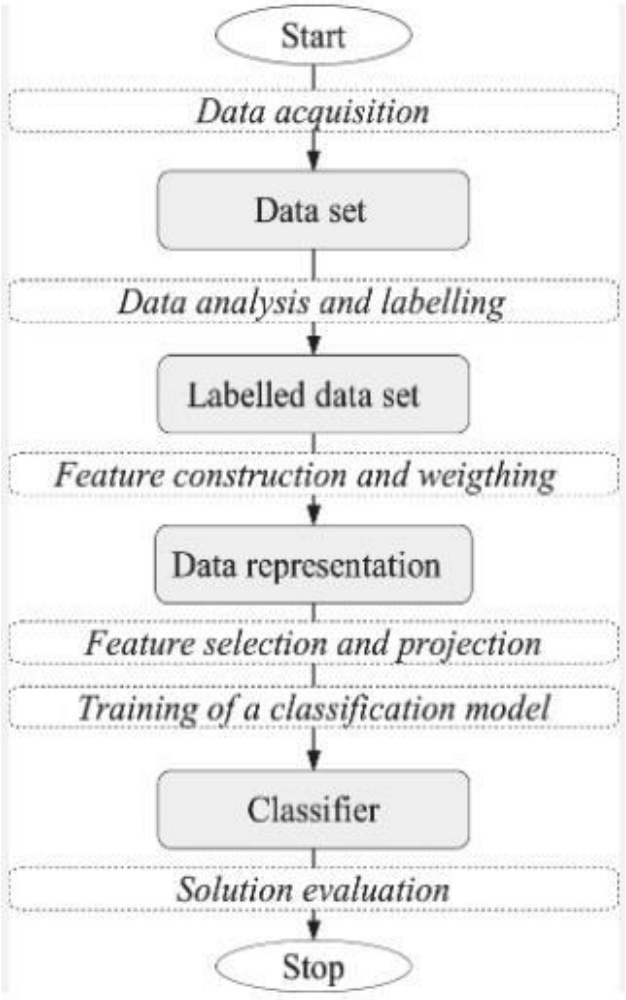

Fig 3. Text Classification

ii. Gender Labeling for both the Datasets : In this step, a 'Gender' column was added to both the datasets, with values indicating whether an article focused on a male or female athlete. Upon observing the configuration of the classifier, it became apparent that the maximum allowable word count for an article input into the classifier is 250

words. To standardize the dataset, all articles from both datasets were padded to reduce their word count to 250, especially considering some articles that even exceeded 1000 words. Using the custom function incorporating the previously created classifier using the pipeline object, the gender of the athlete in each article across both datasets was determined. Corresponding labels ('male' or 'female') were then stored in the 'Gender' column for every article within both the datasets indicating the gender of the athlete on whom a particular article was based.

D. Balancing the Datasets by Downsampling :

To ensure accurate results, both the created datasets were balanced by reducing the number of articles related to male athletes to match the count of articles related to female athletes. As the count of articles related to male athletes exceeded those related to females in both datasets, downsampling the male article samples was carried out using the sample function from the pandas library. This method involved adjusting the number of male articles to align with the number of female articles present in both the datasets. These balanced datasets were used for investigating gender bias as they would offer equitable analysis and unbiased conclusions can be drawn from the data.

