Subreddit Classification

E.R. Schultz

The Questions:

Can a submission title predict the source subreddit thread?

Can comments predict the source subreddit thread?

What similarities in language exist between related and unrelated threads?

Science & Conspiracy

Methodology

1. Scrape Data
2. Investigate Data
3. Clean and Feature Engineer
4. Design and Fit Models
5. Evaluate Results

The Data

“Temporal Dynamics in Viral Shedding and Transmissibility of Covid-19”

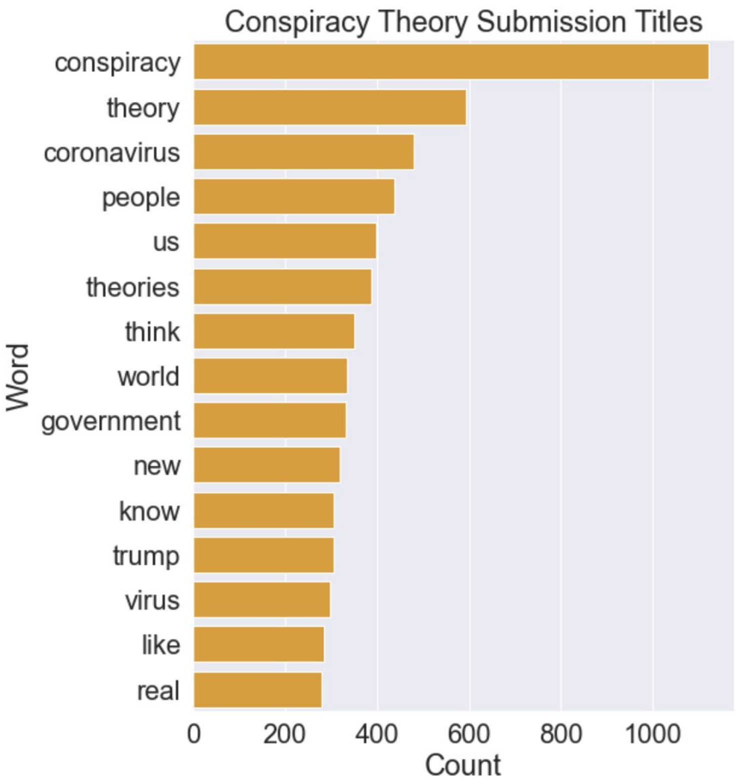
“Mangoes, Covid-19, and Aliens”

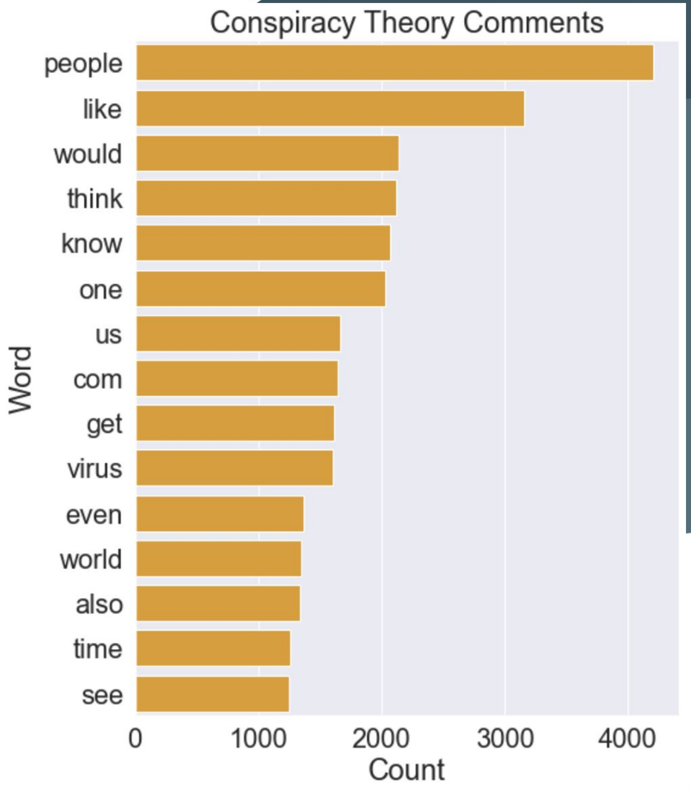
“The Wizard of Wuhan, starring Bill Gates.”

“Source?”

