Gender Bias in Sports Dialogue

Federico Cimini and **Ryan Martin** and **Khushi Shelat** and **Kyle Sullivan**
University of Pennsylvania - CIS 5300

Abstract

For our project, we set out to extend the research from a paper concerned with gender bias in sports interviews, specifically in tennis. First, we attempted to replicate the findings of the paper, calculating the perplexity of questions asked toward men and women relative to the corpus of game commentary transcripts assumed to be unbiased using n-gram models. Next, we continued studying perplexity, this time with an LSTM to encapsulate the context of the commentary and the questions. Finally, we performed experiments with LLMs and topic modeling to see if they would reinforce or contradict our findings with perplexity. Overall, we did not identify a significant trend of bias in questions toward either gender. We believe that further research could reveal bias in certain cases, such as how interviewers may ask questions unrelated to tennis to high-profile celebrity players for social media and tabloids.

1 Introduction

Being avid sports fans, we can't help but notice the differences in language when journalists and the general public speak about male and female athletes. In tennis, for example, for every question Rafa Nadal gets about his game, Serena Williams seems to get a question about her fights with referees. Is this anecdotal evidence, or is there some inherent bias on the language people use when talking about athletes? In this paper, we aim to use NLP to analyze gender bias in journalist's post-game interviews with players, improving and further developing upon current studies on the topic.

We are starting our project by replicating the study "Tie-breaker: Using language models to quantify gender bias in sports journalism" Fu et al. (2016). We will utilize the same dataset used in the paper and build our own language model to analyze the text. We will measure bias using perplexity regarding the interview's relevance to the sport, as done in the paper, and we will compare our results to those of the original researchers. We then proceed to replicate the strategy with more modern language model algorithms, and finally we suggest and experiments with methodologies that don't involve the use of perplexity.

1.1 Baseline Paper

A graphic representation of our baseline paper can be seen in Figure 1.

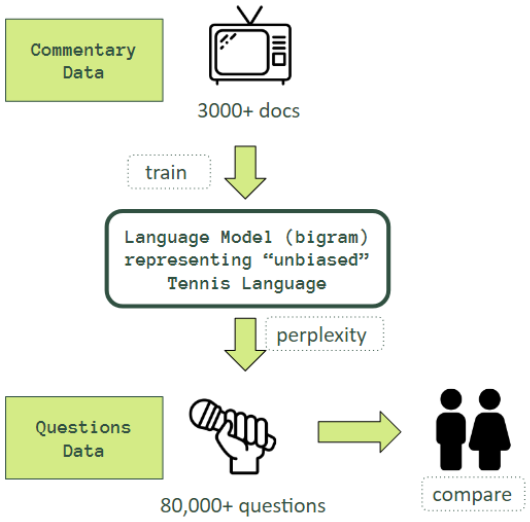

Figure 1: Illustration of baseline procedure as described in Fu et al. (2016)

The basic idea involves building a language model with "pure" tennis vocabulary. The authors of the paper hypothesize that the data from the match commentaries represents an unbiased representation of tennis terminology. Then, perplexity is calculated on questions from interviewers to athletes. The formula for perplexity used by the paper is:

$$P(w_1, w_2, ..., w_n) = \sqrt[N]{\frac{1}{P_{source}(w_1, w_2, ..., w_n)}}$$

It is argued that if the question is related to tennis, the perplexity will be low, but if the question is unrelated to tennis, then perplexity will be high. If there's a significant difference between the average perplexity of questions to male athletes vs questions to female athletes, it can be argued that the questions are gender biased. An example of questions of different perplexities, according to the paper, can be seen in Table 1.

Perplexity	Sample Question (from paper)
Low	What about your serve, Rafa? The tiebreak, was that the key to the match?
High	Who designed your clothes today? Do you normally watch horror films to relax?

Table 1: Sample questions from paper hypothesis

The paper finds that in all the different segments there is a significant difference between the perplexity numbers of male vs female questions. This would indicate that female questions are less related to tennis. We will try to recreate these results and expand the analysis of the paper.

For the first expansion, we build different types of language models and apply the same technique of measuring bias through perplexity. For the second expansion, we attempt to measure bias through other methods, arguing that maybe perplexity isn't the best approach possible.

