



CIS 5300 Natural Language Processing

# Gender Bias in Sports Dialogue

Federico Cimini, Ryan Martin, Khushi Shelat, Kyle Sullivan

# Agenda

Introduction & Motivation

Baseline Paper

EDA

Simple Baseline

Complex Baseline

Perplexity Analysis

Project Extensions

Conclusion

**‘Novak Djokovic is definitely the most motivated – no one can compete with his self-discipline’**

TENNIS NEWS



**Carlos Alcaraz honoured  
Tour's Sportsmanship**

ATP TOUR



**Leading coach thinks there are ‘two explanations’ for Rafael Nadal deciding to play Australian Open**

TENNIS NEWS



## Naomi Osaka gets brutally honest about becoming a mother

TENNIS NEWS



James Rickert/Getty Images

Sport > Tennis

## TENNIS CONTROVERSY US Open drama as unusual outfit change causes huge rift during women's quarterfinal clash

Terence Scott

Published: 19:33 ET, Sep 6 2023 | Updated: 8:06 ET, Sep 7 2023

EXPLAINERS SPORTS CULTURE

## Naomi Williams's US Open fight with umpire Ramos, explained

Players have been celebrated for snapping at umpires. Serena Williams is punished for it.

by alex@vax.com | Sep 10, 2018, 12:50pm EDT

Money Tech Motors


placing Golf Basketball Football

SHARE



the 2018 US Open final. | Julian Finney/Getty Images





Determine quantitatively if  
there is gender bias present in  
journalist questions to tennis  
athletes

PROJECT OBJECTIVE

Minot, Joshua R., et al. "Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance." arXiv preprint arXiv:2103.05841 (2021).

Babaeianjelodar, Marzieh, et al. "Quantifying gender bias in different corpora." Companion Proceedings of the Web Conference 2020. 2020.

Rao, Prashanth, and Maite Taboada. "Gender bias in the news: A scalable topic modelling and visualization framework." Frontiers in Artificial Intelligence 4 (2021): 664737.

# BASELINE PAPER

SUMMARY AND RESULTS

# Tie-breaker: using language models to quantify gender bias in sports journalism

Liye Fu and Cristian

Danescu-Niculescu-Mizil and Lillian Lee

2016



# PAPER SUMMARY

## COMMENTARY

1

### DATA

A gender-balanced set of 3000+ commentaries

2

## BUILD LANGUAGE MODEL ON COMMENTARY

Bigram Language Model built on all commentary data

3

## INTERVIEW

### DATA

Questions and answers from 6400+ interviews (total 80k questions)