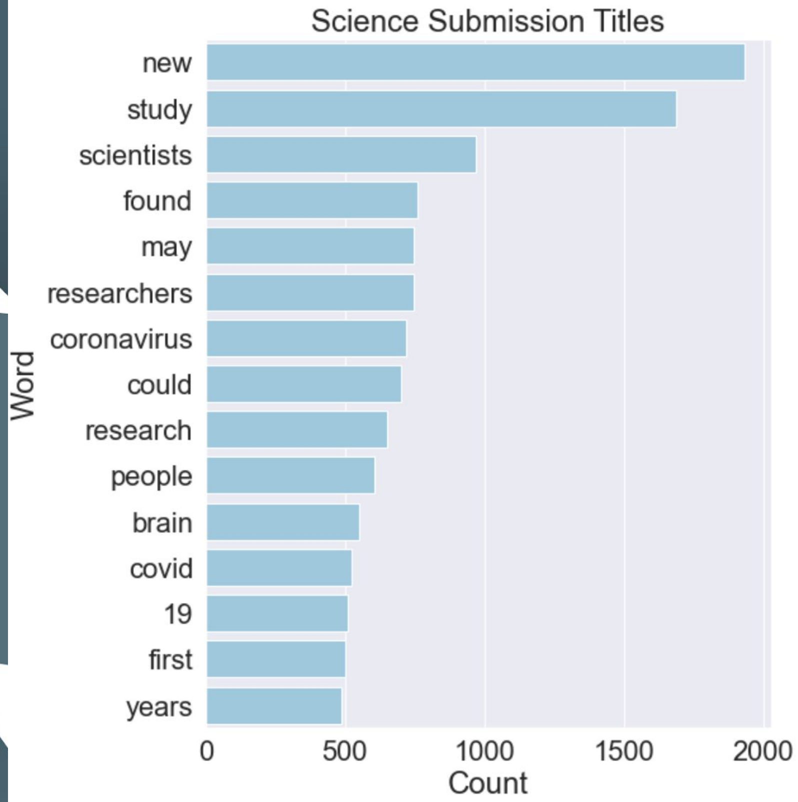
“Enjoy.”