2 Literature Review

Outside of this baseline paper, we examine the current literature around gender bias in sports as well as generalized literature around understanding bias in text corpus. Minot et al. (2021) examine textual medical data similarly to our exploration of sports interview data. It focuses on detecting "gendered" text and uses a BERT-based model to "de-bias" the dataset. It aims to de-bias the data in a way that is also able to preserve important health-related information. The authors use two open-source datasets, which include more than 1300 clinical documents and information on 26,000 males and 20,000 females approximately.

The authors use a rank-turbulence divergence (RTD) method to identify gender bias in the input data. The RTD method is applied by first extracting the n-grams in the text and their corresponding frequency distributions. They calculate the frequency of usage of n-grams within each class (Male and Female). They quantify "'genderdness' of n-grams by comparing their frequency of usage in notes for male and female patient populations," which is used to rank the most biased 1-grams from highest to lowest frequency. These 1-grams are sequentially removed from the data; after removal, to preserve sentence structure, regular expressions are used to replace the removed 1-grams with a space character. After this process is completed, the authors use a BERT-based classifier for gender and health condition classification on the test data. The results note that even removing the top 10% of 1-grams (i.e. debiasing) can lead to a near random classifier for gender, with only a small reduction in accuracy for the health condition classification. The paper also ultimately uses general BERT rather than clinical BERT, while comparing both. This paper illustrates an interesting and efficient approach to debiasing medical text. In the context of our work, it proposes (1) a new method for identifying biased text (RTD), (2) debiasing the text (through iterative removal of 1-grams and (3) a way to compare the results before and after trimming (BERT classification accuracy). Applying this approach to our dataset to improve the perplexity-based approach of the original paper is a direction we are considering. Below is an image taken from this paper, visually illustrating the authors' approach.

A second paper by Sun et al. (2019) is a 2019 literature review of (1) recognising bias in NLP models and (2) discussions of gender bias based on four forms of representation bias. This paper is selected by our team as one to highlight in our literature review because, while our task is not to debias an NLP model, it is focused on understanding definitions of gender bias and how to debias text, which can be informed by this prior literature.

The paper highlights three methods of observing gender bias (1) Adopting Psychological Tests to show biases in word embeddings, (2) analyzing gender subspace in (3) embeddings and measuring performance differences across genders. The third is particularly interesting because it relates to the methodology from Paper 1. Paper 2 states that "in an NLP task, a model's performance should not be heavily influenced by the gender of the entity mentions or contexts in the input." It therefore identifies differences in model performance across classes as a measure of gender bias. Similarly, Paper 1 identifies success as meaning that the classification model is able to effectively classify health condi-

tions even once the dataset is debiased and if the gender classification model is essentially random. This bolsters the usage of this method as a benchmark for debiased text.

Secondly, the paper highlights six methods for debiasing, including (1) data augmentation (2) gender tagging, (3) bias fine-tuning, (4) learning gender-neutral embeddings, (5) constraining predictions and (6) adversarial learning. The second half of these methods are focused on reducing bias through directly changing the NLP model itself. The methods that we believe would link best to our study are data augmentation and removing gender subspace. Data augmentation involves enhancing an unbalanced dataset with more balanced data; this is certainly something we could explore through synthetic text generation or something similar. Regarding removing gender subspace, Paper 1 applied a similar method, where it utilized cosine similarity between the gender classes as a measure for bias and as an indication of 1-grams needing to be removed. Both augmentation and subspace removal could indicate strong directions for our work to take when aiming to reduce bias in our textual data.

Similarly, Babaeianjelodar et al. (2020) focuses on the question of gender bias measurements related to NLP and textual data more broadly, which we hope to work with in the sports context. The most interesting thing about this paper is its in-depth discussion of a quantitative definition of gender bias in text.

The authors analyze data from numerous different domains, including fiction books, online news, Quora and Wikipedia. They defined a measurement of gender bias focused on the proximity of a gendered term (she/her/daughter vs he/him/son etc.) to words describing professions. There exists a crowd-worker sourced list of 340 different professions that this paper utilized as a way to measure and compare bias levels of each of the corpuses. While this may not be directly applicable to our context, considering the fact that sportswomen and men are all within the same broader profession, we could similarly build a bias score with the same logic and using sports terminology. This would expand on the original paper's measure of perplexity, giving us a proximity of a gendered term to a sporting reference.