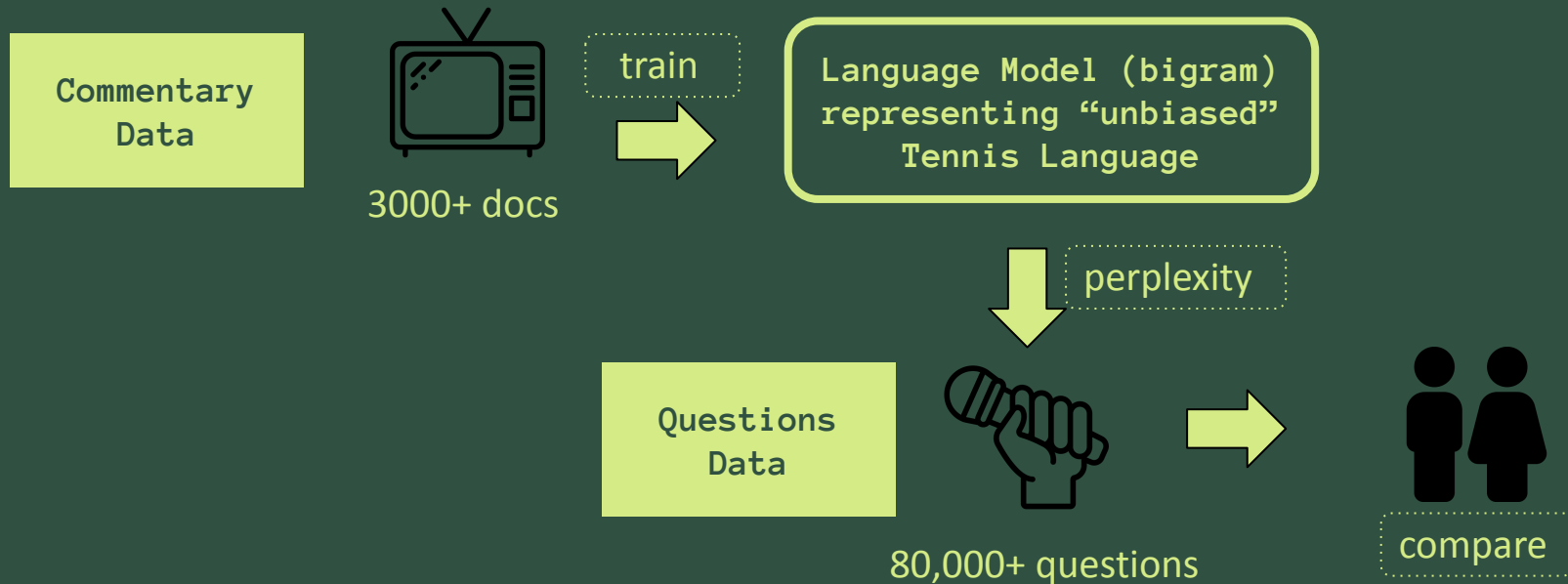
4

## PERPLEXITY OF LM ON QUESTIONS DATA

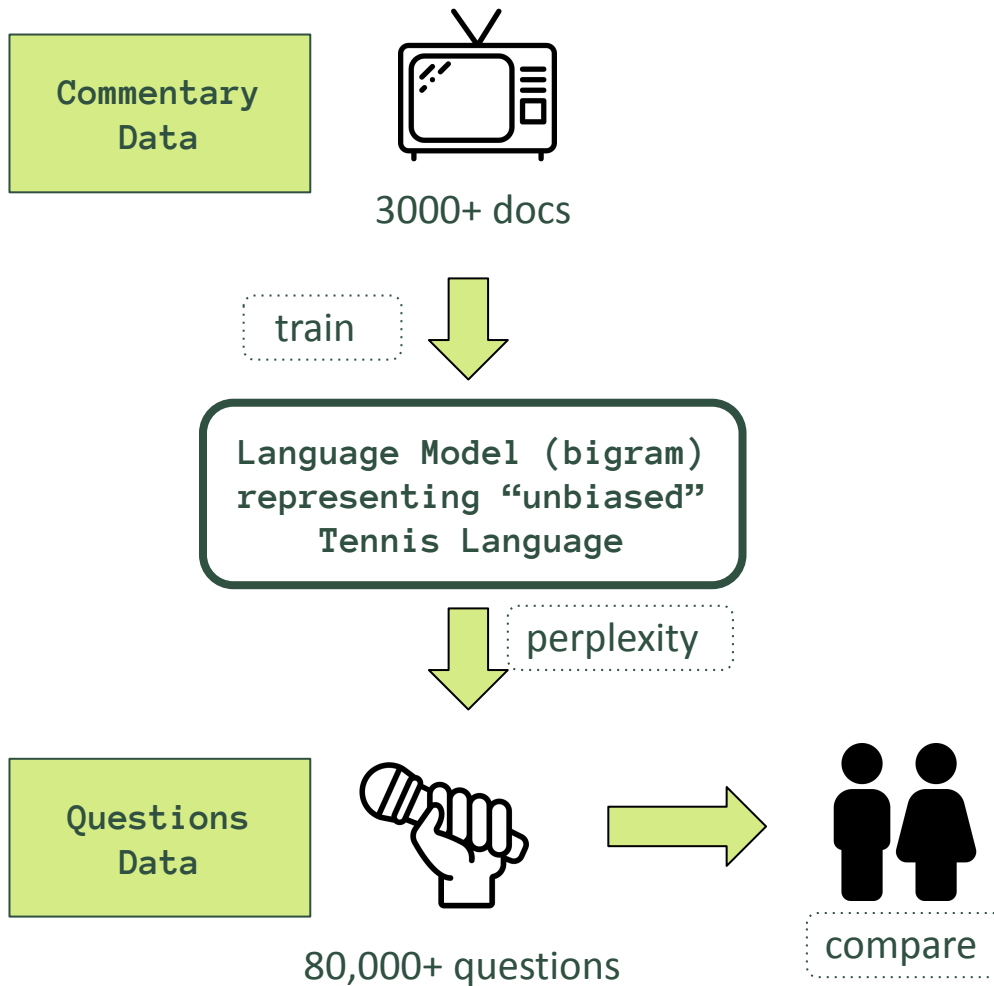
Hypothesis is that this can show if question language isn't tennis related

Tie-breaker: using language models to quantify gender bias in sports journalism

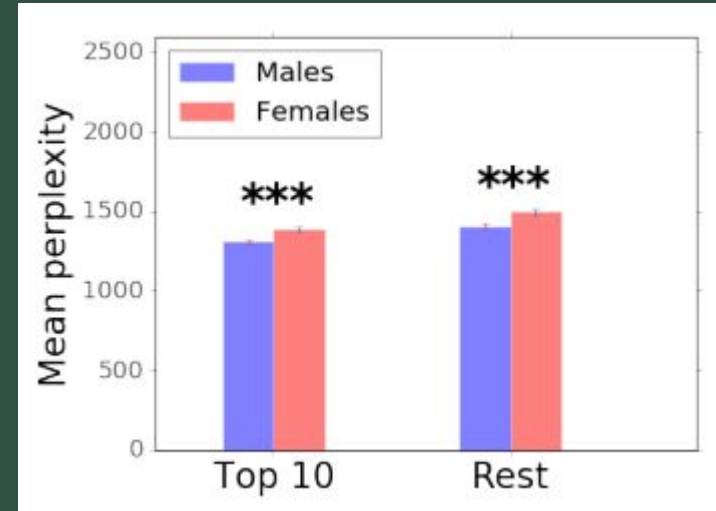
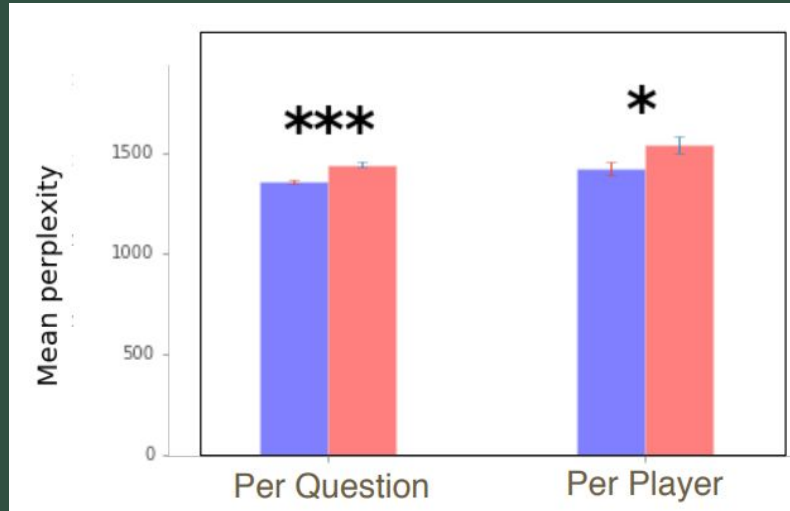
# PAPER SUMMARY



Tie-breaker: using language models to quantify gender bias in sports journalism



# FINDINGS



Tie-breaker: using language models to quantify gender bias in sports journalism

## PERPLEXITY APPROACH

Is this approach replicable  
with other, more modern  
language models?

## OTHER APPROACHES

Can we measure gender bias  
through other, more reliable  
methods?



