# Why do People Care about Sports?

An analysis of reddit posts

Graham Taylor

# Our Goal (goooollllllllll?):

- To gain insight into what drives engagement on reddit across sports of all kinds (team, individual, hobby)
  - *Success: Significant improvement over the baseline in predicting whether a post on reddit will exceed the median number of comments*
- To use this insight to better understand the common threads among sports that people like so much
  - *Success: Finding specific aspects of sport (word-related) that are generally important in predicting engagement*

# Why do this?

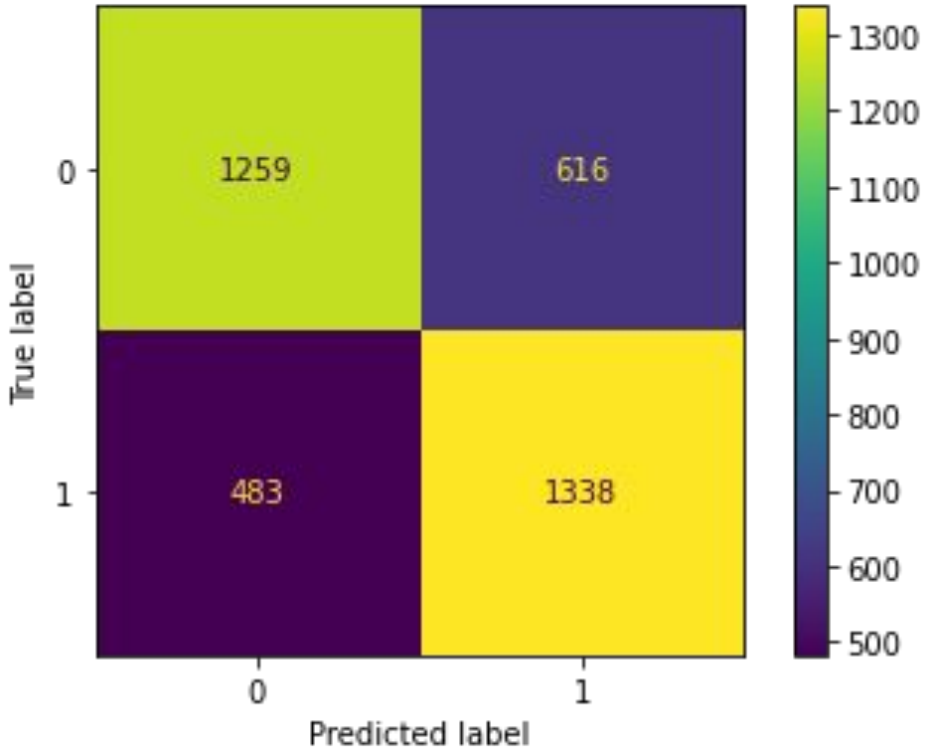
- Sport affects Government:
  - Freakonomics Radio ep. 506: Sportswashing
  - Soft power through sport: Qatar World Cup, LIV Golf
- Small business, big interest
  - All sports combined as large as Johnson & Johnson
  - Yet sport a section in every major newspaper
- Evangelism
  - If you read “Why?!?”: Exploring why sport appeals to others can change understanding, appreciation

# Methods:

- Gather up to 1000 recent posts from 20 most popular sports-related subreddits via Reddit API
- Count-vectorize text using English “stop words”
- Two Models:
  - Logistic Regression
  - Random Forest Classifier
- Evaluate Accuracy
- Identify Important Words

# Best Model (RFC): Metrics

- Accuracy: 0.703
- Precision: 0.685
  - Reliability of Pos. Preds
- Recall: 0.735
  - ID True Pos.
- Specificity: 0.671
  - ID True Neg.



# Best Model (RFC): Useful Predictors

- *Title descriptiveness* increases engagement (title_len, age, has_authtext, has_linktext)

- *Titles designed to encourage engagement* do so (thread, discussion)

- *Titles related to current teams* generate engagement (team, 2022)