The methodology in a paper by Rao and Taboada (2021) involves using Latent Dirichlet Allocation, a topic modeling technique, to identify topics and distributions of keywords in news articles. The study analyzes a two-year dataset of news articles from mainstream Canadian media in English. The findings reveal that certain topics in news articles prominently feature either women or men and exhibit different types of language. This methodology can be used for further study to analyze gender bias in media and contribute to efforts to achieve gender parity.

It demonstrates the applicability of topic modeling as a quantitative approach to view gender bias between texts. It is possible that topic modeling could also be applied in the sports media context for our work, though it is a more qualitative approach than other directions of literature.

There are a range of other miscellaneous papers that explore this area of literature in different ways. Particularly, in the sports context, Pereira Fernandez (2021) asks two interesting questions that could help us explore the definition of sports gender bias differently (1) does POS tagging reveal a difference in the structure of the questions asked and (2) does seeding reveal a difference in the type of sports language used (football). An older study by Aull and Brown (2013) on WNBA vs NBA, looks at "blame" measurement in NBA commentary compared across gender. This could be particularly interesting as a measure in our dataset, which contains journalistic questions and live commentary in games. Finally, Harrison et al. (2023) explore both visual and textual data for gender bias across different sports in a unique study; they note that their key limitation is a smaller sample size.

3 Experimental Design

3.1 Data

The data from the baseline paper consists of two different datasets:

1. **Commentary data**: this includes full transcripts of gender/balanced, live commentary of tennis matches. This data is used to train an "unbiased" language model that only has language related to the sport. The assumption is that this data would not show signs of bias and faithfully represents tennis vocabulary and lingo. This dataset also contains

match information (players and result) and whether it was a female or male tennis match. There are 3962 documents of text data.

2. **Questions data**: this consists of questions asked by journalists in post-game interviews. It also contains information about the player to which the question is directed to: name, gender, tennis rank, and whether they had won or lost the match when the question was asked). This is the data that is analyzed to see if questions are biased against female players or not. It covers a total of 6467 post-match press conferences, and if counting individual questions, we have over 80,000.

This data was analyzed and cleaned. The full results of data analysis can be seen in the Appendix, but here we'll show some highlights:

- **Word clouds reveal differences between genders**: if we look especially at the disproportionately occurring words for male vs female athletes, we see stark differences between them (Figure 6). For male athletes, some words that stand out are '*investigation*', '*supervisor*', and '*antidoping*'. For female athletes, however, we see words like '*shopping*', '*boyfriend*', and '*manicure*'.

- **Word characteristic scores by gender**: if we measure the valence, arousal, and dominance of the words in the journalists' questions, we see some differences between male vs female athletes (Figure 7). In particular, female questions have higher valence(the pleasantness of a stimulus) than male questions.

- **Number of questions according to rank**: We see that the top 10 male athletes have a much higher number of questions in the dataset than the top 10 female athletes, and for lower ranks, the numbers are more similar, with female athletes having a slightly higher number of questions. (Figure 8)

These results indicate already that there seems to be a difference between the nature of language in male vs female questions.

4 Experimental Results

4.1 Simple Baseline

For our simple baseline, we also used a similar concept as perplexity as a measure of bias, but instead of measuring relative to the commentary which might also be biased, we generated a long list of tennis-related words using ChatGPT to determine whether there was an obvious difference in questions containing these words. More specifically, we first identified which 30 of the generated words appeared most commonly across all the questions. We proceeded to count how many of the questions asked to each gender contain at least one of these words, and found that 30% of questions asked to men contain these words while 20% of questions asked to women contain these words. There are only 15% more questions for men than for women, so it is unlikely that this difference simply arises from all the common words being pulled from questions toward men since the majority is not that significant. These results can be seen in Table 2.

Gender	% of words related to tennis for top 30 ranked players
Male	32.48 %
Female	21.28%

Table 2: Results from Simple Baseline analysis

4.2 Complex Baseline

For our complex baseline, we attempted to recreate the bigram method of the paper in order to see whether our results matched theirs, with a bigram model as well as a trigram model. At first, we ran into issues with smoothing, leading to infinite perplexity for questions containing bigrams that had never been seen before, but solved this by implementing an unknown token which would be assigned to bigrams seen fewer than two times as opposed to only those that have never been seen. Our results appear to match those of the paper for most models, as they show more perplexity for women than for men ranked 10 or higher, but not all. And none of our model give the same values that the paper got using their KenLM model.