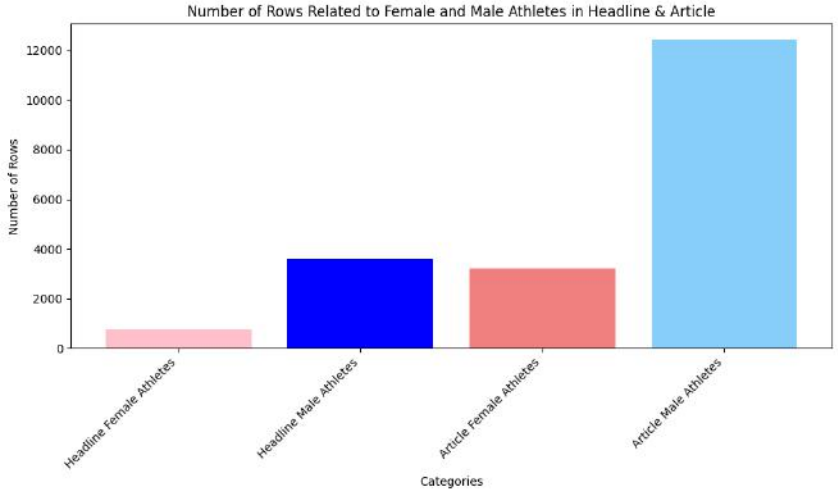

Fig 4. Imbalance in the custom Kaggle dataset

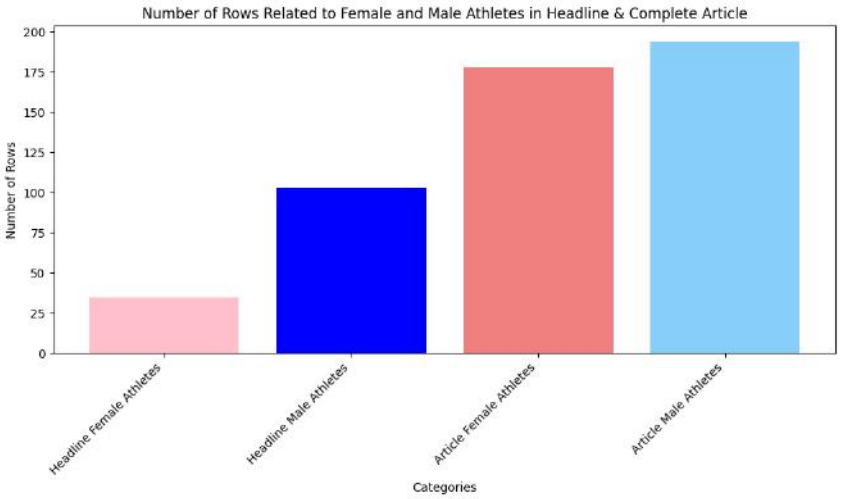

Fig 5. Imbalance in the scraped articles dataset

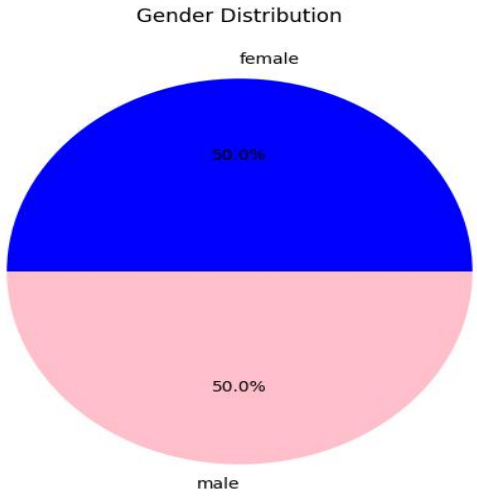

Fig 6. Balanced Datasets Gender Samples Distribution

E. Text Analysis :

i. Sentiment Analysis : This step involved using Python's TextBlob library, which offers functions to calculate sentiment scores for textual content, indicating whether the sentiments of the textual content are positive, negative, or neutral. First, all the rows from both datasets that contained articles labeled as 'male' in the Gender column were stored in a separate Pandas dataframe. Then, sentiment polarity scores were calculated for each of these articles within both datasets using the sentiment polarity attribute of the TextBlob object. The obtained sentiment polarity scores were stored in a new column named 'Sentiment Score' in both datasets. Based on the polarity scores, in another new column named 'Sentiment,' a label indicating whether the sentiment is 'Positive,' 'Negative,' or 'Neutral' was stored. Categorizing sentiment scores as positive, negative, or neutral allowed for a nuanced understanding of the sentiments conveyed. This exact same process was repeated to obtain the sentiment polarity scores and corresponding labels for all articles related to female athletes within both datasets. The obtained polarity scores and corresponding labels for articles related to female athletes were also stored in the 'Sentiment Score' and 'Sentiment' columns in both datasets.

