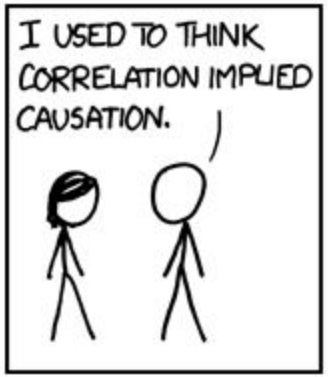
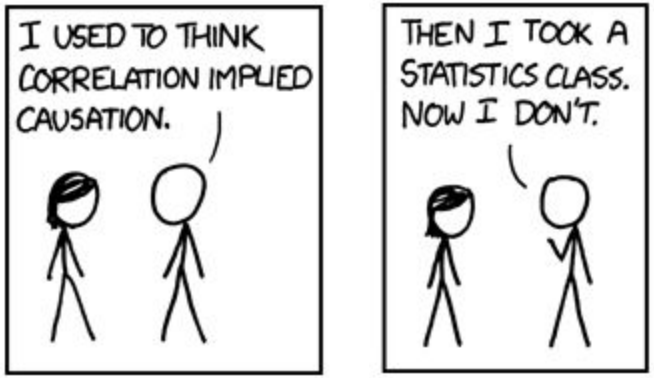
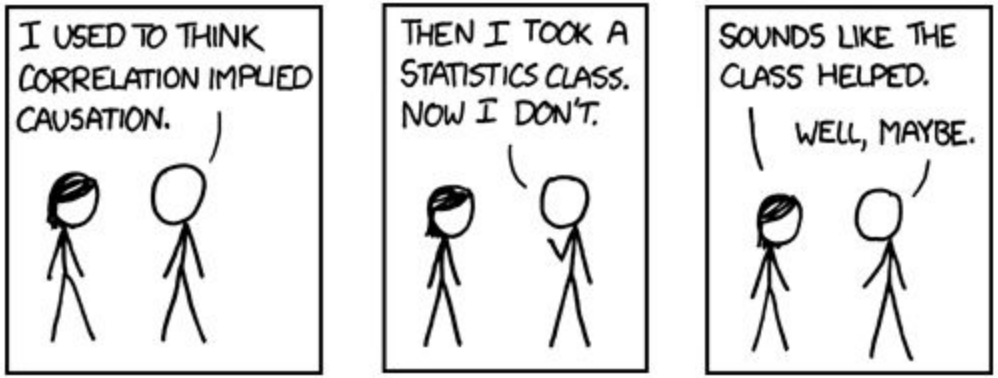
Classifying Subreddits: Jokes vs. Riddles



https://xkcd.com/552/



https://xkcd.com/552/

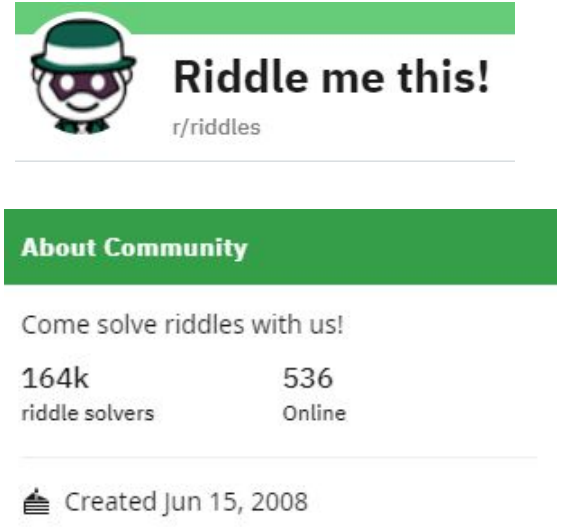


https://xkcd.com/552/

Problem:

Can we create a model to identify which subreddit a post is from?