# EDA

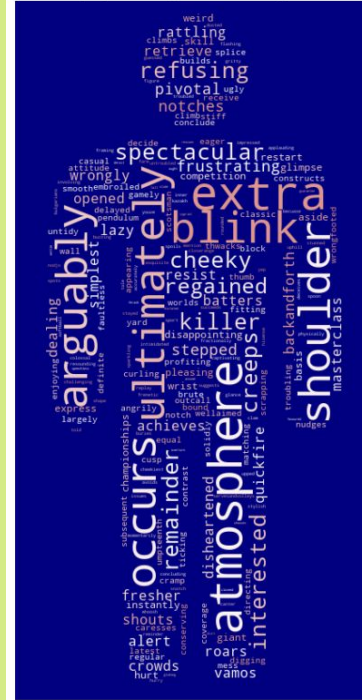
COMMENTARY AND QUESTIONS DATA

# COMMENTARY DATA

While the paper argued that this data was an unbiased representation of tennis vocabulary, we found that there are differences between men and women matches (especially for uncommon words)

## Commentary Word Clouds by Gender (removed intersection words with freq above -3 stds)

Disproportionately Male-Occurring Words



Disproportionately Female-Occurring Words



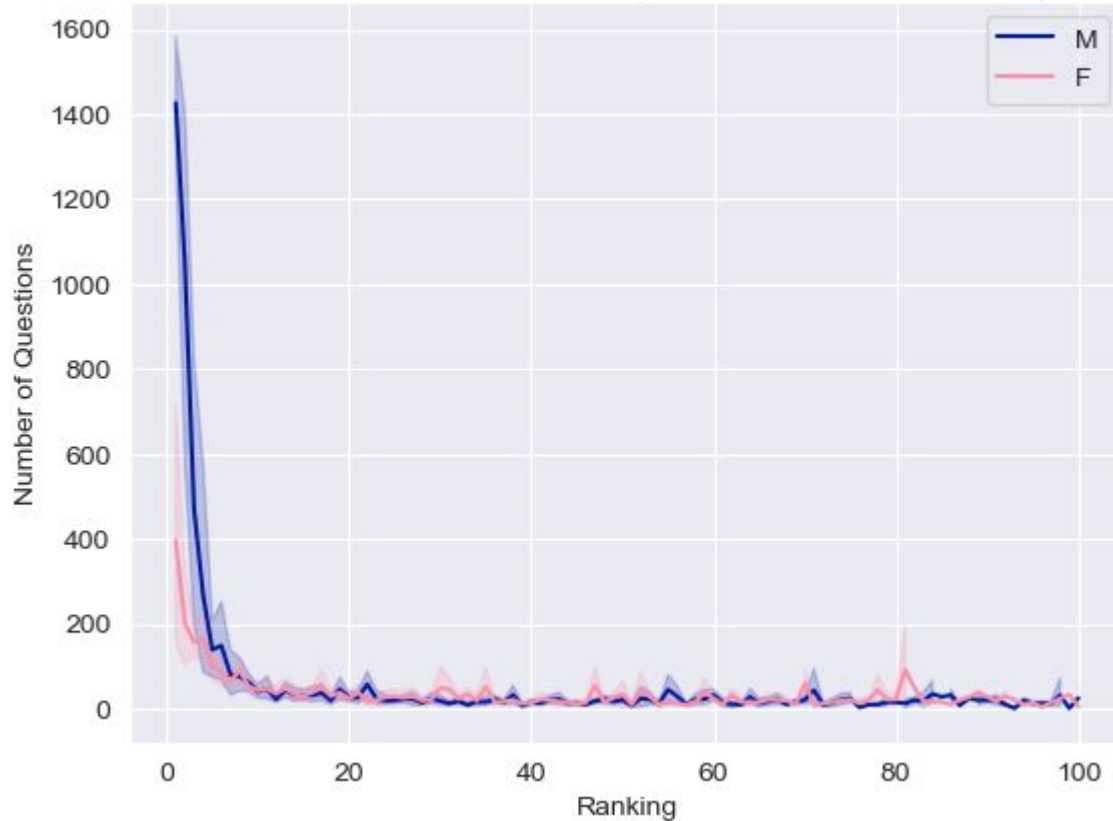
# QUESTIONS DATA

Disproportionately Female-Occurring Words



For the questions dataset, we see that the differences in gender are very prominent, especially for uncommon words. Female players get questions that on the surface appear to be more gender-biased.

Number of Questions Per Ranking for Male and Female Players



Top male athletes had more questions on the dataset compared to top ranked female players, while lower ranked female athletes had more questions than their male counterparts.