Upon closer inspection of questions with high perplexity, we noticed that not all of these were biased or unrelated to tennis. For example, one game-related question with extremely high perplexity reads, "*That last set seemed like a faultless*

performance. Is that the way you saw it?" Our initial guess is that the bigrams (*a, faultless*) and (*faultless, performance*) did not appear anywhere in the commentary transcripts, leading to high perplexity even though it is semantically related to tennis. We are curious whether the authors of the paper manually inspected enough of the questions they identified as high-perplexity to determine that perplexity relative to the commentary was an accurate indicator of a biased question. Our results can be seen in the graph in Figure 2.

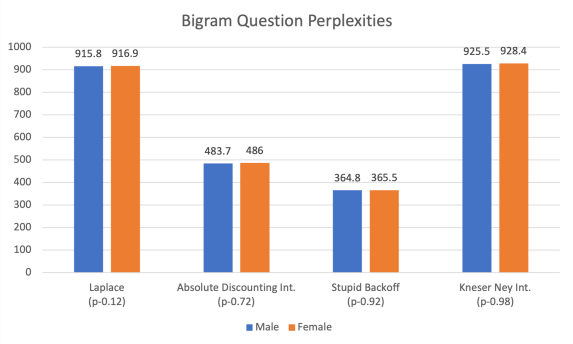

Figure 2: Perplexity results using different bigram models

4.3 Perplexity Extensions

For the first perplexity extension, we repeated the same methodology but using **trigram models** rather than bigram models. And for these more complex models, the difference between genders is still not significant. Results can be seen in Figure 3.

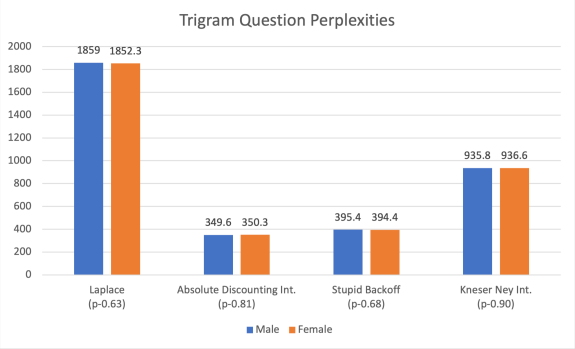

Figure 3: Perplexity results using different trigram models

These results further confirm that either perplexity doesn't seem like the best metric to measure bias, or that the questions aren't very biased. We decided to do one final try with perplexity but building an **LSTM model**.

To replicate the methodology of the paper, we trained our LSTM on the entire corpus of match commentaries, converting all words to lowercase in order to calculate perplexity most accurately. The LSTM revealed that there were no significant perplexity differences between men and women, even when comparing the top-10 ranked players in each gender as the paper did. As mentioned in the previous milestones, the paper was able to identify a significant difference in perplexity in questions asked to those players only. Having to cherry-pick certain players to prove their result should have been an indication that perplexity may not be the best metric for identifying bias, and we view these results from a more advanced model as a sign to move away from perplexity and try a different method to confirm or deny that there is a significant difference in the amount of identifiable bias in questions asked to each gender. Figure 4 below displays the small difference in perplexity identified by the LSTM, and statistical tests again showed that this difference is statistically insignificant.

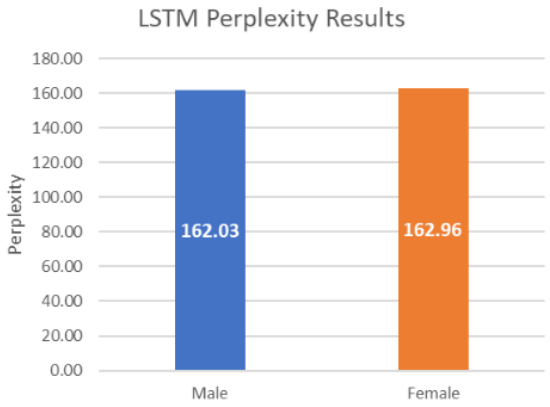

Figure 4: Perplexity comparison between male and female questions with LSTM

4.4 Non-perplexity Extensions

After performing these experiments with perplexity as a metric, which did not yield significant differences between questions asked to men and those asked to women, we decided to move on to other metrics to see if those might reveal bias better than perplexity or if those would reinforce our findings that showed a lack of bias in the interviews. We primarily experimented with topic modeling and large language models, pre-trained natural language tools which may help reveal trends better than our methods built from the ground up and trained only on the provided commentary data.