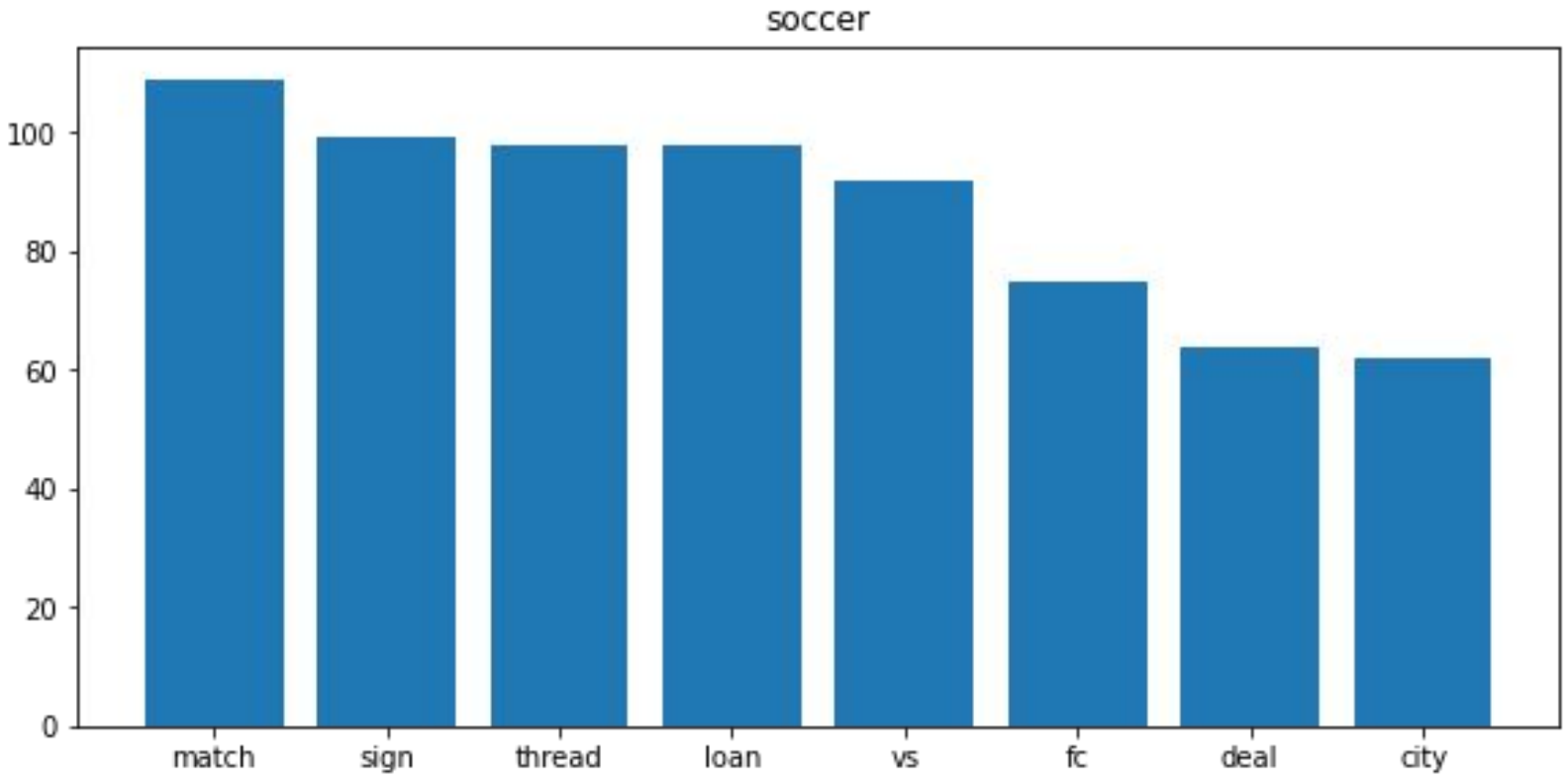
- *Subreddit* matters a lot.

First question partially answered, but not general question.