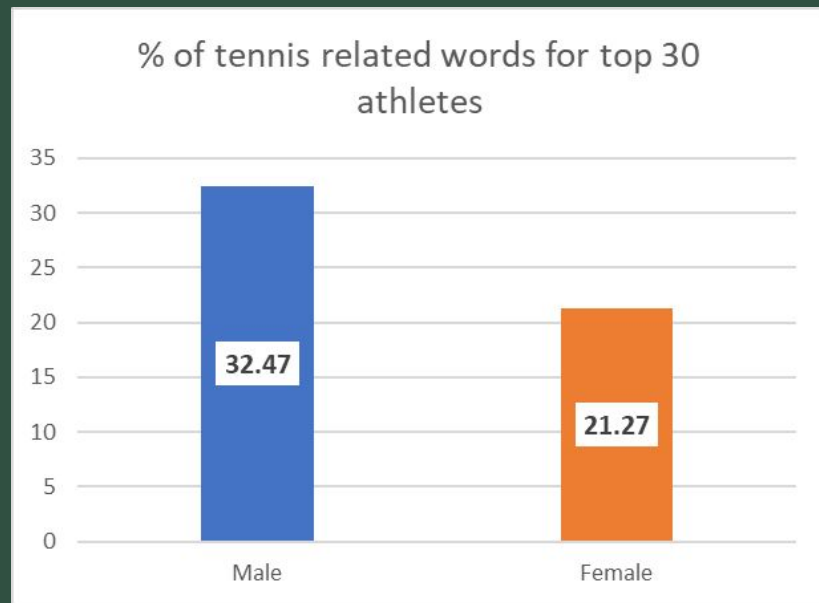
# BUILDING BASELINES

MEASURING RESULTS



# SIMPLE BASELINE

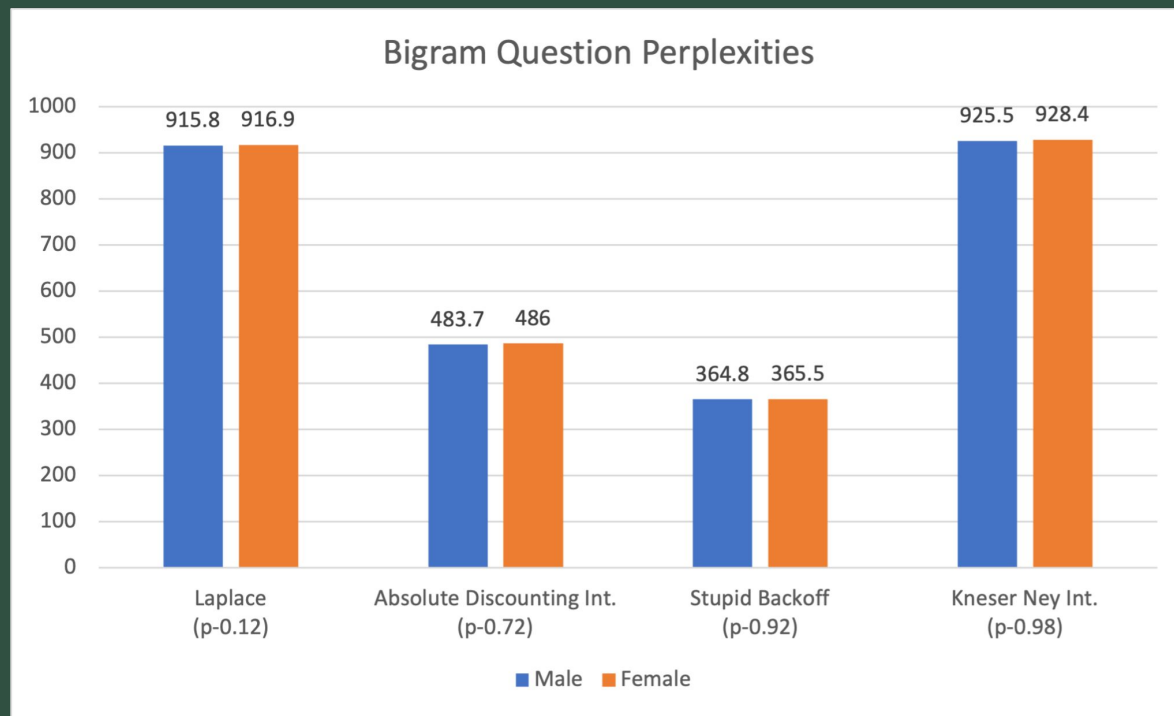
For our simple baseline, we asked Chat GPT for tennis related words, and measured how often they appeared in male vs female questions



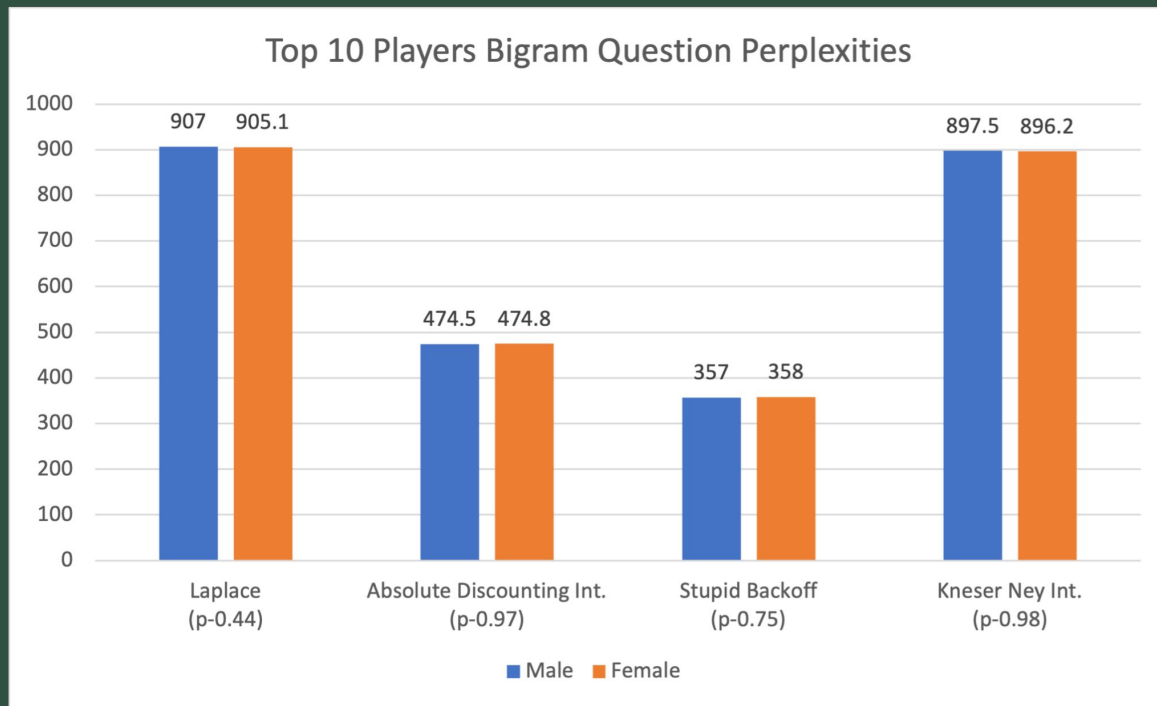
# COMPLEX BASELINE

For our complex baseline, we  
replicated the methodology of the  
paper with different bigram and  
trigram models