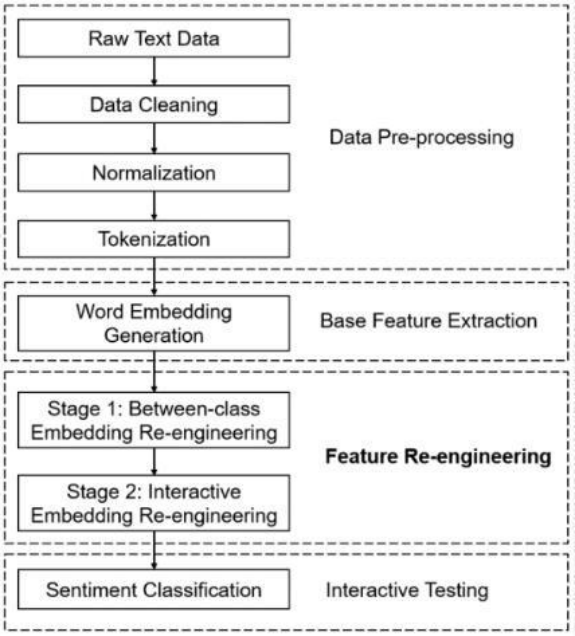

Fig 7. Sentiment Analysis Process Flowchart

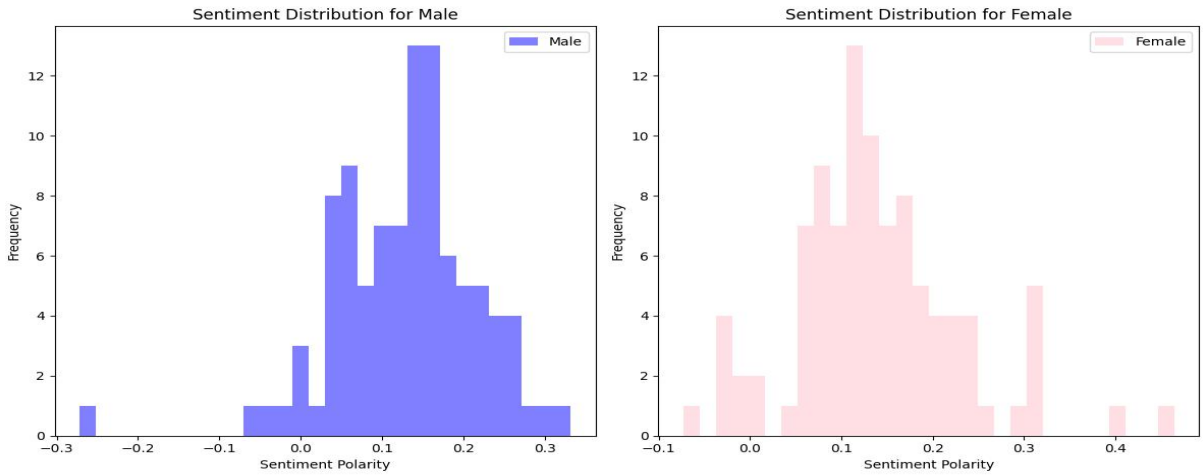

Fig 8. Sentiment Score distribution for scraped articles dataset

ii. Word Frequency & Topic Modeling : The 'NLTK' NLP library's stopwords module was used to remove stopwords from all the articles in both datasets. The articles' content was split into words, enabling iteration over each word from every article, making it easier to identify and remove stopwords. These words, often lacking significant meaning, encompass common occurrences such as articles (e.g., "the," "a," "an"), prepositions (e.g., "in," "on," "at"), conjunctions (e.g., "and," "but," "or"), and some common verbs (e.g., "is," "are," "was"). After the removal of stopwords, the Counter object from Python's collections module was utilized to calculate the frequency of occurrence of the most common words in articles related to male and female athletes. Topic modeling using the Latent Dirichlet Allocation (LDA) model from Gensim's Python library was conducted on both datasets. Articles labeled as 'male' and 'female' in the Gender column of both datasets were segregated. Stopwords had been previously removed, and the articles were divided into words stored in lists. Two separate dictionaries ('male_dictionary' and 'female_dictionary') were created using Gensim's corpora.Dictionary to convert words from the articles into numerical IDs. A corpus, representing bag-of-words for each document, was generated using doc2bow() for both 'male' and 'female' articles. Topic modeling was done separately for articles related to male and female athletes using Gensim's LDA Model**.** To identify major themes in sports articles on male and female athletes across both datasets, five topics were identified, by iterating over the data 15 times while doing topic modeling.

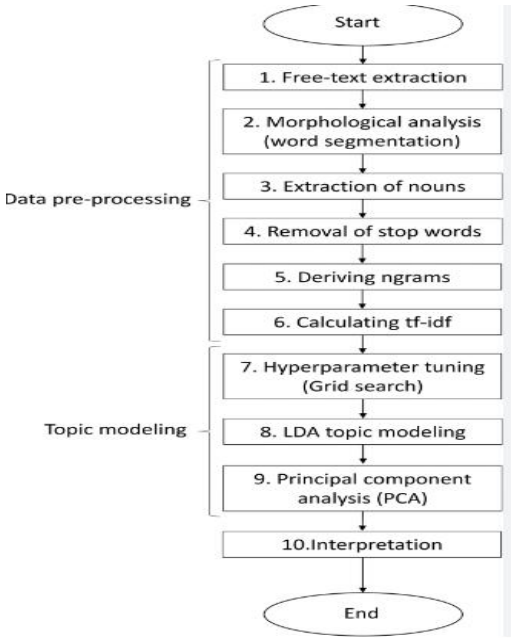

Fig 9. Topic Modeling using Latent Dirichlet Allocation

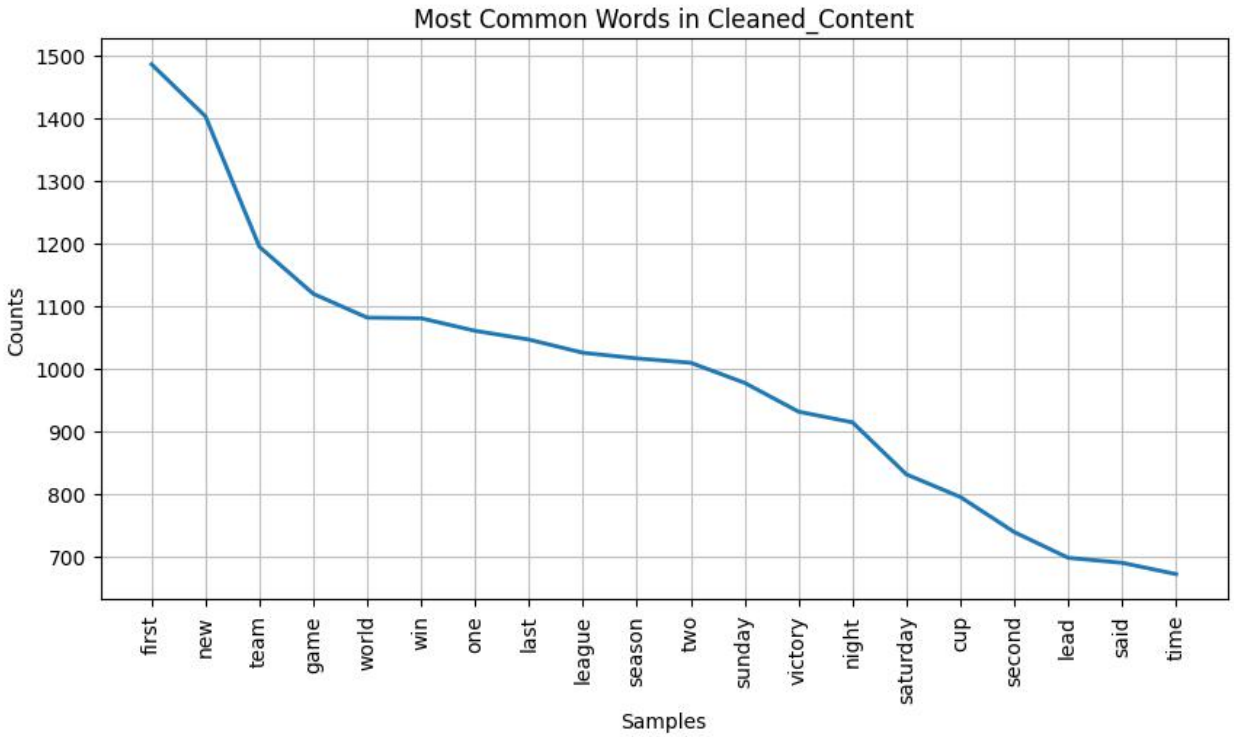

Fig 10. Frequency of the most common topics in articles from custom kaggle dataset