Our goal with topic modeling was to see if the topics picked out from the questions for women were significantly different from those in the questions for men. Additionally, we wanted to examine whether those topics were relatively more unrealted to the sport as compared to the topics asked to men. To perform the topic modeling, we used the sklearn Latent Dirichlet Allocation model to detect the topics and the Seaborn library for visualization. We believe this topic modeling approach can be further improved in two ways, namely (1) cleaning the dataset further by removing player names and nationality descriptions and (2) utilizing seed words to initialize topics that can separate the tennis-related words more clearly. However, in comparison to statistical tests on the perplexity methods which revealed no differences between the genders in terms of that measure, we see some notable differences in Graph 1 and Graph 2. Namely, words like felt, feeling and mental are observed more in the female corpus topic extraction. Words like victory and confidence are only viewed in the male top words. Finally, coach and dad only show up in the women's top words, indicating that women are potentially asked more questions about how others have helped or supported them. We also ran LDA on the entire questions data, and measured the prevalence of topics extracted from this overall LDA model on both genders' corpuses. This demonstrated a small but statistically significant difference in the frequency distributions of topics at $p < 0.05$. These topic modeling results indicate that more language-level analyses of bias may allow us to identify the true gender differences in this questions data set, as opposed to measuring tennis-related perplexities.

One alternative to perplexity that we thought of involved turning this problem into a classification problem. If we can use NLP to classify questions according to what gender the athlete being asked has, we can have an additional sign of gender bias. If the model can do a good job identifying the gender of the athlete (with things like gendered words and pronouns removed), then that would indicate that questions between genders are different and differentiable. If, on the other hand, the model does a poor job, and guesses near randomly, then that would indicate that questions are similar and indistinguishable, and the data isn't biased. For this task, we used large language models, as the large corpora on which they are pre-trained would likely help identify trends in our data which might exist elsewhere.

One language model we utilized for classification was an instruction-tuned version of GPT-3 known as `text-curie-001`. We selected this model because it was what we had used on Homework 5. We decided to try few-shot prompting because of its success on that assignment, providing the model with 5 examples of questions asked toward men and 5 questions asked toward women labeled with the gender of the interviewee. To avoid adding our own bias, we randomly sampled the male and female questions instead of trying to select them ourselves, alternating male and female questions in the prompt. An example prompt with one question for each gender is shown in Figure 5.

```
The second set your errors came from concentration or a
bit tired, or what do you think was the reason? : Male

Is it disappointing sometimes, Jelena, in a quarter, you
play two matches against her in the fall where you want
to test yourself that the match ends that way? : Female

{input} :
```

Figure 5: Example GPT Few-Shot Prompt

We then ran this prompt with other questions in the dataset as `{input}`, sampled at random. Unfortunately, due to limited OpenAI credit, we were unable to test all of the questions, but we were able to try about 10,000 of them. Using this prompt, GPT-3 predicted the gender of the interviewee with an accuracy of 54.4%. This indicates a lack of bias, as it demonstrates that there was little about the questions in the prompt that reveals the target of the question, even without the removal of any gender-related words such as names and pronouns.

Another language model we used as a gender classifier is BERT. We ran this model on two versions of the questions, one intact and one with gendered words such as names and pronouns removed. For this test, we fine-tuned BERT on the dataset of questions using gender as the label. We used 72% of the dataset for training, 8% for validation, and 20% for testing. The entire question was input to the model and we asked it to return the probability it was asked to each gender. For the intact questions, the model was able to correctly identify the gender of the interviewee $68 - 70\%$ of the time on the training and test datasets. However, with the gendered words removed, the model's accuracy dropped to $60 - 62\%$. This $\sim 10\%$ drop demonstrates that the gendered words were signifi-

cantly helping BERT identify the target of certain questions.

4.5 Analysis of Results

As seen in the figures above, our intial experiments from Milestones 1 and 2 did not reveal significant perplexity differences in questions asked to the two genders, meaning that the content of each group of questions was equally perplexed relative tot he commentary data. This indicates to us that questions asked toward women were not significantly more biased or unrelated to tennis as compared to the questions asked to men. Additionally, the large language models could only correctly predict the gender of the interviewed player with $50 - 60\%$ accuracy, demonstrating that there were no broad trends across the data which the models could take advantage of to identify the interviewee with high accuracy. However, we found it interesting that the topic modeling did reveal certain notable differences across the top words in each corpus. We are curious if GPT-3 could have picked up on these if provided more (e.g. 1000+) randomly sampled examples of questions from each gender, potentially increasing its classification accuracy. The results of the BERT experiment show shows that, while these trends may have been the cause of its improved performance, they were clearly not significant enough for 75%+ accuracy without using gender-associated words as a crutch.