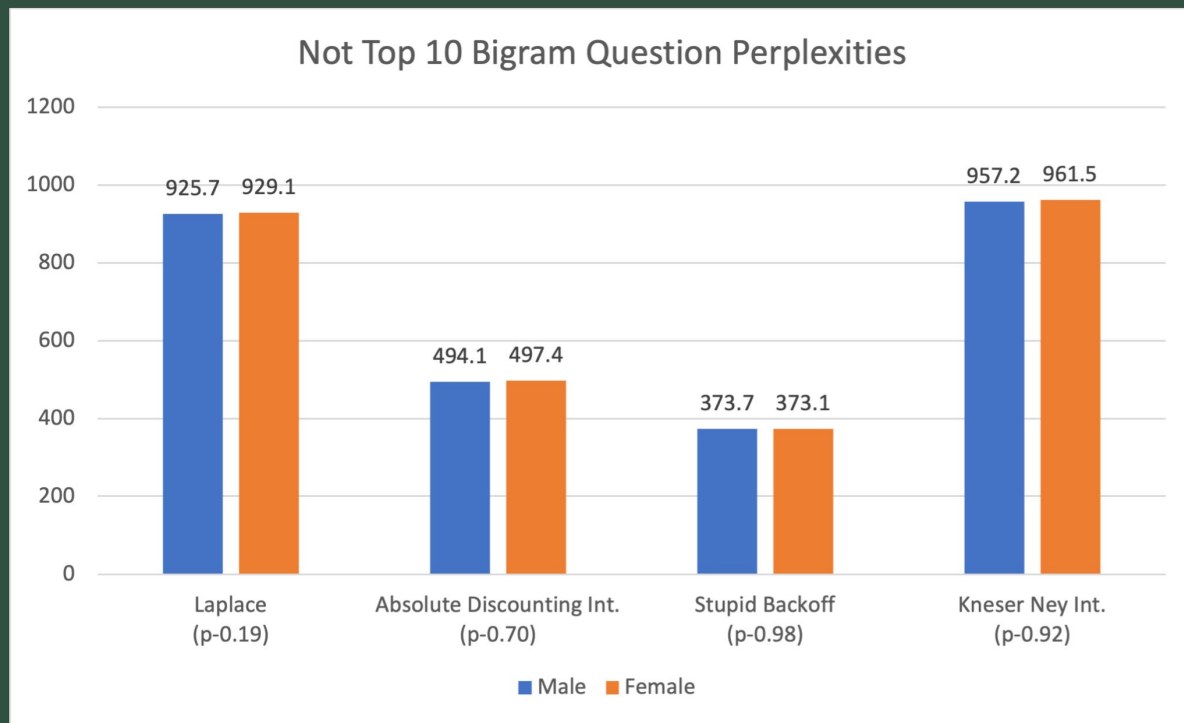
# BIGRAMS COMPARISON – All Players



# BIGRAMS COMPARISON – Top 10 Players



# BIGRAMS COMPARISON – Other Players





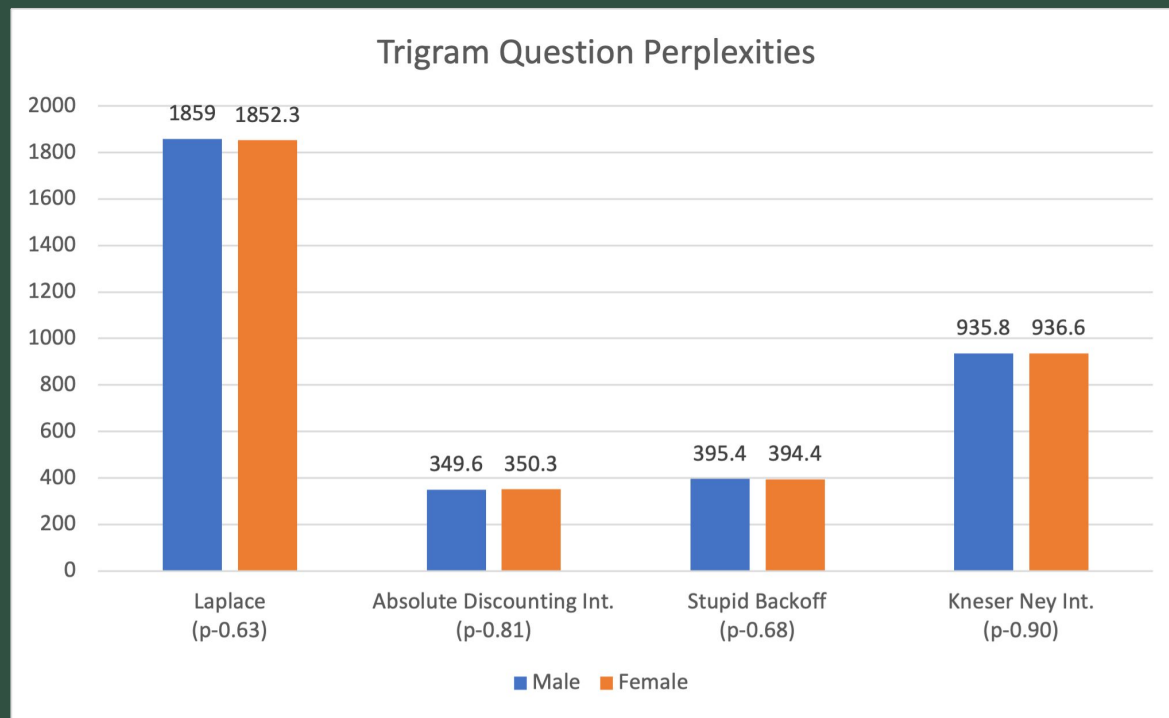
# BASELINE ANALYSIS

- Simple baseline supports paper statement.
- Conflicting results for complex baseline:
  - Perplexity for lower ranked players is higher than of top 10 players, as shown in the paper.
  - But, between men and women, there is not a significant difference in perplexities.
- Perplexity is limited as a measure of bias.