Defining the Categories



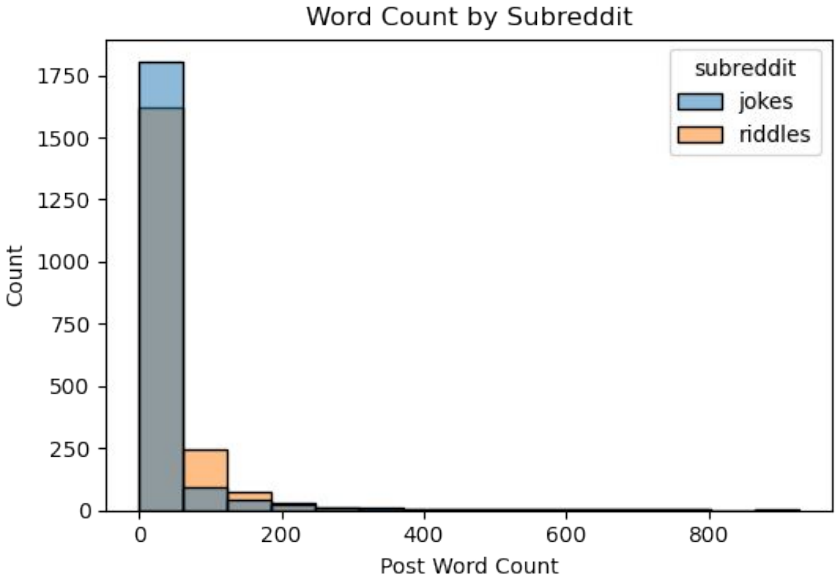
Baseline accuracy: 50%

Features & Model

Selection and Process

Feature Selection - Post Text

	Jokes	Riddles
Avg Word Count	23	45
Avg String Length	126	247



Model Selection

Logistic Regression Model:

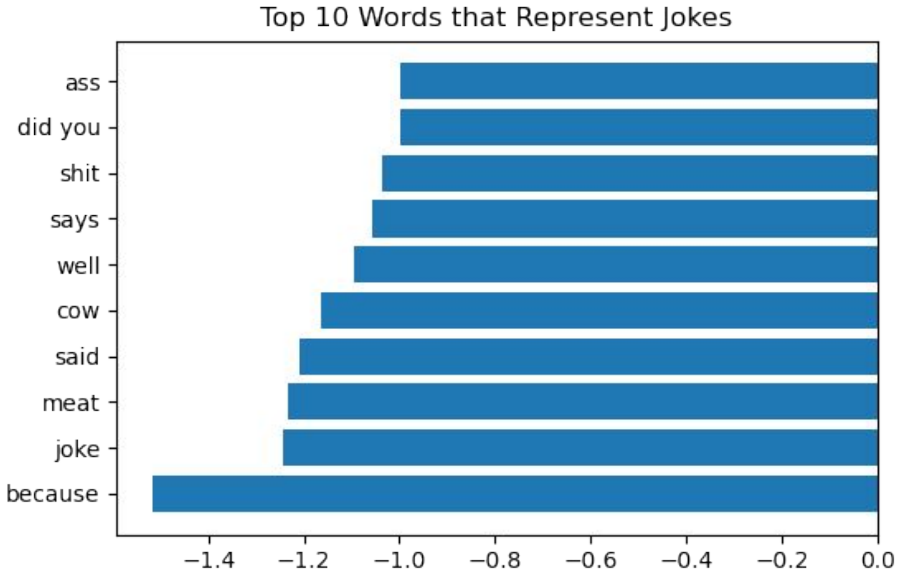
- Lemmatization
- CountVectorizer
  - N-gram: (1, 2)
  - Keep stop words
  - Max Features at 2,500