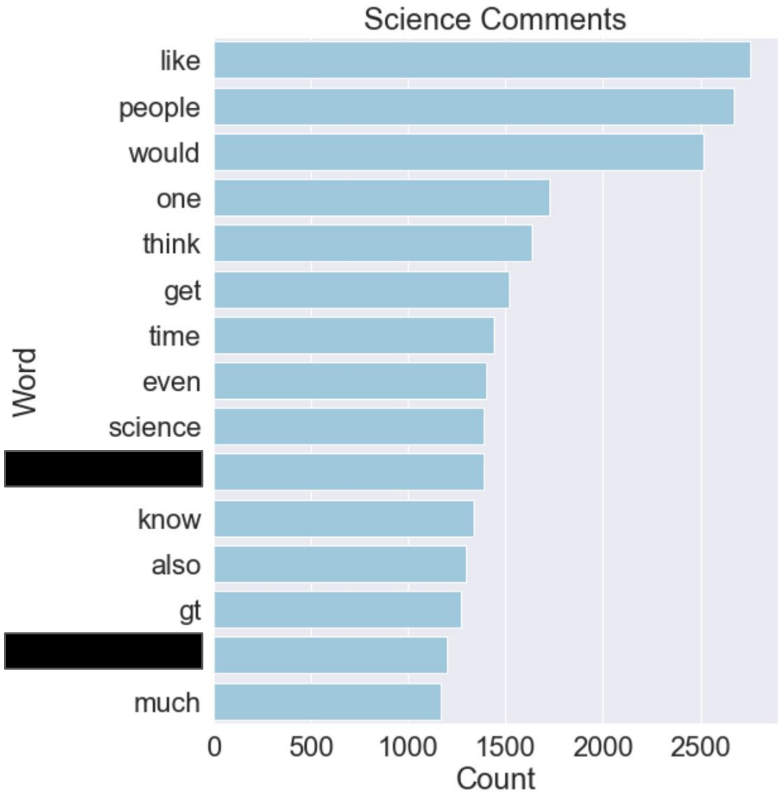
Top Words: r/conspiracies



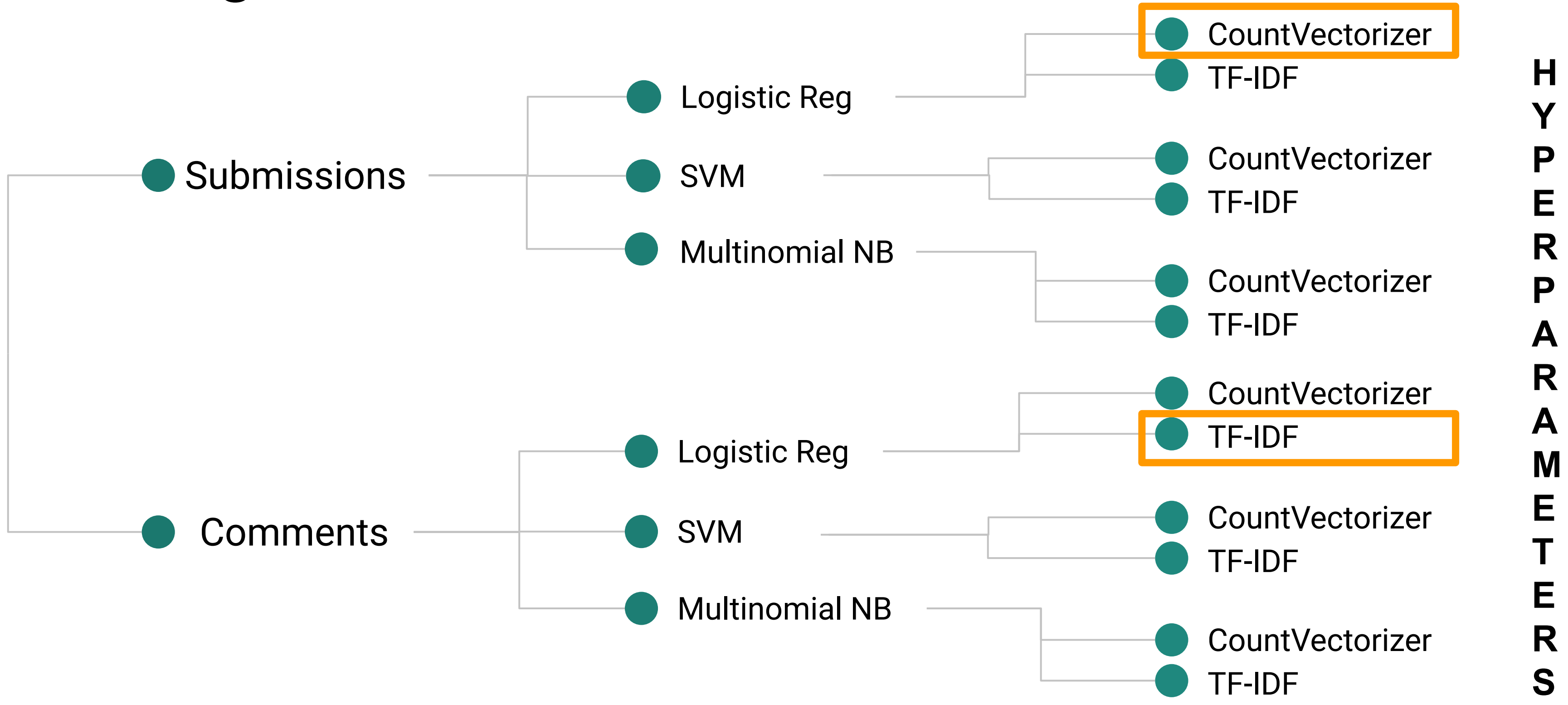


Top Words: r/science


Science Submission Titles


Science Comments

Modeling and Model Choice:

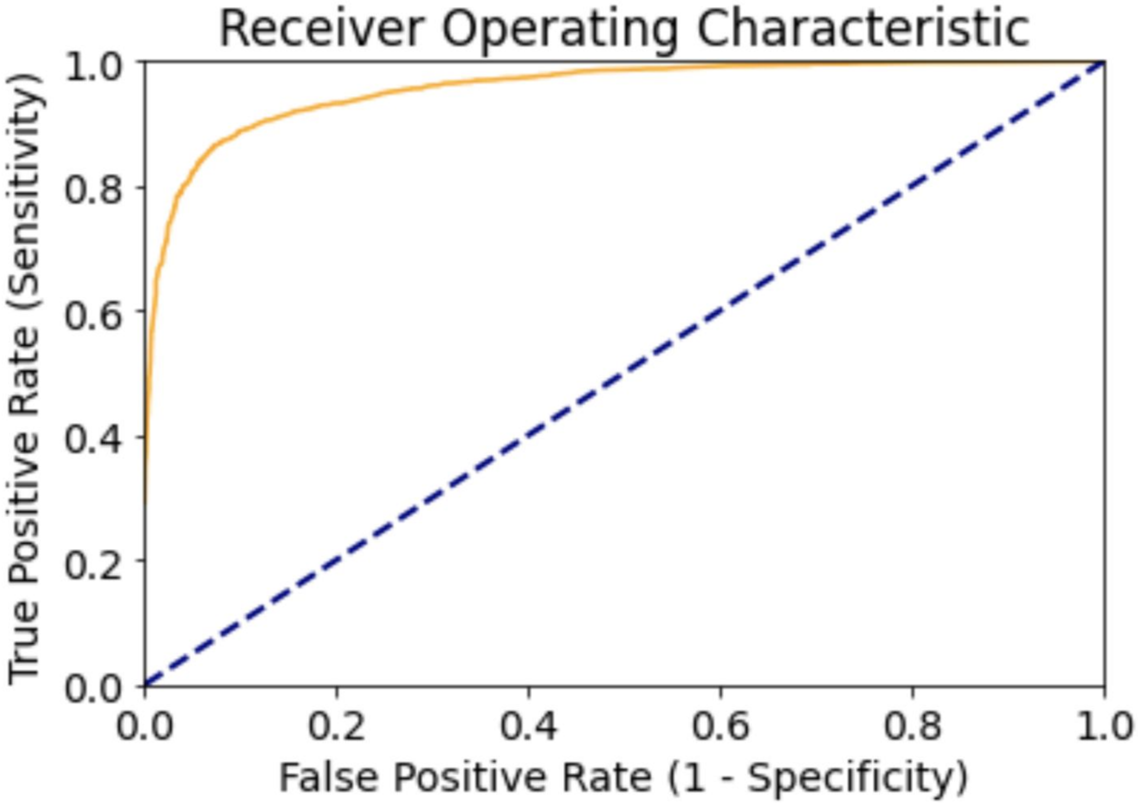


Submissions Model:

Logistic Regression

Accuracy Score: .895

AUC: .895

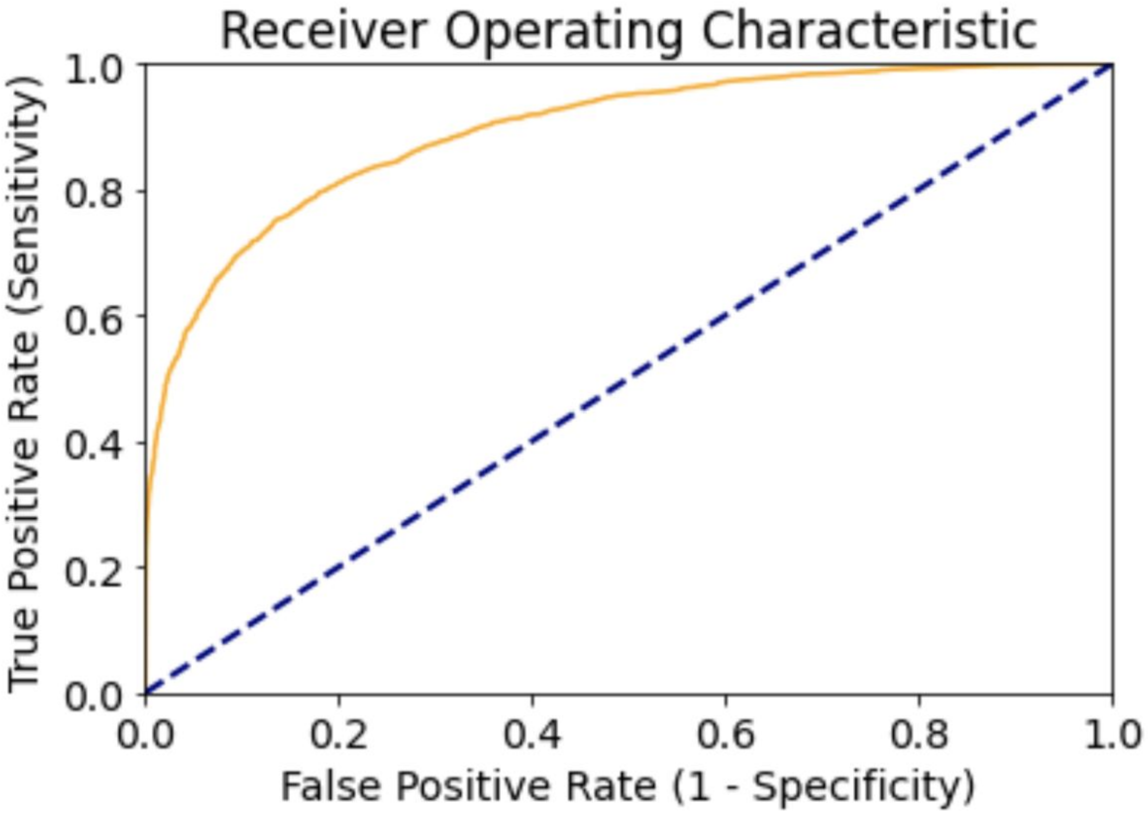


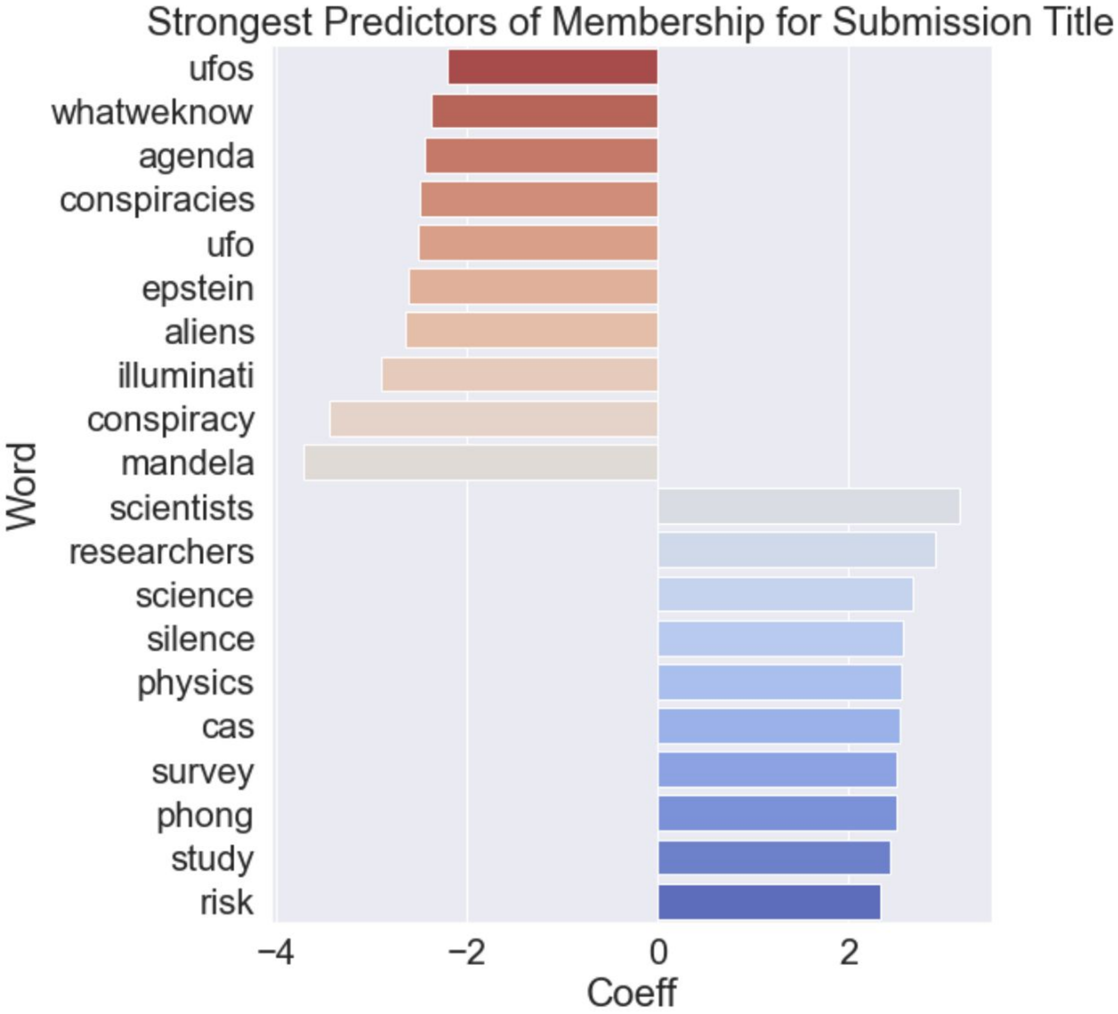
Comments Model:

Logistic Regression

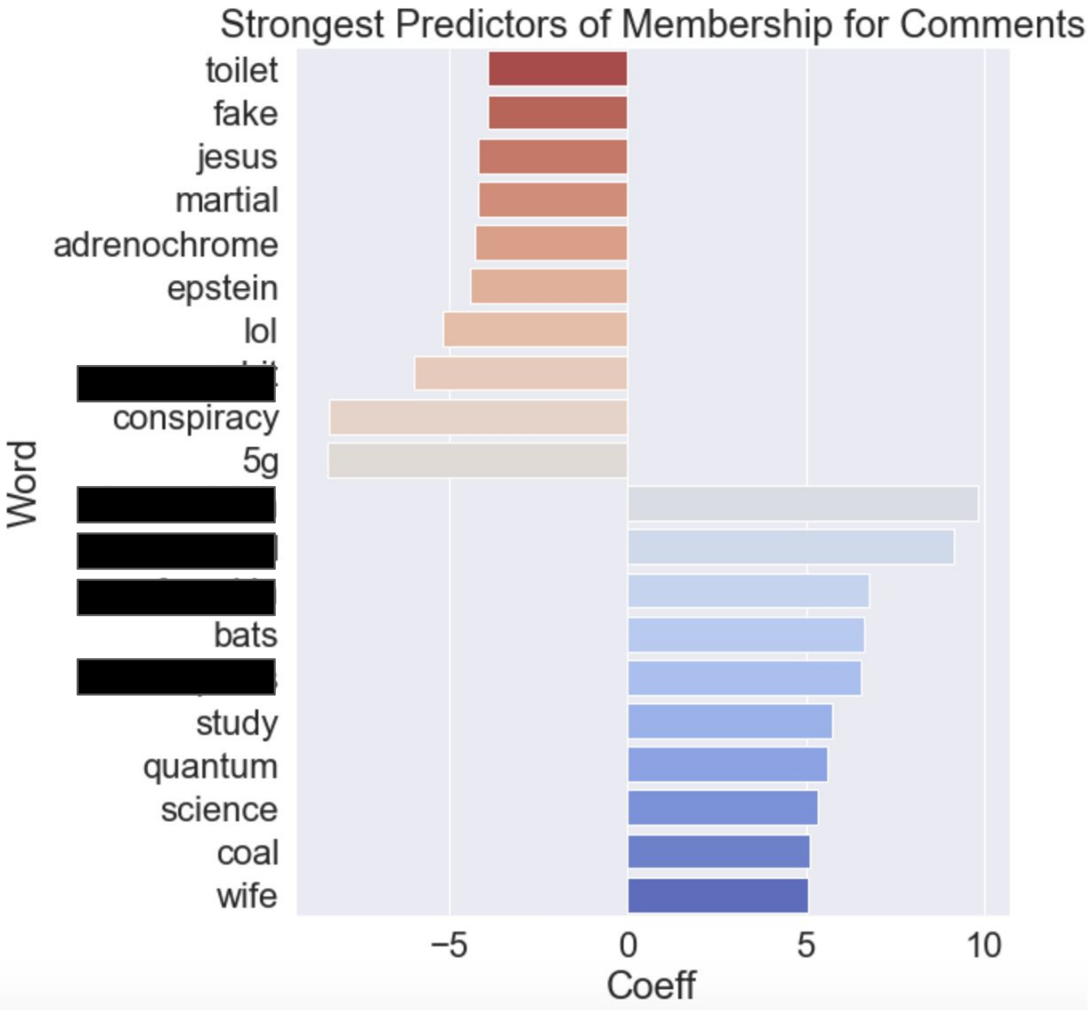
Accuracy Score: .807

AUC: .807





Strongest Predictors of Membership for Submission Title


Strongest Predictors of Membership for Comments
toilet
fake
jesus
martial
adrenochrome
epstein
lol
[redacted]
conspiracy
5g
[redacted]
[redacted]
[redacted]
bats
[redacted]
study
quantum
science
coal
wife
Word
−5
0
5
10
Coeff

Answers and a Recommendation:

- **Can we predict membership?** An accuracy score of 89% isn’t bad, but remember Reddit’s daily comments based on a 10-year average total ~470,000. That means we’re mislabeling 47k posts a day.

- **Are there similarities?** Similarities do seem to exist in the frequencies of used words, but not in the words most predictive of membership.

- It’s tempting to start thinking about extrapolating this method for user classification. Before using any of this data for user categorization or classification, higher model accuracy, a deep understanding of inherent Type I and Type II error, and serious ethical considerations are required.

Questions?