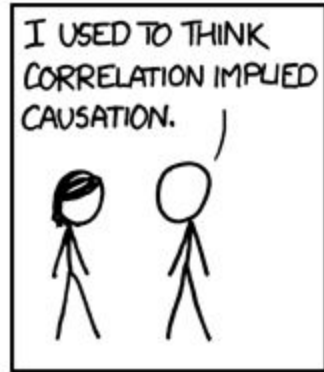
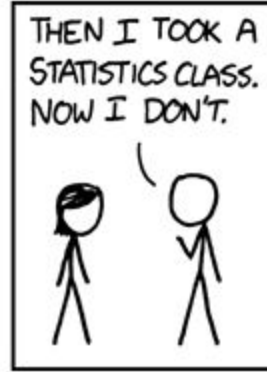
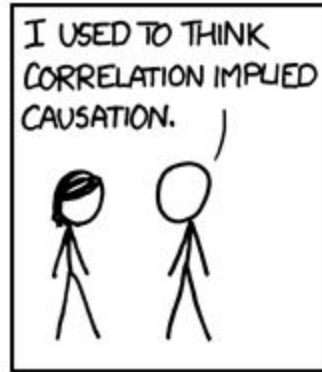
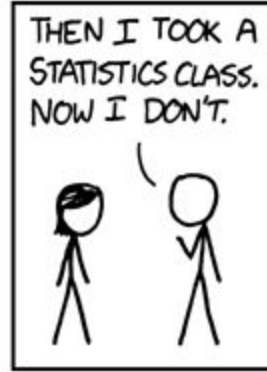
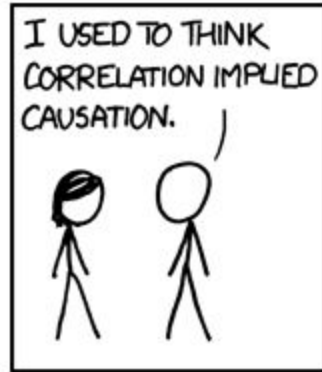


# Classifying Subreddits: Jokes vs. Riddles







# Problem:

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

# Defining the Categories



## Jokes: Get Your Funny On!

r/Jokes

### About Community

The funniest sub on reddit. Hundreds of jokes posted each day, and some of them aren't even reposts!

20.5m  
Members

24.8k  
Online

🏠 Created Jan 25, 2008



## Riddle me this!

r/riddles

### About Community

Come solve riddles with us!