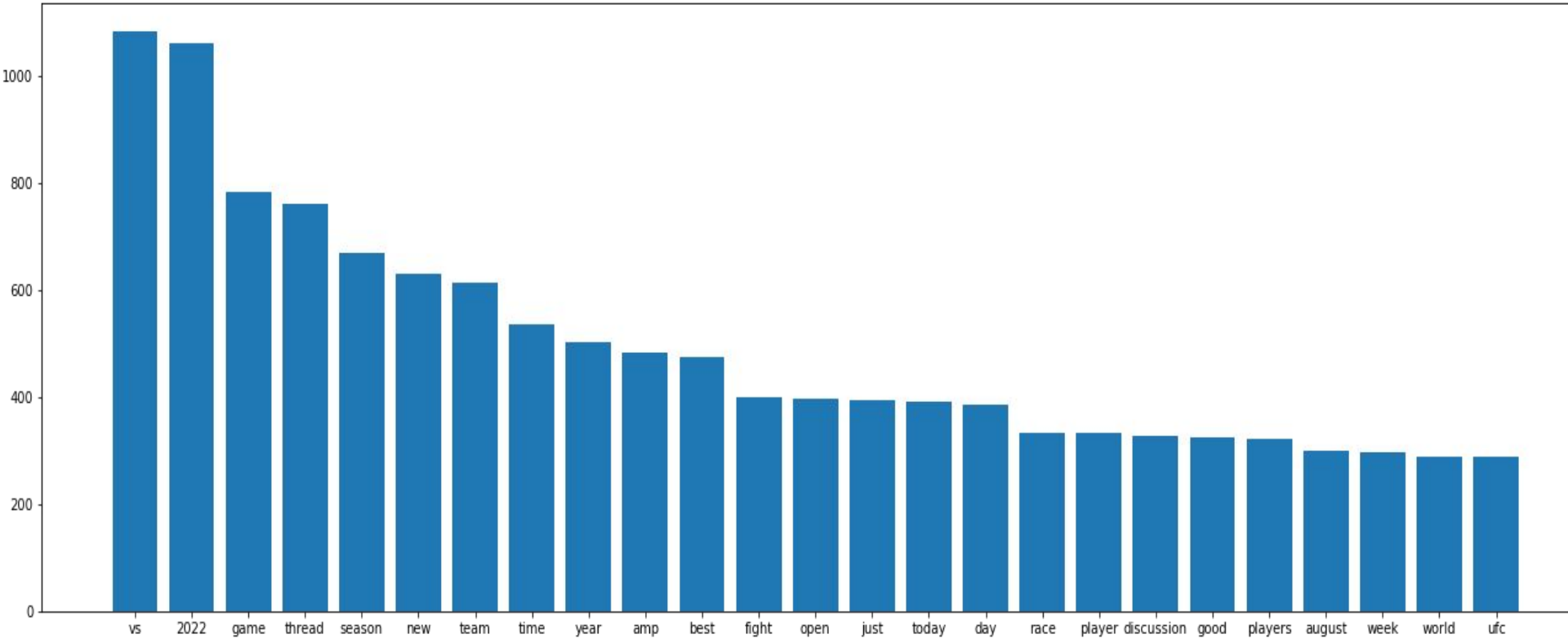
# Model Wrap-up

- With 70% accuracy, the best model was fine.
- Challenge: Commenting behavior differs significantly across subreddits
  - Nature of Sport
  - Moderation / Rules
- How are people really engaging?
  - Title word frequencies might yield more insight into “engagement” and why people care about sports