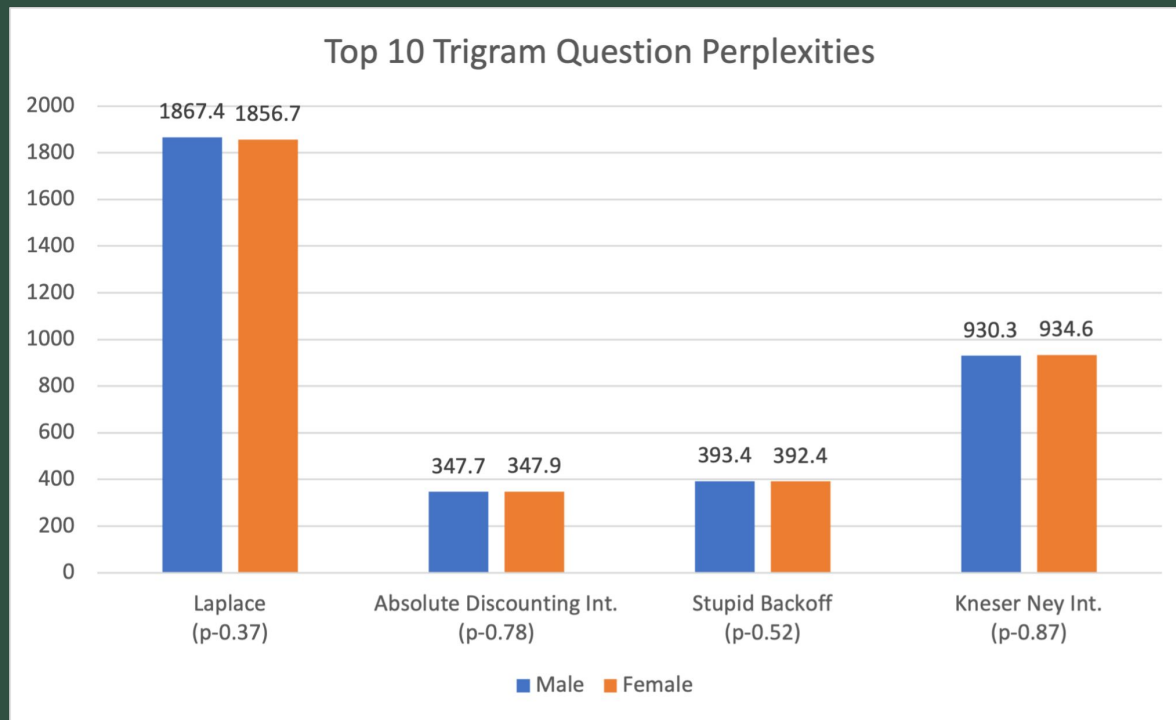
# EXPERIMENTATION

MEASURING RESULTS

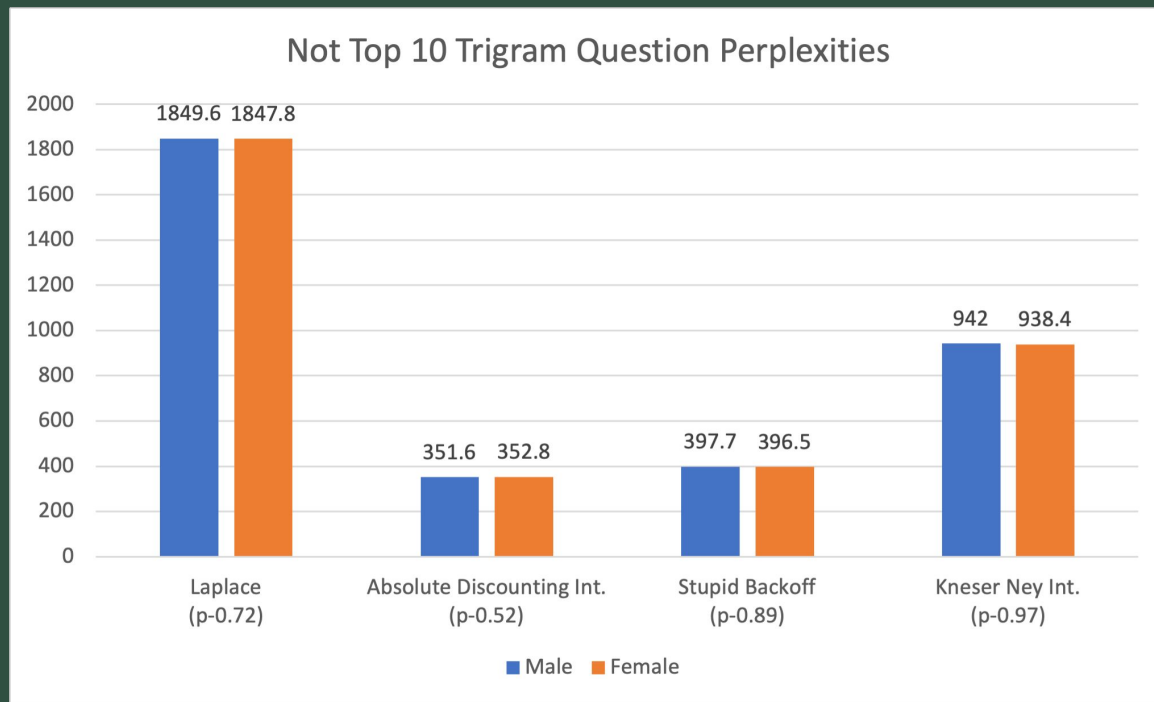
# TRIGRAM RESULTS



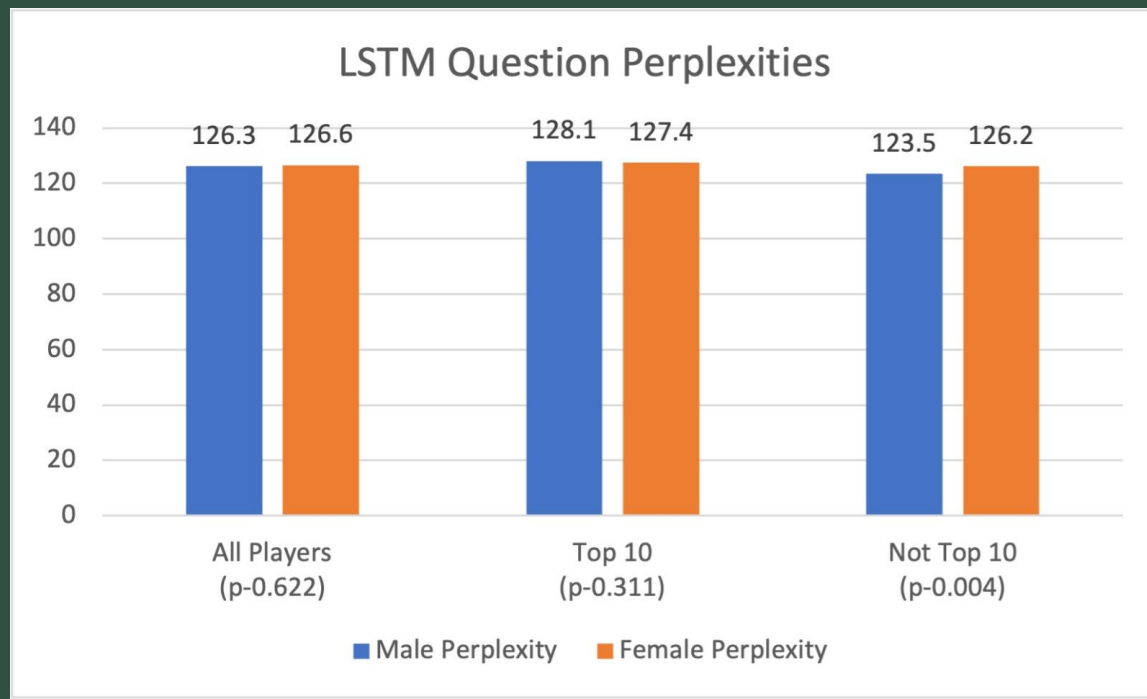
# TRIGRAM RESULTS – Top 10



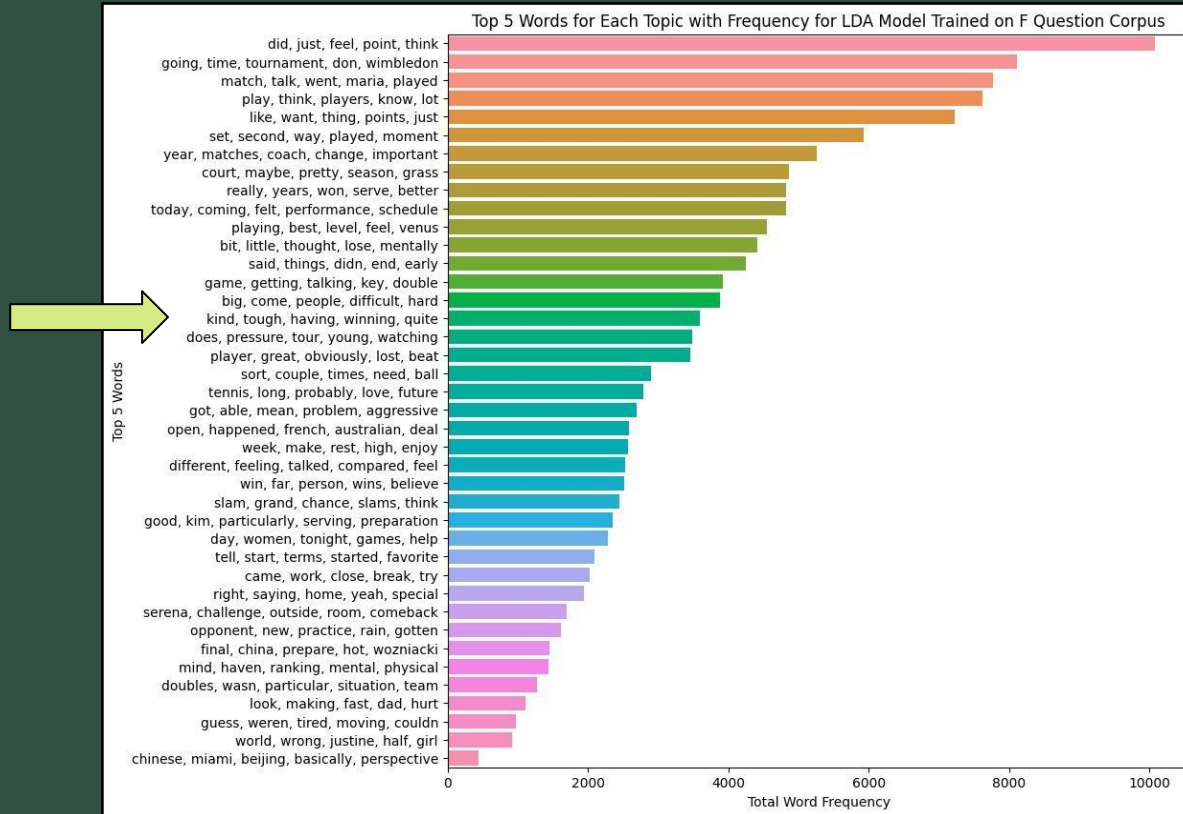
# TRIGRAM RESULTS – Others



# LSTM RESULTS

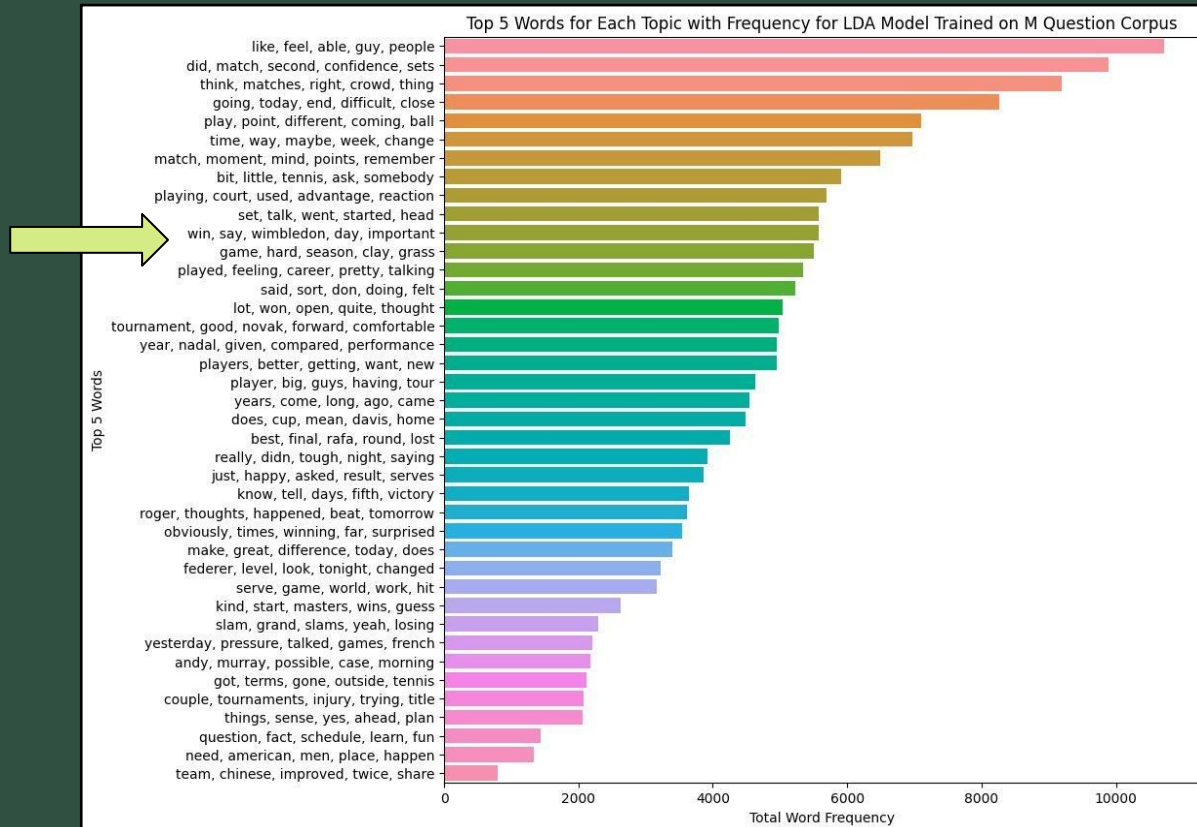


# TOPIC MODELING – WOMEN

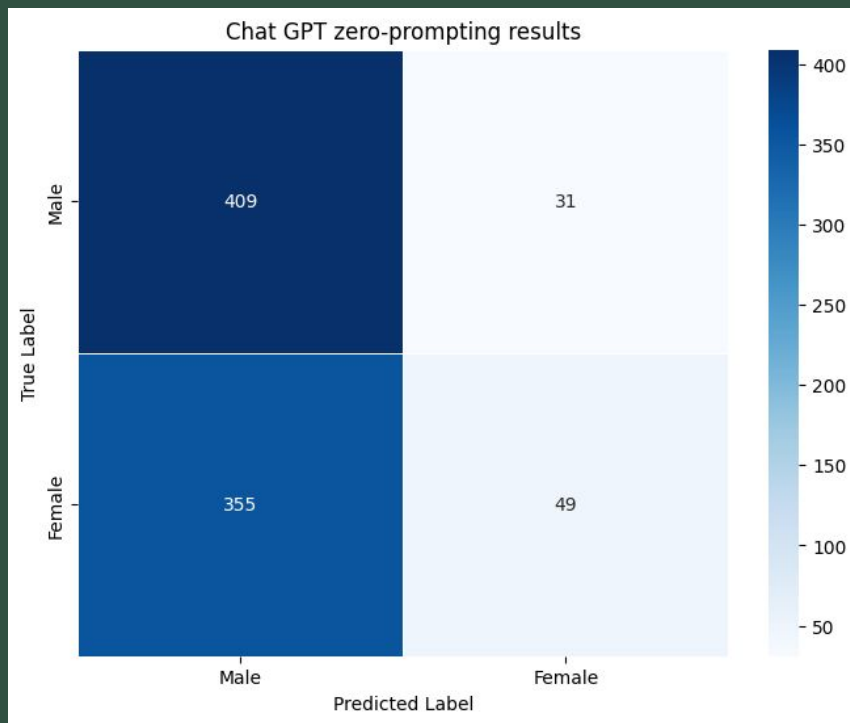




# TOPIC MODELING – MEN



# GPT PROMPTING



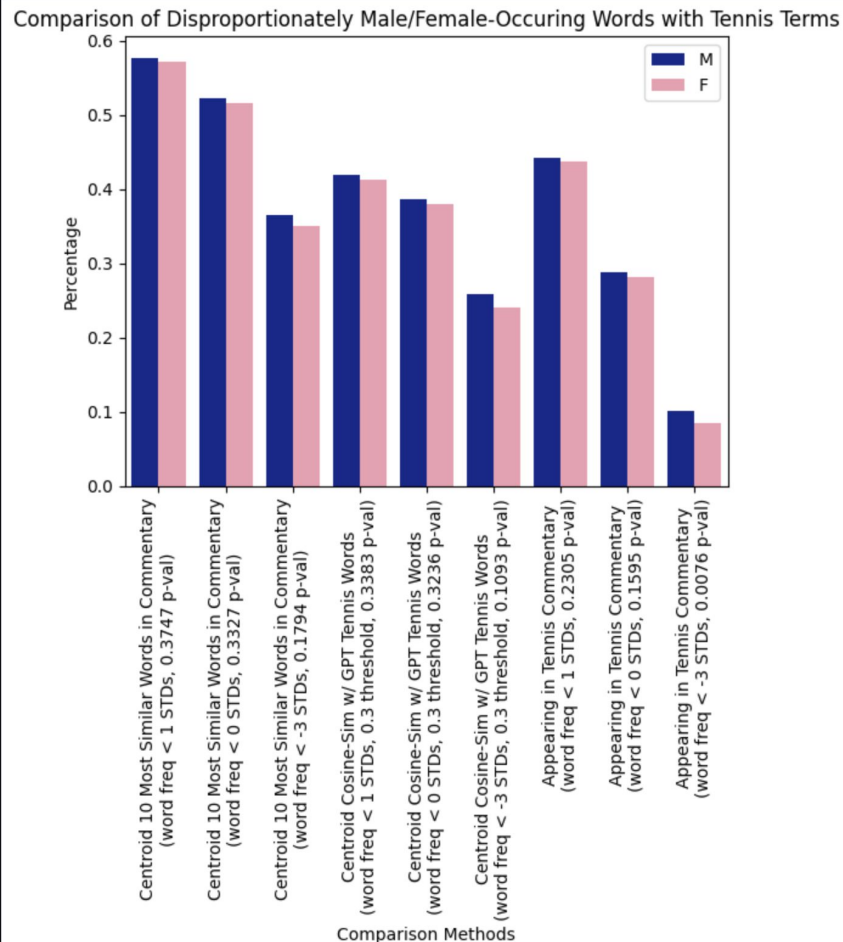
Correct: 5168/9485

Accuracy: 54.5%

# BERT and SportsBERT – CLASSIFICATION

Model	Dataset	Accuracy
