Who Said That?

Reddit API, NLP, & Classification Models

By: Tofer Kim

Real Housewives and the NBA

Similarities

- Televised programming
- Franchises represent cities
- Entertainment value--Drama
- Wealthy, aspirational
- Seasonal Contracts/MVP/Rookies

Differences

- Occupation
- Gender
- Performance stats
- Fan base*

r/BravoRealHousewives or r/NBA?

— — —

- Haven't you heard? __Ramona__ is the MVP of every season.
- What a bitch! Does anyone actually like _Joakim Noah_?
- My problem this whole season is that they've all been assholes and there's no one to root for. (Sonja Morgan)
- I'm sure the contracts are different now compared to 2011/12 when _Alex McCord_ wasn't asked back.
- _Scott Brooks_ is the dumbest human being alive and that's not hyperbole at all.
- God damn the difference in height! (Erika Girardi & Kandi Burruss)
- What a terrible parent letting their kid drink champagne. (Kemba Walker)
- As long as _Steph Curry_ took back those statements I don't care what the circumstances are. [He/she is] not stubborn enough to keep up the ignorance and is making an effort to learn and adapt [his/her] beliefs.

Problem Statement

- Can we use NLP and classification methods to determine whether a post or comment came from a particular subreddit?
- Are there any similarities between fans of Bravo's Real Housewives franchise and fans of the NBA?

Top 20 Most Frequent Words

Unique words highlighted

NBA	Real Housewives
game	like
player	think
team	season
nba	show
play	get
point	one
season	know
like	housew
year	episod
would	would
get	realli
one	discuss
time	say
lebron	go
shoot	make
think	look
pt	watch
good	love
pm	time
win	friend

Top 20 Most Important Words

NBA

Real Housewives

Gradient Boost

game
team
player
housew
nba
play
show
episod
fuck
look
carolin
jax
bravo
shit
know
lebron
guy
kim
trade
realli

Soft Voting Classifier

Logistic Regression	Random Forest	AdaBoost	Gradient Boost	Multinomial NB
15%	15%	15%	15%	40%

87%

Accuracy score on unseen data (baseline of 51%)

Further Exploration

- Include more features (higher n-grams), remove common words, and tune different hyperparameters for our models.
- Analyze user patterns such as: length of text, sentiment scores, use of numbers, emojis, and hyperlinks.
- Correct spelling errors and identify variations in spelling.

Thank you!