iii. Keyword Analysis : Various keyword analyses were conducted to unveil specific patterns and biases within the articles from both datasets. This encompassed examining the frequency of male and female pronoun occurrences using dedicated lists. Additionally, lists containing words like 'masculine,' 'sportsman,' etc., were employed for male athletes, and 'feminine,' 'sportswoman,' etc., for female athletes, to analyze athlete-related language. Stereotypical terms such as 'strong,' 'powerful,' 'strength,' 'explosiveness,' etc., were scrutinized for male athletes, while words like 'graceful,' 'elegant,' 'slender,' 'flexibility,' etc., were observed for female athletes. Moreover, words indicating gender bias, including 'weak,' 'sensitive,' 'emotional,' 'attractive,' etc., for female athletes, and 'tough,' 'dominant,' 'aggressive,' 'macho,' etc., for male athletes, were analyzed. The frequency of popular male and female athlete names was measured using extensive lists containing their names. Furthermore, the distribution of articles and headlines related to male and female athletes, based on specific word lists, was quantified.

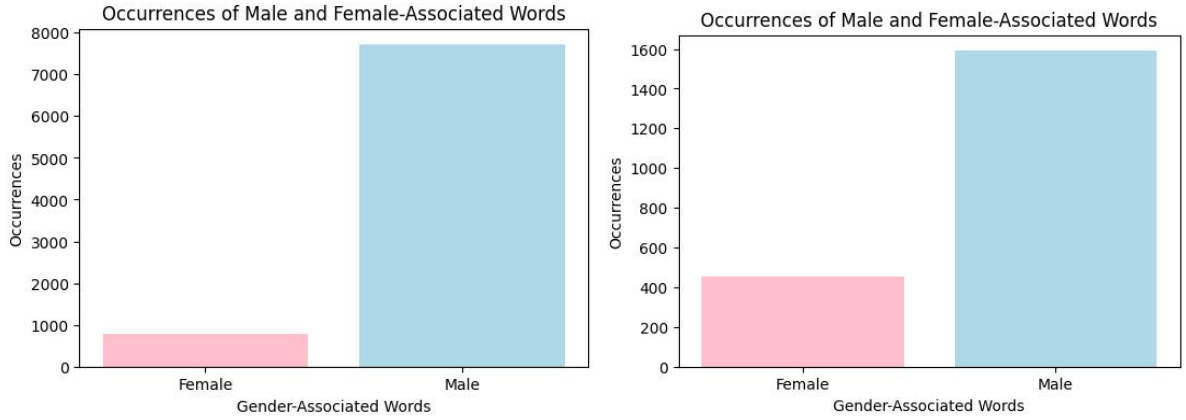

Fig 11. Frequency of the words associated with male & female athletes in articles from both the datasets

iv. Visualizations : The findings derived from the text analysis conducted on the sports articles within both datasets were presented through a variety of graphs and visualizations. Bar graphs were generated to display the word probabilities of the prominent topics identified in articles related to male and female athletes from both datasets after topic modeling. Line graphs were used to illustrate the word frequencies of the most common words identified in articles for male and female athletes separately in both datasets. Histograms were employed to represent the distribution of sentiment scores for articles featuring male and female athletes. Additionally, pie charts were created to depict the distribution of sentiment labels (positive, negative, and neutral) within articles related to male and female athletes across both datasets. Bar graphs were utilized to exhibit the frequency of occurrence of male and female pronouns within the dataset. Similarly, bar graphs were created to visualize the occurrences of stereotypical words associated with male and female athletes, as well as words indicating gender bias. Furthermore, a bar graph displaying the distribution of names of popular male and female athletes was also included. These visualizations offered clear and concise depictions of the results obtained throughout the analysis.

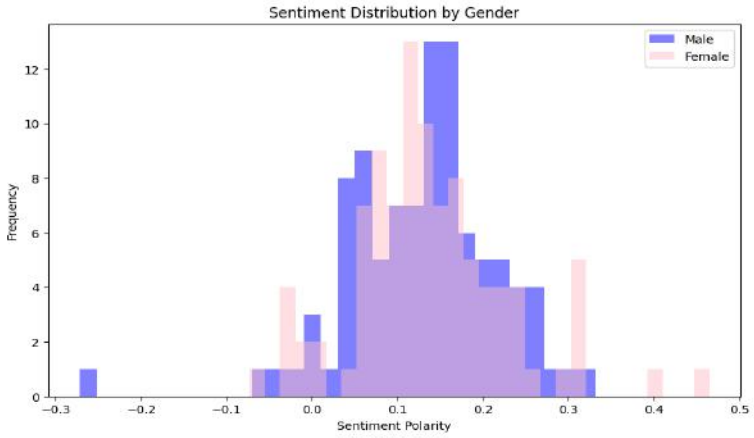

Fig 12. Sentiment score distribution in articles related to male & female athletes in scraped articles dataset