BERT Classifier (M/F)	With Gender Pronouns	68.79%
BERT Classifier (M/F)	No Gender Pronouns	62.24%
BERT Classifier (Win/Loss)	With Gender Pronouns	74.58%
SportsBERT Classifier (M/F)	With Gender Pronouns	68.27%
SportsBERT Classifier (M/F)	No Gender Pronouns	61.73%
SportsBERT Classifier (Win/Loss)	With Gender Pronouns	73.31%

# VECTORIZATION RESULTS



# RESULTS AND CONCLUSIONS

MEASURING RESULTS

## PERPLEXITY

- We challenged perplexity as a measure for bias
- We suggest Cosine Similarity and BERT Classification as potential measures to quantify bias further

## BIAS

- Gender bias exists in tennis journalism, but modelling it without tennis-specific domain knowledge is a challenge.
- Gender bias is most apparent when comparing the *least common words* asked to both genders.

# Limitations & Next Steps

## Key Limitations

- Questions data are very specific to the tennis context.
- Commentary data and chatGPT-generated tennis terms were the only tennis-context related data we had.

## Next Steps / Future Extensions

- Building a fine-tuned BERT model specific to the tennis context; sportsBERT model to not increase accuracy significantly
- Topic modelling with a method that has the ability to allocate tennis-related topics