5 Conclusions and Limitations

From our analysis, we conclude that there is not a significant difference in the amount of bias contained in interview questions when comparing interviewees of either gender, at least overall or on average. This matches the paper's initial findings, as they too did not detect a significant difference in bias except when examining 'uncommon' questions or the top-10 ranked players in their respective competitions. We believe that the perplexity difference in 'uncommon' questions could be explained by the fact that the terminology in less common questions asked to women tends to be slightly less sport or activity related than the terminology in questions asked to men, and the difference in the top-10 players could be due to the fact that many high-profile female athletes tend to be sponsored by or model for various companies, so they may be asked about those deals which are less common for men. Female celebrities also tend to be more public about their families than male celebrities, so interviewers may tend to ask questions about that to women more than they ask men. In short, we conclude that there is no significant overall gender bias in the way interviewers pose questions to women, but there may be significant differences for certain players due to external societal factors.

6 Acknowledgements

We would like to thank our professor Dr. Mark Yatskar and our TA Yu Feng for guiding us throughout this nontraditional project. We would also like to acknowledge Dr. Chris Callison-Burch for advising us on potential paths forward using large language models for Milestone 4.

References

Laura L. Aull and David West Brown. 2013. Fighting words: a corpus analysis of gender representations in sports reportage. *Corpora*, 8(1):27–52.

Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying gender bias in different corpora.

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism.

Sophia Harrison, Eleonora Gualdoni, and Gemma Boleda. 2023. Run like a girl! sport-related gender bias in language and vision. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14093–14103, Toronto, Canada. Association for Computational Linguistics.

Joshua R. Minot, Nicholas Cheney, Marc Maier, Danne C. Elbers, Christopher M. Danforth, and Peter Sheridan Dodds. 2021. Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance.

Isabel Pereira Fernandez. 2021. Investigating gender stereotypes in the media : A natural language processing approach to understanding gender disparities in the reporting of football. Master's thesis, Linköping UniversityLinköping University, The Institute for Analytical Sociology, IAS, Faculty of Arts and Sciences.

Prashanth Rao and Maite Taboada. 2021. Gender bias in the news: A scalable topic modelling and visualization framework. *Frontiers in Artificial Intelligence*, 4.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review.

A Appendix

A.1 EDA graphs

Figure 6: Word cloud of disproportionately occurring words by gender

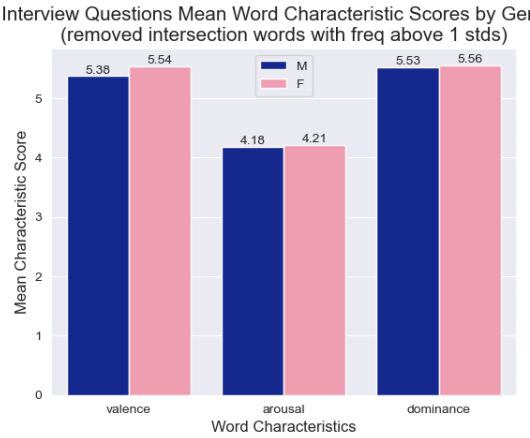

Figure 7: Word characteristics of male and female athletes

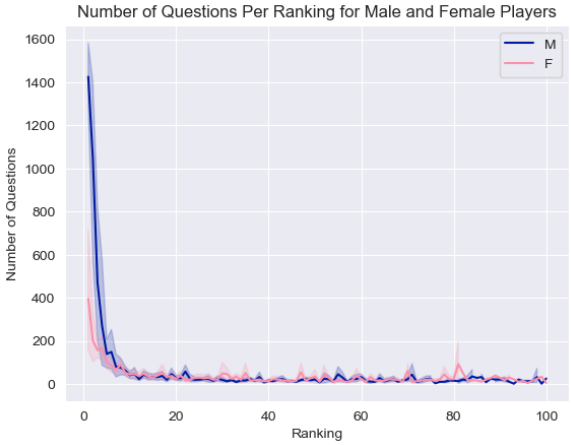

Figure 8: Number of questions per player rank