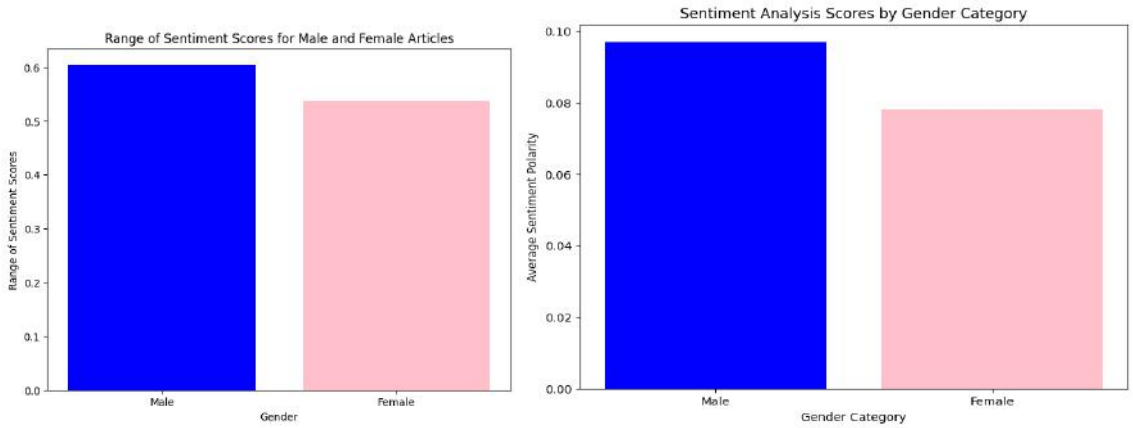

Fig 13. Range of Sentiment scores for articles related to male & female athletes in scraped articles dataset & custom kaggle dataset

4. Results

The examination of athlete-related articles revealed a significant imbalance in coverage between male and female athletes across both datasets, so balanced datasets were created to get proper insights. Pronoun frequency analysis revealed a consistent trend of male pronouns surpassing female pronouns in both datasets. Sentiment analysis conducted on the balanced Kaggle dataset showcased an average sentiment score of 0.0970 for articles featuring male athletes, compared to 0.0795 for articles related to female athletes. Notably, the lowest sentiment score recorded for male athlete-related articles stood at -0.2718 in the scraped dataset, while the same for articles centered on female athletes yielded a lower value of -0.0733. Analysis of word frequencies revealed different

themes, with male athlete articles emphasizing terms like 'won,' 'gold,' and 'medal,' while female athlete articles featuring terms like 'new,' 'last,' 'body,' 'upper,' and 'looking.' Topic modeling using LDA revealed prevalent topics in male athlete articles such as 'win,' 'victory,' 'first,' 'points,' 'champion,' 'lead,' 'league,' 'team,' and 'coach,' contrasting with topics in female athlete articles like 'sports,' 'last,' 'us,' 'open,' 'season,' 'new,' and 'league.' Additionally, sentiment distribution showed a higher positivity bias in articles about male athletes with 94.8% articles from the dataset being positive, 4.1% articles being negative and 1% articles being neutral compared to slightly lower positivity in female athlete articles with 92.8% articles being positive, 7.2% articles being negative, 0% articles being neutral. Examination of stereotypical word usage indicated a prevalence of terms like 'strong,' 'powerful,' 'aggressive,' and 'dominant' in male athlete articles, whereas 'emotional,' 'weak,' 'sensitive,' and 'attractive' were more frequently associated with female athletes. Furthermore, a noticeable disparity in the frequency of athlete names was observed, with a higher occurrence of male athlete names across both the datasets.

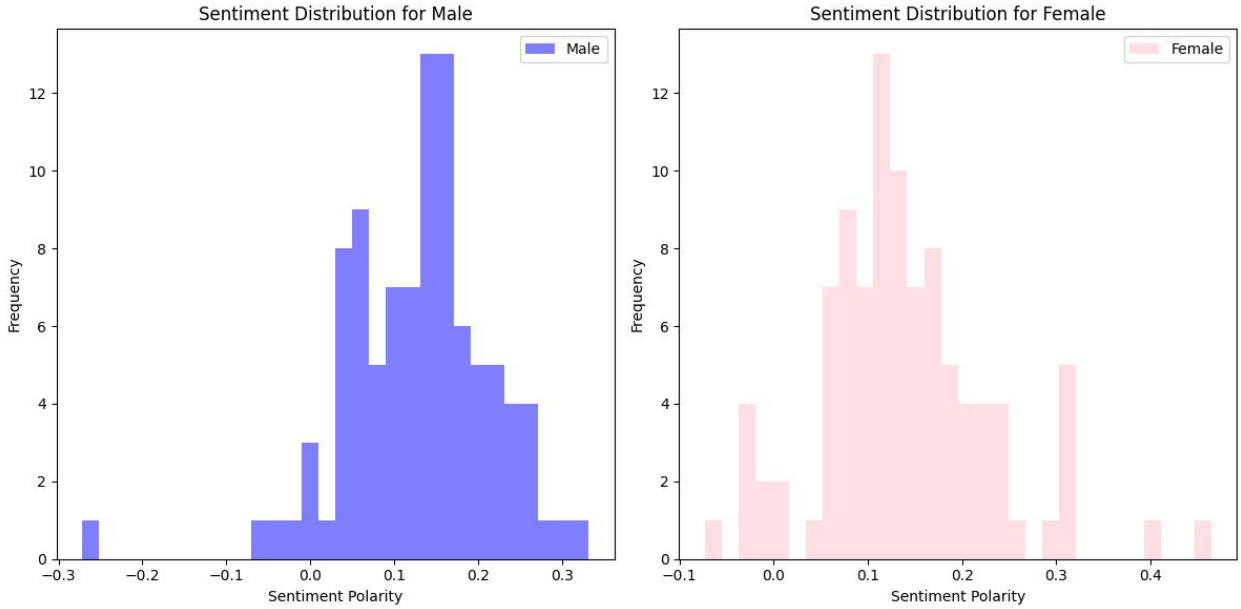

Fig 14. Sentiment score distribution in articles related to male & female athletes in scraped articles dataset