# Soccer: News and Rumors



# Common Words Overall

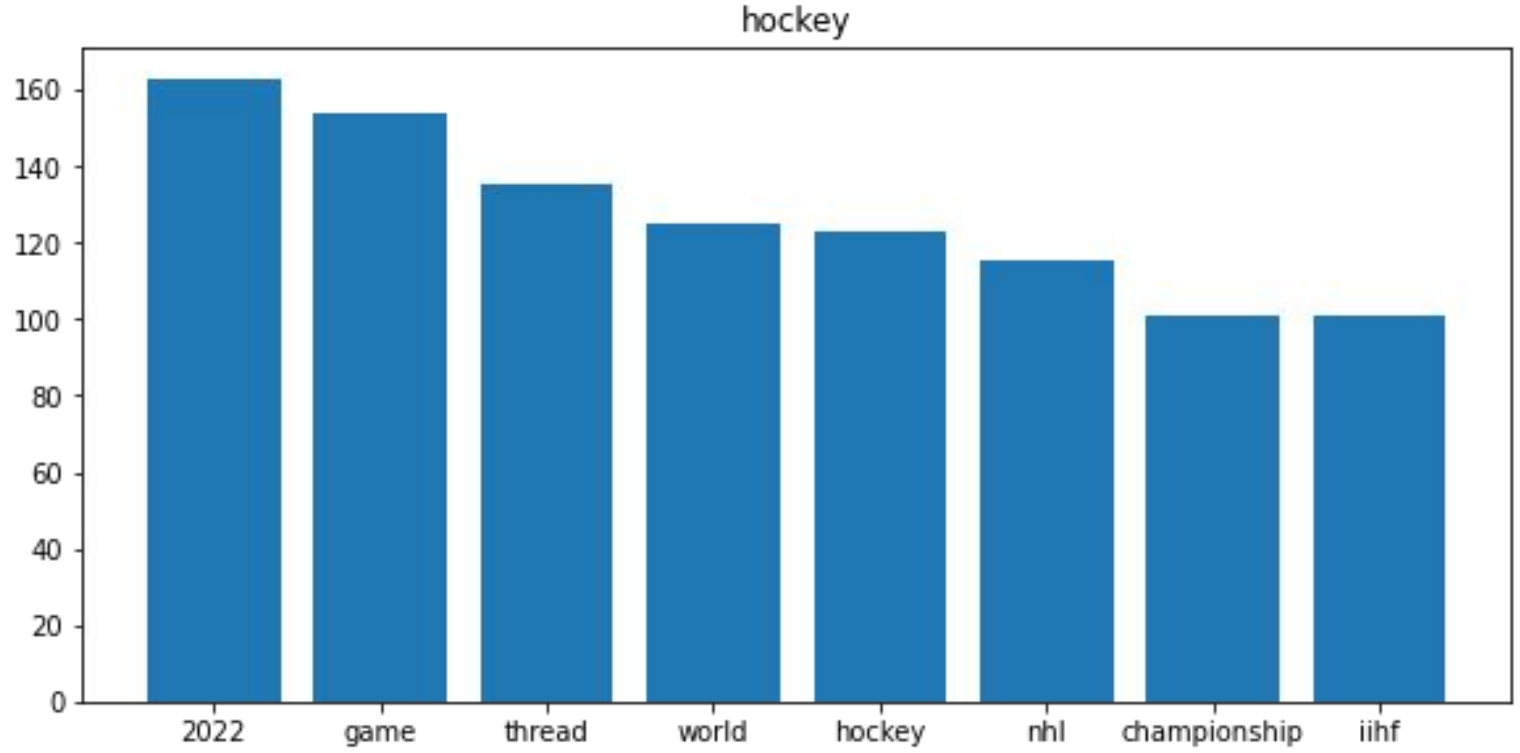


