Soccer Talk: Exploring Differences in r/soccer Comments from Men's and Women's World Cup Match Threads

Kaitlin Swinnerton

Background

- In the United States, women athletes experience bias and microaggressions that can cause negative biological, cognitive, and behavioral effects (Kaskan et al, 2014)
- Professional female athletes trail behind their male counterparts in income and working conditions
    - Serena Williams is the only woman on Forbes' 2019 list of the 100 highest paid athletes
    - The USWNT is currently suing U.S. Soccer for gender discrimination in the workplace

Female economists face similar challenges

Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum

Alice H. Wu*

- Among academics studying economics, women are underrepresented and face difficulty rising up the ranks in their careers
- A 2017 analysis by Alice Wu of comments on the Economics Job Market Rumors Forum found evidence of sexism directed to female economists discussed on the site

Can similar evidence of gender stereotyping be found in online discussions about sports?

- Hypothesis: Online forums discussing soccer will exhibit gender stereotyping, bias, and microaggressions towards female athletes.

Methods

- Scraped comments from reddit.com/r/soccer match threads from the 2018 FIFA World Cup and the 2019 FIFA Women's World Cup
- Preprocessed to remove names of players and countries participating in the tournaments
- Replaced gendered pronouns (he/him/his, she/her/hers, etc) with gender neutral terms
- Built models to classify comments by tournament

Model 1: Mulitnomial Naive Bayes

- Used sklearn's CountVectorizer function to transform comments into a matrix of token counts
- Vocabulary size of the training data was 10,818
- Hyperparameters were tuned with grid search cross validation
- **Accuracy on the test set was 66.86%**
- Same process was repeated using bigrams and trigrams, but the model performance was worse

Model 1: Mulitnomial Naive Bayes

Feature	Men's World Cup	Women's World Cup	coef_diff
thai	-11.429555	-8.276627	3.152928
hd	-11.429555	-8.367599	3.061956
norwegian	-11.429555	-8.367599	3.061956
band	-11.429555	-8.416389	3.013166
bein	-11.429555	-8.467682	2.961872
graham	-11.429555	-8.521749	2.907805
hara	-11.429555	-8.578908	2.850647
vd	-11.429555	-8.578908	2.850647
illegal	-11.429555	-8.578908	2.850647
onside	-11.429555	-8.639532	2.790022
supersport	-10.330942	-7.583480	2.747463
olympics	-11.429555	-8.704071	2.725484
x200b	-11.429555	-8.704071	2.725484
themself	-11.429555	-8.704071	2.725484
racism	-11.429555	-8.773064	2.656491
wwc	-11.429555	-8.773064	2.656491
chilean	-11.429555	-8.847172	2.582383
female	-10.736408	-8.193245	2.543162
offside	-9.126970	-6.616331	2.510639
orange	-11.429555	-8.927215	2.502340

Feature	Men's World Cup	Women's World Cup	coef_diff
russian	-7.270672	-10.718974	-3.448302
vuvuzelas	-8.028357	-11.412121	-3.383764
iranian	-8.133718	-11.412121	-3.278403
mexicans	-8.171458	-11.412121	-3.240663
peruvian	-8.251501	-11.412121	-3.160620
kdb	-8.294061	-11.412121	-3.118061
croatian	-8.338512	-11.412121	-3.073609
itv	-8.385032	-11.412121	-3.027089
2010	-8.385032	-11.412121	-3.027089
colombian	-8.433822	-11.412121	-2.978299
doot	-8.539183	-11.412121	-2.872938
belgian	-8.539183	-11.412121	-2.872938
russians	-8.596341	-11.412121	-2.815780
putin	-8.656966	-11.412121	-2.755155
subasic	-8.721505	-11.412121	-2.690617
honk	-8.721505	-11.412121	-2.690617
ba	-8.790497	-11.412121	-2.621624
bus	-8.133718	-10.718974	-2.585256
2014	-8.171458	-10.718974	-2.547516
columbia	-8.864605	-11.412121	-2.547516

Model 1: Mulitnomial Naive Bayes

```
Top five most poorly classified messages

Message 1:
Actual Label: 0,  Predicted Label: [1]
R ratio: 4081.78

this is either name brilliant tackle or clear penalty depending on the poster , which is why soccer name var in socce
r is always going to be controversial name many calls are name to the refs name made in the context name what has hap
pened in the game already


Message 2:
Actual Label: 1,  Predicted Label: [0]
R ratio: 1716.19

we keep avoiding the germans , 2010 world cup 2014 world cup now 2019


Message 3:
Actual Label: 0,  Predicted Label: [1]
R ratio: 801.15

country name country should play 3 on 3 ot but first give them sticks to hit the ball , then put them on ice with ska
tes , then change ball to name puck


Message 4:
Actual Label: 0,  Predicted Label: [1]
R ratio: 409.72

motion to make yellow country 's primary kit colour it looks name much nicer yes , i know they 're called the red dev
ils but the yellow with the black shorts look like it would make name better first choice kit


Message 5:
Actual Label: 0,  Predicted Label: [1]
R ratio: 344.95

surely on the line doesn t mean outside the box for keepers \? i don t know how the referee could be sure about that
one
```

Model 2: Recurrent Neural Network

- Recurrent neural network with an embedding layer, recurrent layer, a fully connected layer, a dropout layer, and an output layer
- Experiments were run to test the ideal number of neurons and batch size
- The optimal model yielded a test accuracy of 65%, failing to outperform the Naive Bayes model.

Discussion and next steps

- Models were able to predict gender of the tournament at a rate greater than chance
- Investigating the most gender specific words and the most poorly classified comments suggests shows some interesting patterns
- More thorough text cleaning is necessary to draw strong conclusions
- Increasing sample size and looking at a wider variety of discussion types could provide more insight
- Next steps: topic analysis, analyses of specific male/female athletes