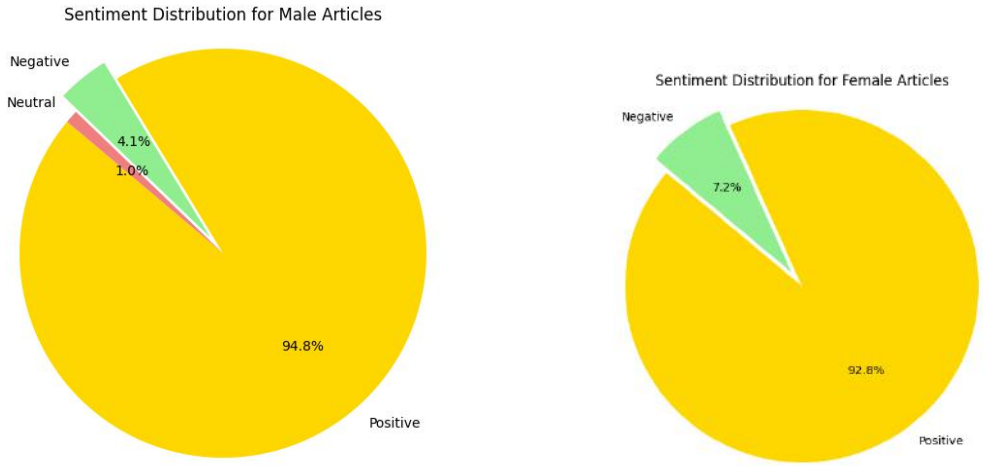

Fig 15. Sentiment distribution in articles related to male & female athletes in custom kaggle dataset

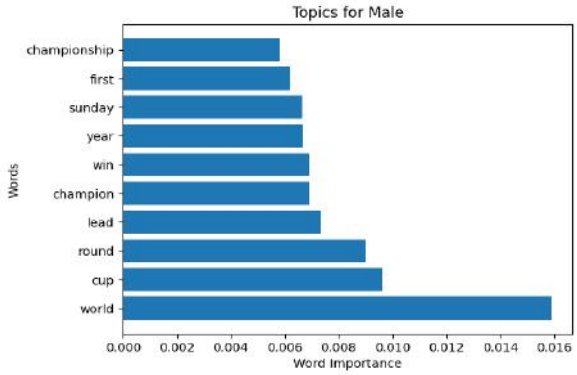

Fig 16. Prominent topics identified in articles related to male athletes in custom kaggle dataset during Topic Modeling using LDA

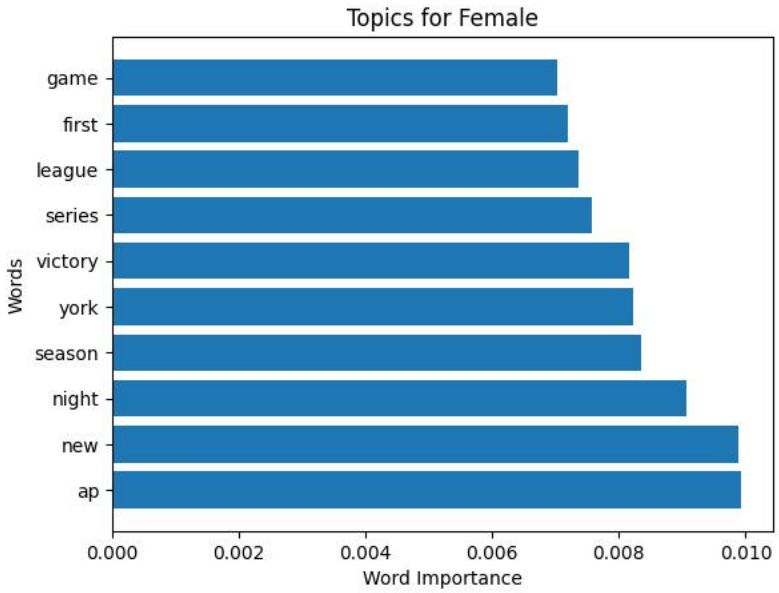

Fig 17. Prominent topics identified in articles related to female athletes in custom kaggle dataset during Topic Modeling using LDA

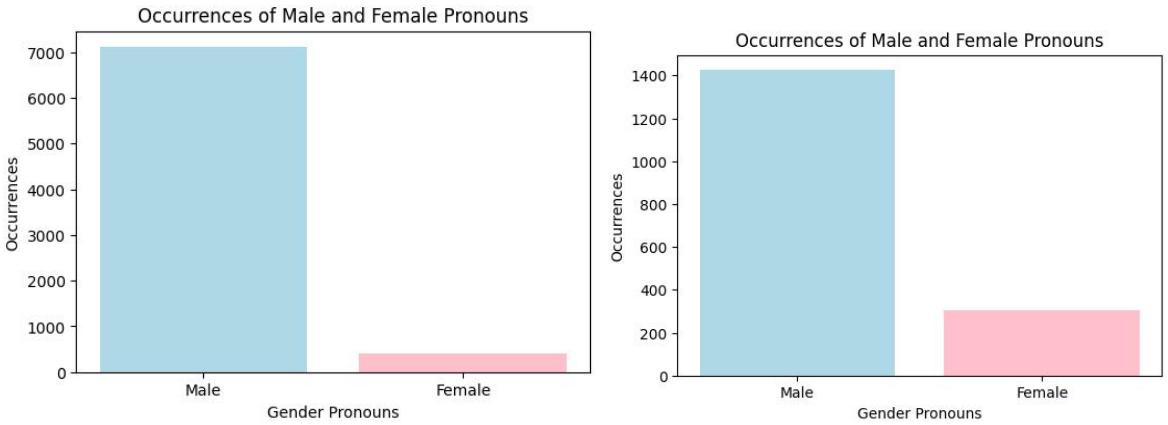

Fig 18. Distribution of male & female pronouns in custom kaggle dataset & scraped articles dataset