# Insights from Exploratory Data Analysis:

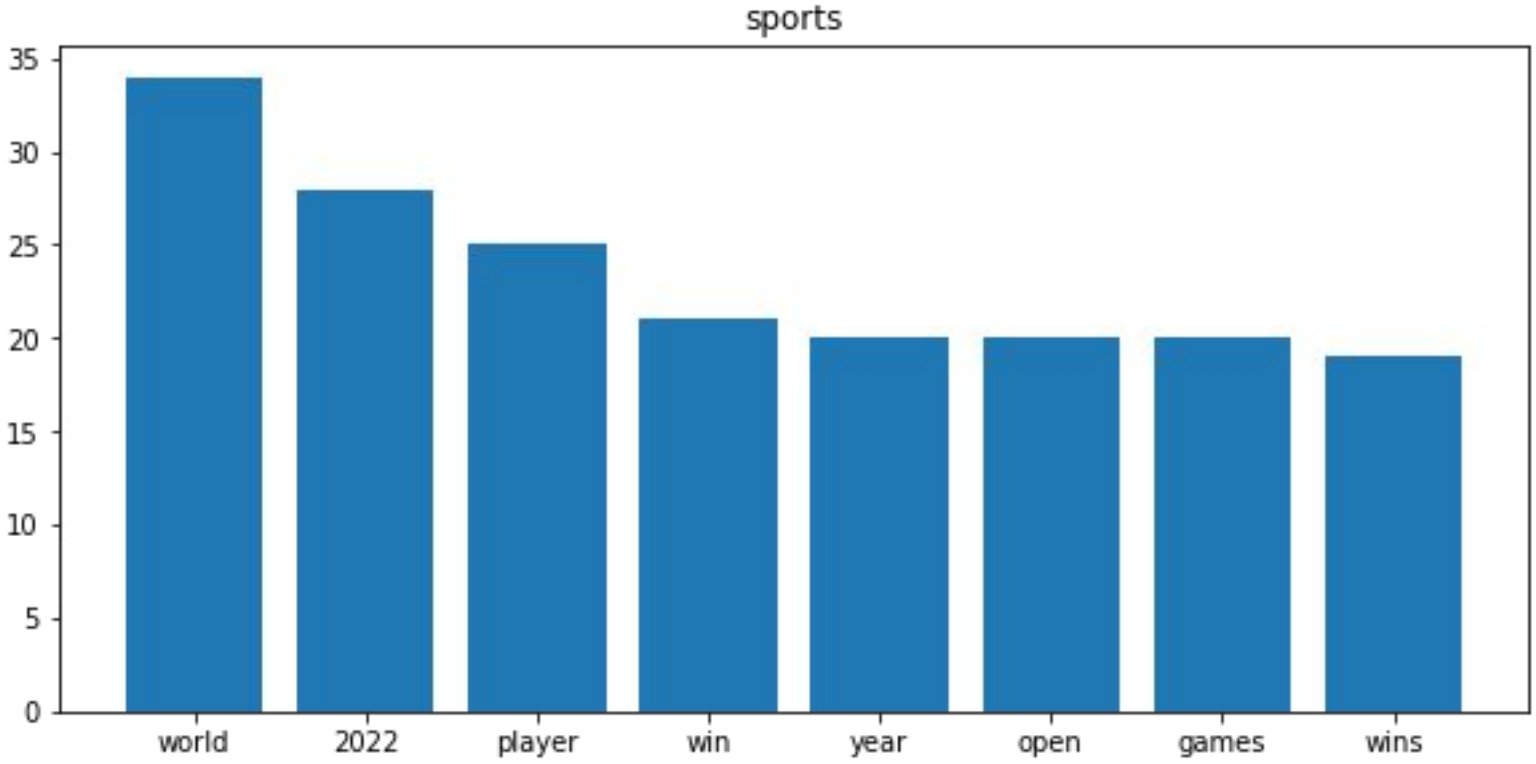
Title word frequency patterns:

- Competition words (vs, game, season)
- Time words (2022, new, year, today, august, week)
- Discussion words (thread, discussion)

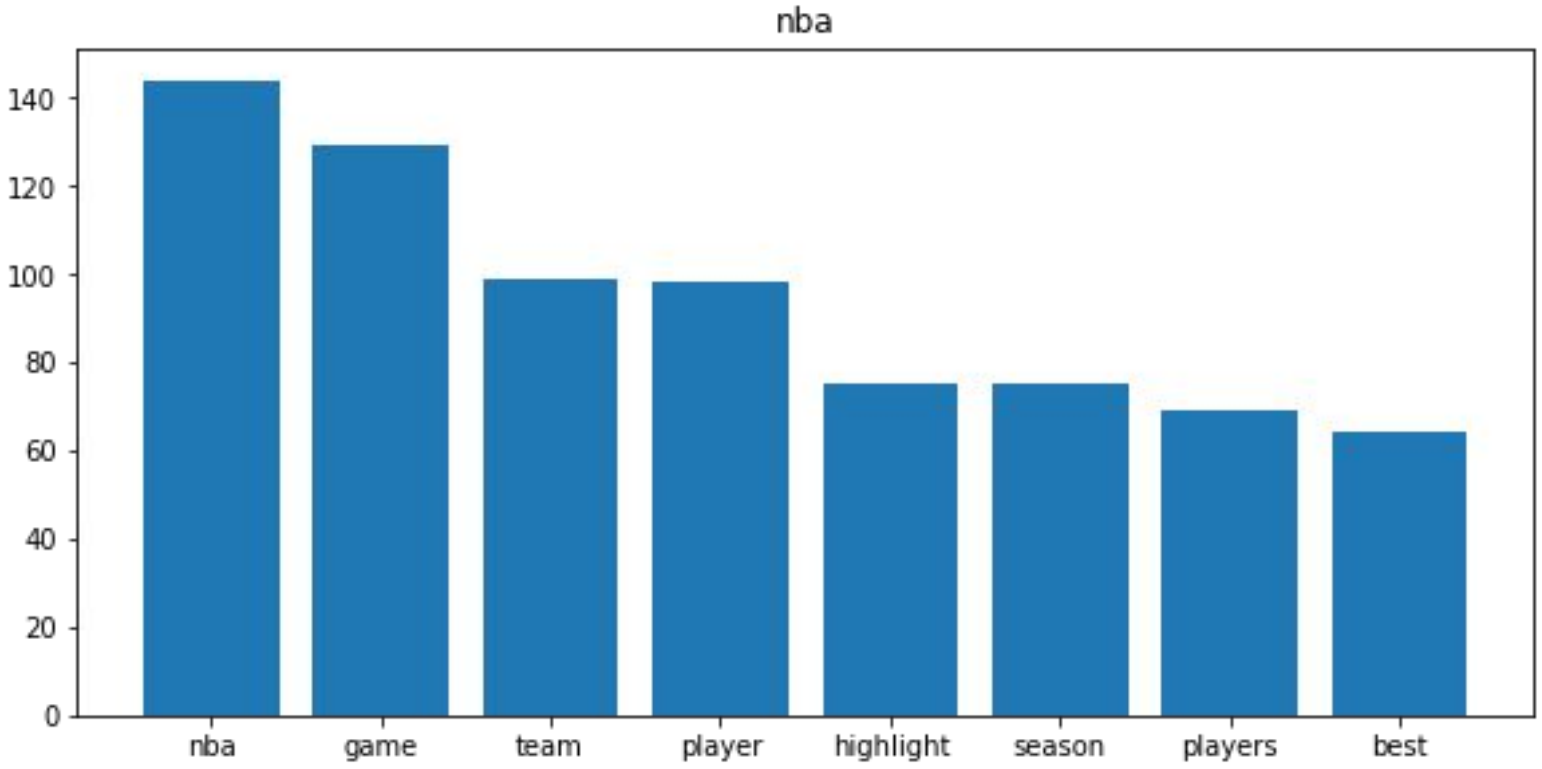