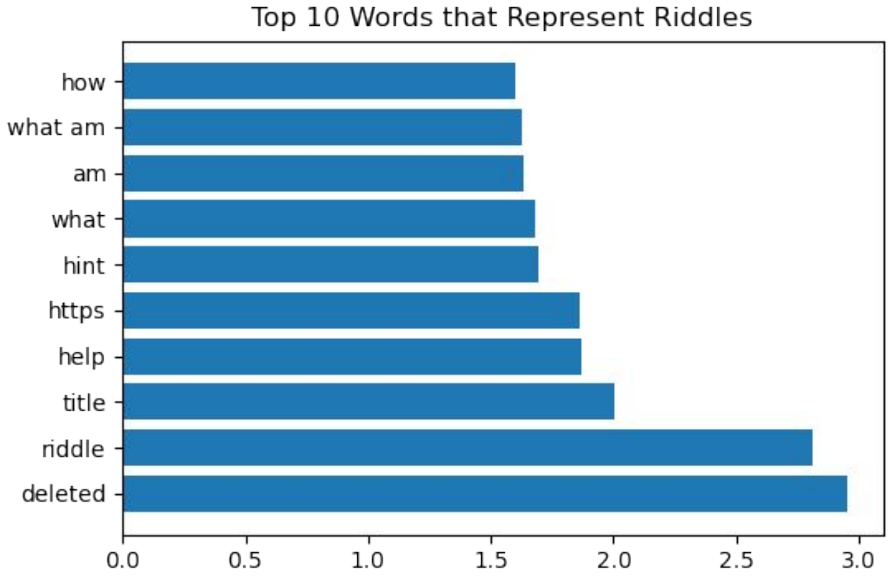
0.88 Accuracy

Model Evaluation

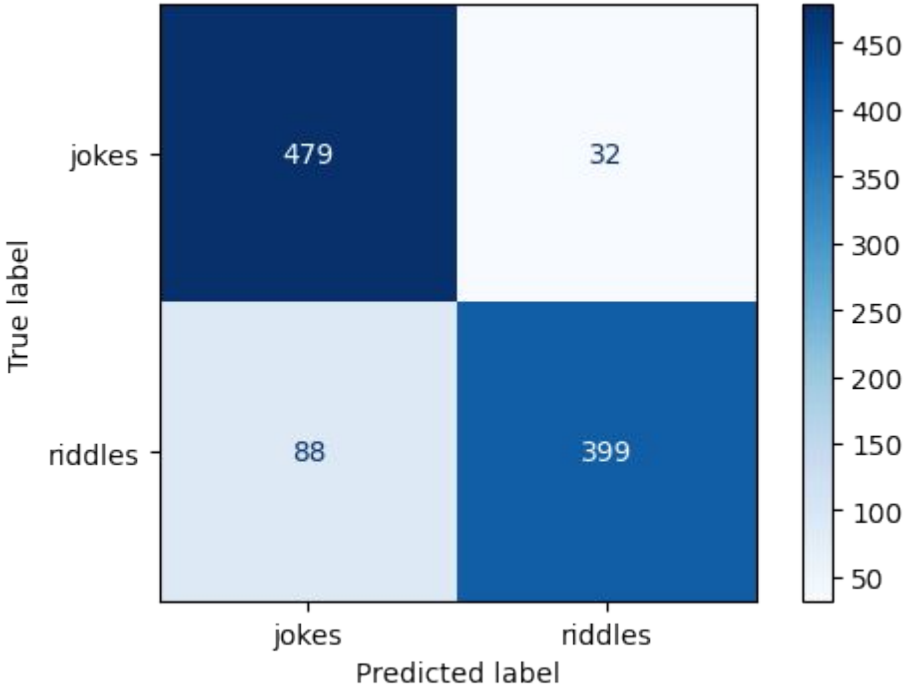
Metrics & Testing

Top Words by Subreddit





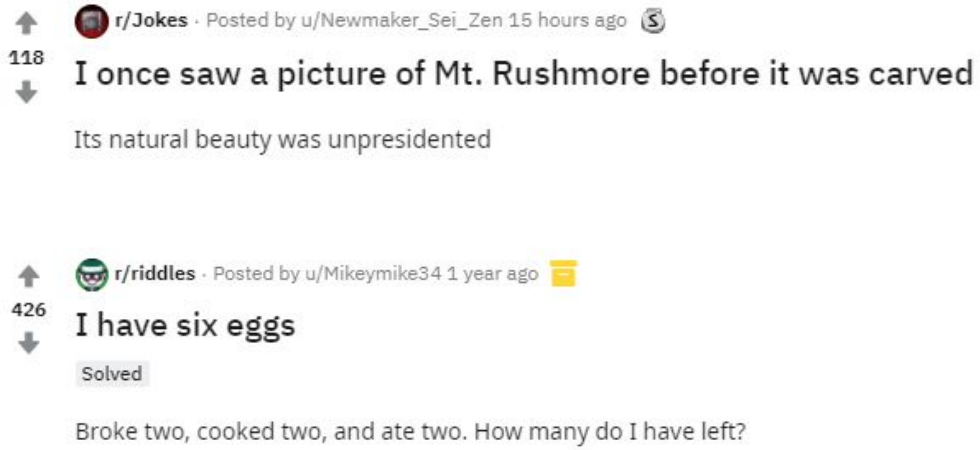
Test Set Results



	Jokes	Riddles
Precision	0.84	0.93
Recall	0.94	0.82
F1-Score	0.89	0.87

Test Set Accuracy = 0.88

Testing with Posts Not in Dataset



When posts outside of the dataset were pulled the model guessed correctly both times.

Conclusion & Recommendations

- Use both title and text columns
- Remove posts that have content deleted
- Clean up riddles data to not show any links