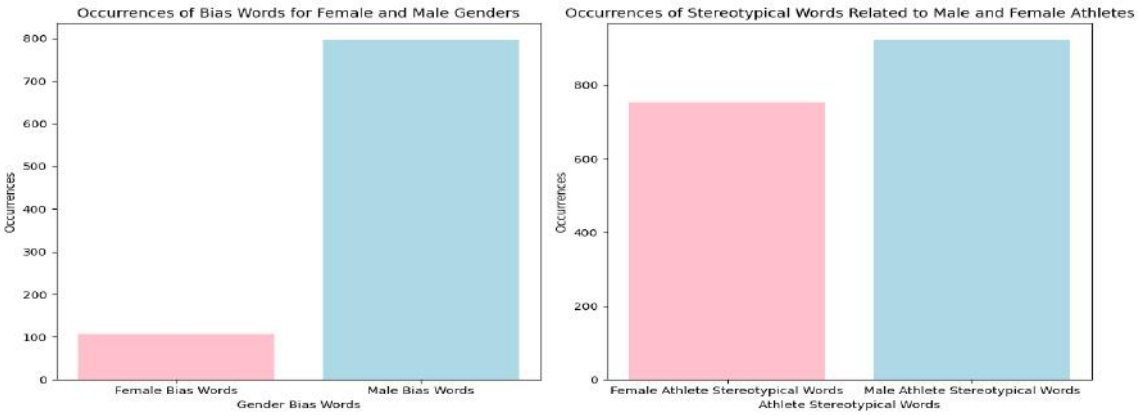

Fig 19. Distribution of stereotypical words & words indicating gender bias in custom kaggle dataset

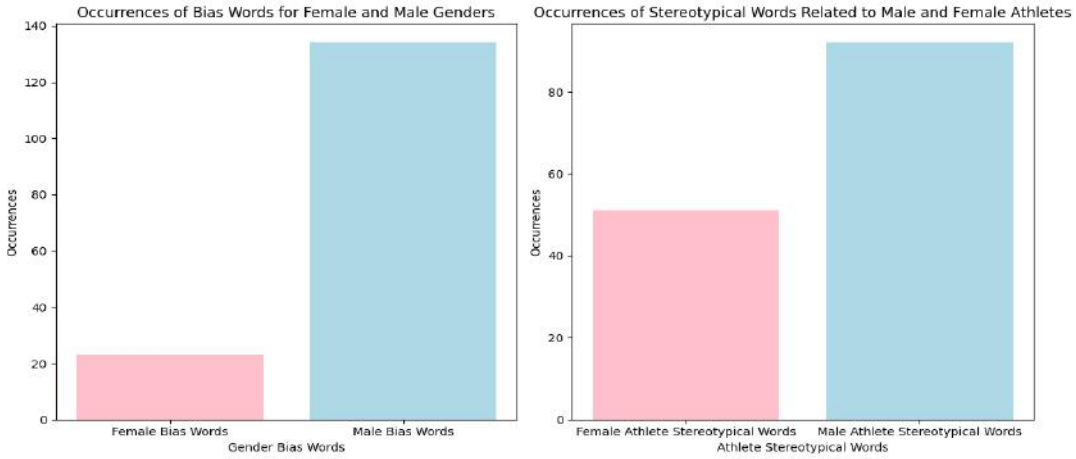

Fig 20. Distribution of stereotypical words & words indicating gender bias in scraped articles dataset