# Hockey: Competition Words



# Sports: Competition Words



# NBA: Competition, Player Words



# Tennis: Player, Competition Words

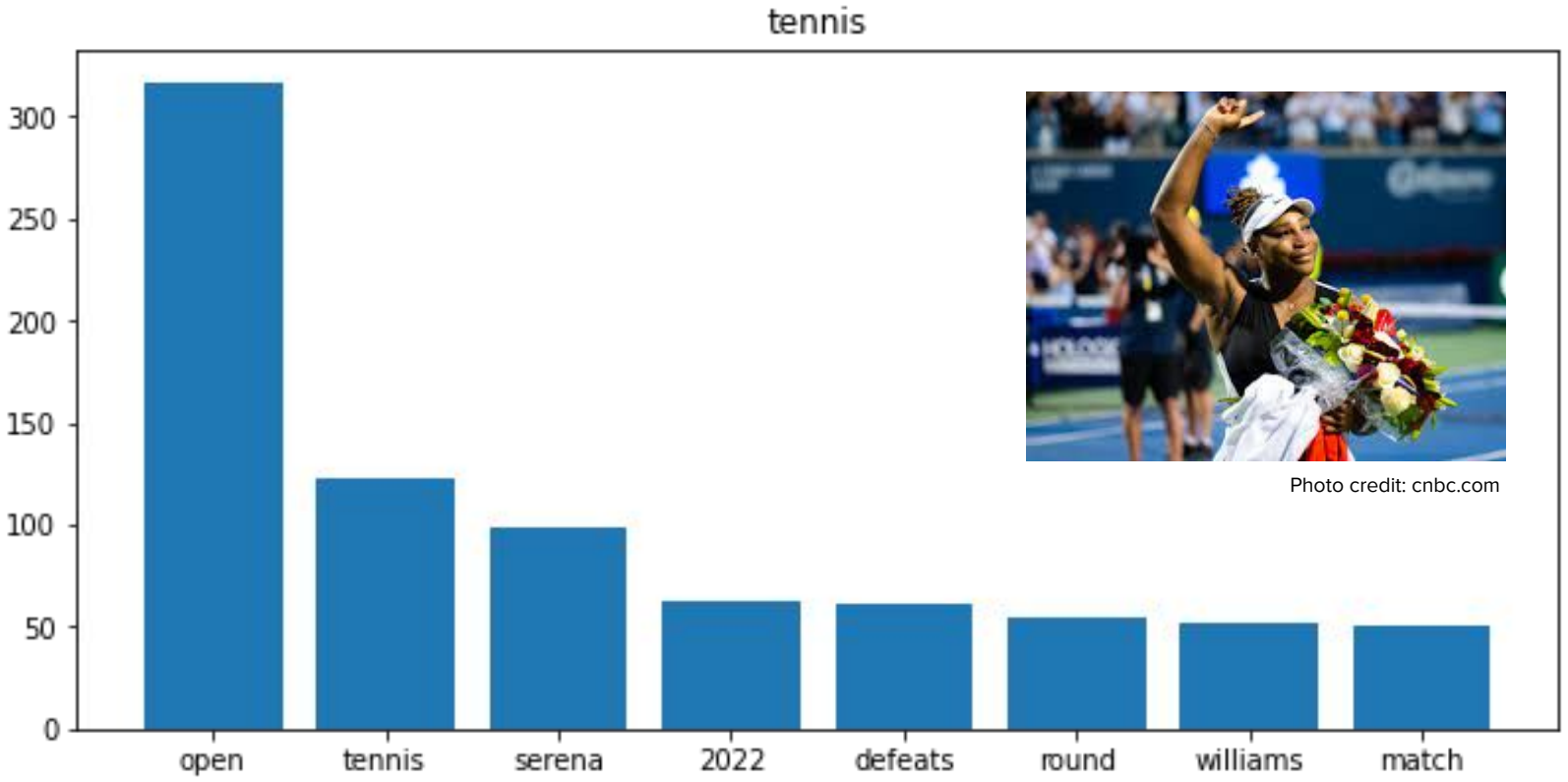
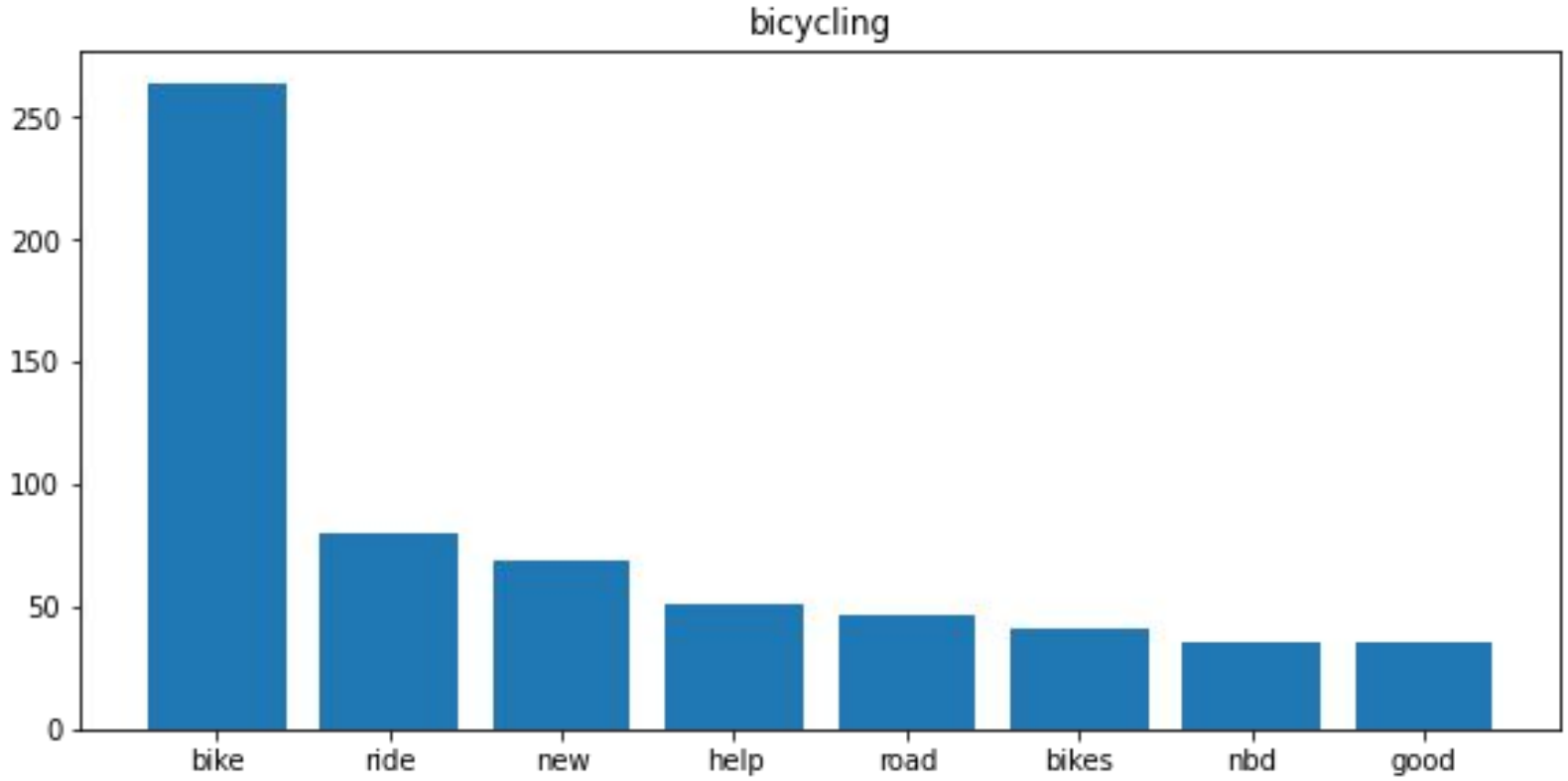