164k  
riddle solvers

536  
Online

🏠 Created Jun 15, 2008

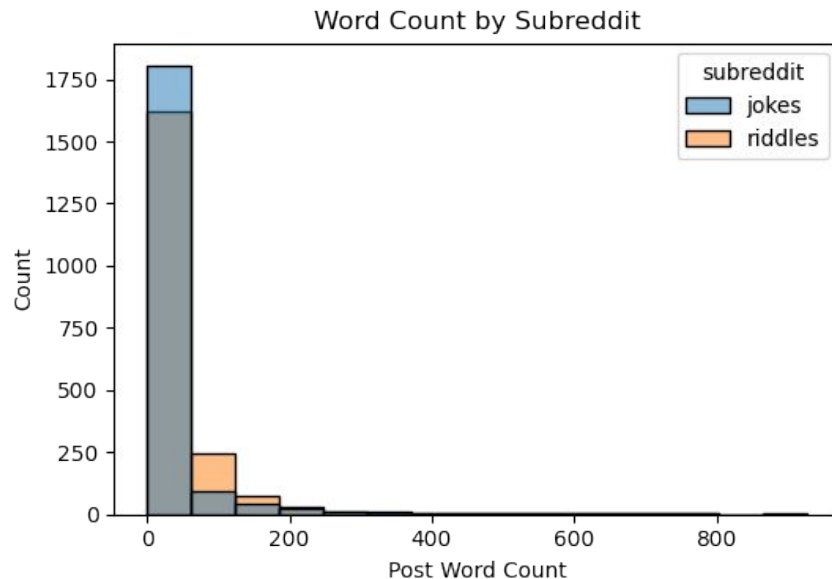
Baseline accuracy: 50%

# 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



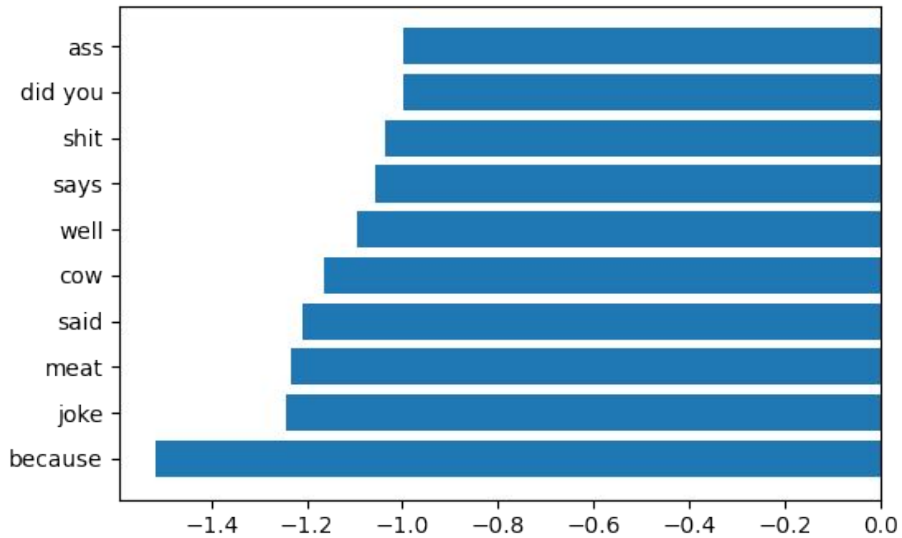
**0.88**  
**Accuracy**

# Model Evaluation

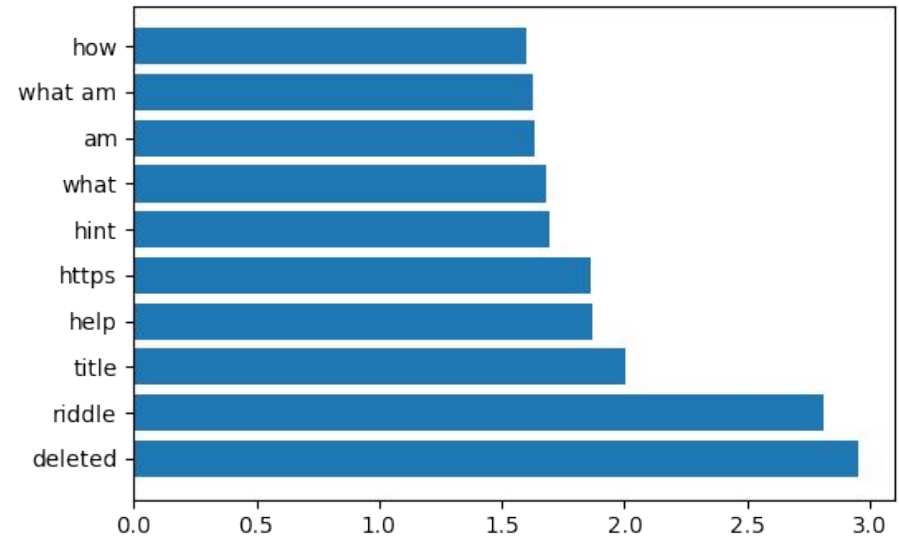
Metrics & Testing

# Top Words by Subreddit

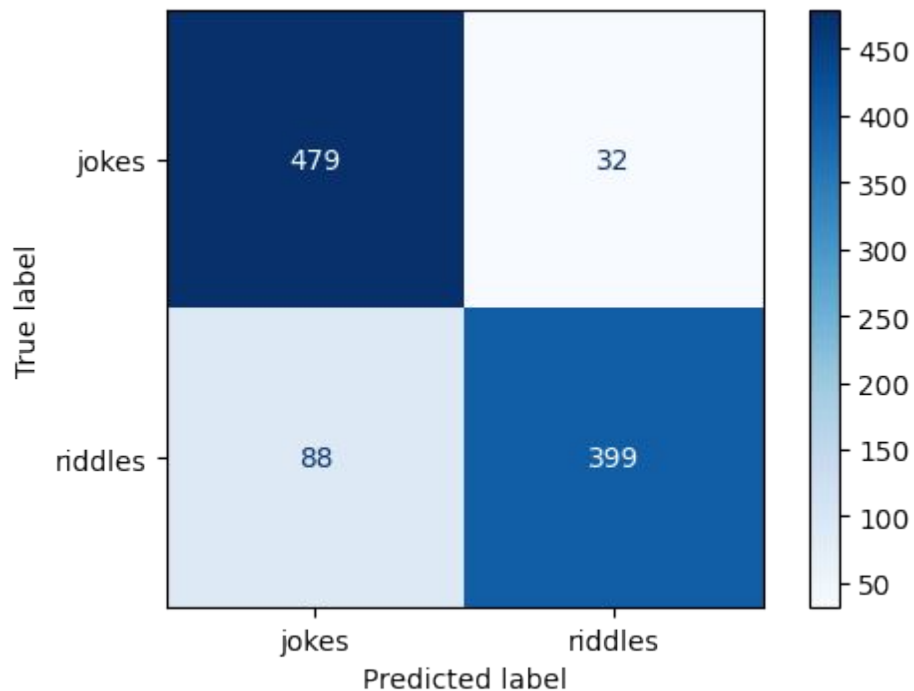
Top 10 Words that Represent Jokes



Top 10 Words that Represent Riddles



## 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

↑ 118 ↓  
r/Jokes · Posted by u/Newmaker\_Sei\_Zen 15 hours ago  
**I once saw a picture of Mt. Rushmore before it was carved**  
Its natural beauty was unpresidented

↑ 426 ↓  
r/riddles · Posted by u/Mikeymike34 1 year ago  
**I have six eggs**  
Solved  
Broke two, cooked two, and ate two. How many do I have left?

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