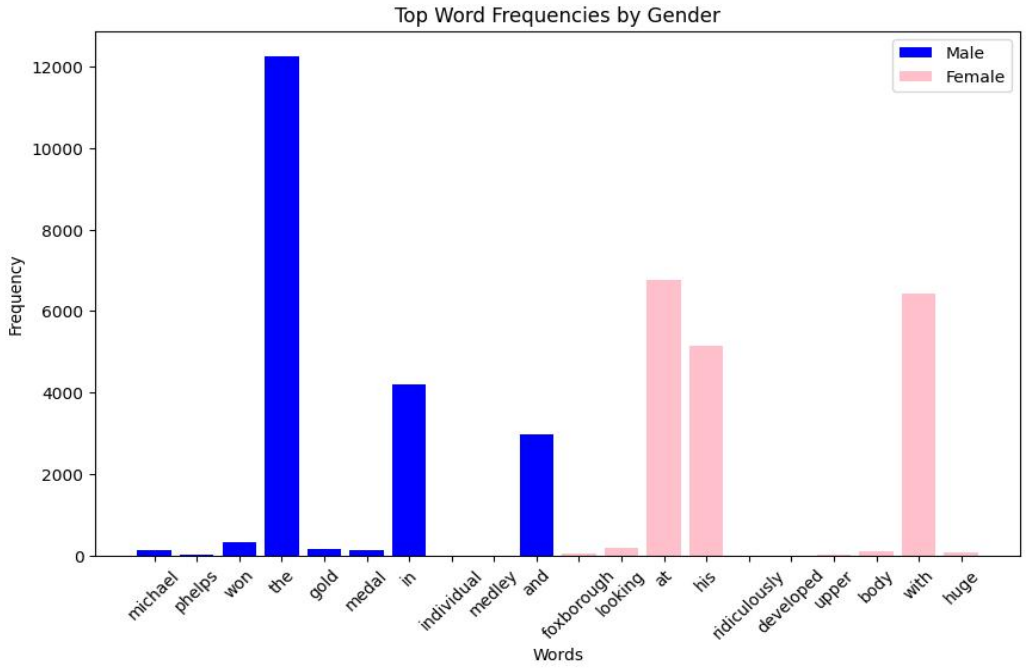

Fig 21. Frequencies of most common words from custom kaggle dataset

5. Conclusion

The conclusion of this study on gender bias in sports journalism reveals a surprising truth that there is a widespread pattern of inequality in how female athletes are portrayed in sports articles. The thorough investigation carried out on a wide range of sports-related articles highlights the need for significant modifications in portrayals of athletes in newspaper articles by the media. There is a clear difference in the sentiment analysis of stories featuring male and female athletes. Interestingly, articles on female athletes regularly scored lower on sentiment than articles about male athletes. This continuous pattern suggests that there is a significant imbalance in the framing of articles, which may have an impact on how the public perceives and feels about athletes of different genders. Even though the difference is slight in the sentiment scores of articles featuring male & female athletes, it contributes to the unfair representation of female athletes in the media and the upholding of gender stereotypes. Subtle patterns were also revealed by the careful analysis of word frequencies, topic modeling, and keyword analysis. These analyses throw light on word choices and thematic emphases that might be involved in the uneven portrayal of genders in sports journalism. These results highlight the need for an abrupt shift in the media's discussion of sports. Outside of the sporting arena, the repercussions of these disparities reflect societal norms and strengthen gender-based stereotypes. Because gender biases limit female athletes equal opportunity and recognition for their achievements, they impede the creation of a more inclusive sports environment. To remove these deeply embedded biases, media stakeholders, sports organizations, and society at large must work together in concert. Initiatives to eradicate established prejudices in sports media, create diverse and inclusive representations, and advance fair coverage are imperative to support. Such programs are necessary to ensure equitable representation and to foster an environment that celebrates sports accomplishments regardless of gender.

Acknowledgement

I would like to express my gratitude to the mentors Mrs.Reetu Jain & Mr. Suraj Sharma of On My Own Technology Pvt. Ltd. for extending their help in carrying out this project.

References

1. Asr, Fatemeh Torabi, et al. "The gender gap tracker: Using natural language processing to measure gender bias in media." PloS one 16.1 (2021): e0245533.
2. Eastman, Susan Tyler, and Andrew C. Billings. "Sportscasting and sports reporting: The power of gender bias." Journal of Sport and Social Issues 24.2 (2000): 192-213.
3. Vincent, John. "Game, sex, and match: The construction of gender in British newspaper coverage of the 2000 Wimbledon Championships." Sociology of sport journal 21.4 (2004): 435-456.
4. Ihle, Holger. "How gender affects the newsworthiness of sports news on German TV: An application of the news-factors approach to understanding gender-biased sports news presentation." International Review for the Sociology of Sport 58.2 (2023): 253-277.
5. Ponterotto, Diane. "Trivializing the female body: A cross-cultural analysis of the representation of women in sports journalism." Journal of International Women's Studies 15.2 (2014): 94-111.
6. Cooky, Cheryl, Michael A. Messner, and Robin H. Hextrum. "Women play sport, but not on TV: A longitudinal study of televised news media." Communication & Sport 1.3 (2013): 203-230.
7. Yip, Adrian. "Deuce or advantage? Examining gender bias in online coverage of professional tennis." International Review for the Sociology of Sport 53.5 (2018): 517-532.
8. Koivula, Nathalie. "Gender stereotyping in televised media sport coverage." Sex roles 41.7-8 (1999): 589-604.
9. Godoy-Pressland, Amy. "'Nothing to report': a semi-longitudinal investigation of the print media coverage of sportswomen in British Sunday newspapers." Media, Culture & Society 36.5 (2014): 595-609.

10. Chattopadhyay, Dhiman. "Gender Bias in India's Newspaper Coverage of Male and Female Athletes at the 2014 Incheon Asian Games." Global Media Journal: Indian Edition (2016).
11. Shifflett, Bethany, et al. "Gender bias in sports-media analytics." Journal of Sports Media 11.2 (2016): 111-128.
12. Sainz-de-Baranda, Clara, Alba Adá-Lameiras, and Marian Blanco-Ruiz. "Gender differences in sports news coverage on Twitter." International Journal of Environmental Research and Public Health 17.14 (2020): 5199.
13. MacKay, Steph, and Christine Dallaire. "Campus newspaper coverage of varsity sports: Getting closer to equitable and sports-related representations of female athletes?." International Review for the Sociology of Sport 44.1 (2009): 25-40.
14. King, Christopher. "Media portrayals of male and female athletes: A text and picture analysis of British national newspaper coverage of the Olympic Games since 1948." International review for the sociology of sport 42.2 (2007): 187-199.
15. Dacon, Jamell, and Haochen Liu. "Does gender matter in the news? detecting and examining gender bias in news articles." Companion Proceedings of the Web Conference 2021.