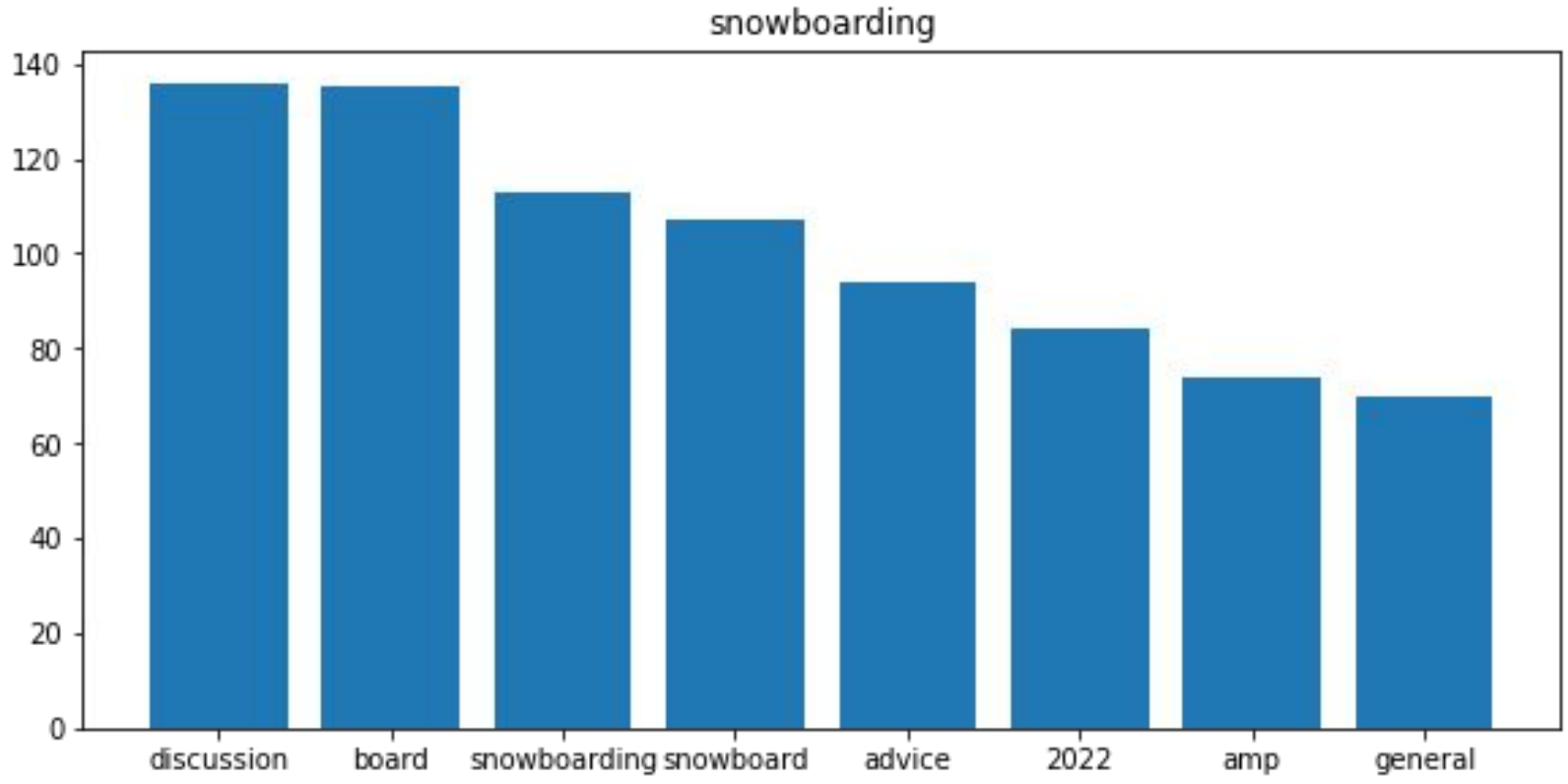



Photo credit: cnbc.com

# Bicycling: Community Words



# Snowboarding: Community Words



# Conclusions and Recommendations:

- Conclusions:
  - Modeling provided limited information
  - EDA much more interesting, showing clear connections across sports
- Where to Go from Here:
  - Analyze individual sport subreddits (get more data)
  - Classify sports, analyze classes
    - Team vs. Individual, Game vs. Hobby, etc.
  - Use custom list of “stop words” and better tokenization to improve model performance