CLASSIFYING SUBREDDITS

STEPHEN GODFREY

PROBLEM STATEMENT

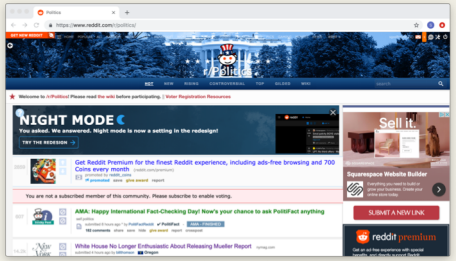
Given a Reddit post, identify its source subreddit

Data

- 7,199 posts from four subreddit categories
- Collected Saturday March 30, 2019 8:30 am - 10:30 am, 6:15 pm – 8:15 pm and Wednesday April 3 9:00 – 9:30 am
- High degree of duplication

subreddit	Posts
politics	2,379
woodworking	2,270
relationships	1,672
DIY	878
Total	7,199

Process



API

Posts
title, selftext, comments

PostgreSQL database (AWS)

Data Cleaning

Remove punctuation
Lower case

Lemmatize
Remove stop words

Vectorizing

Count Vectorizer (cvec)

Term Frequency Inverse Document Frequency (tfidf)

Modeling

Multinomial Naïve Bayes (bayem)

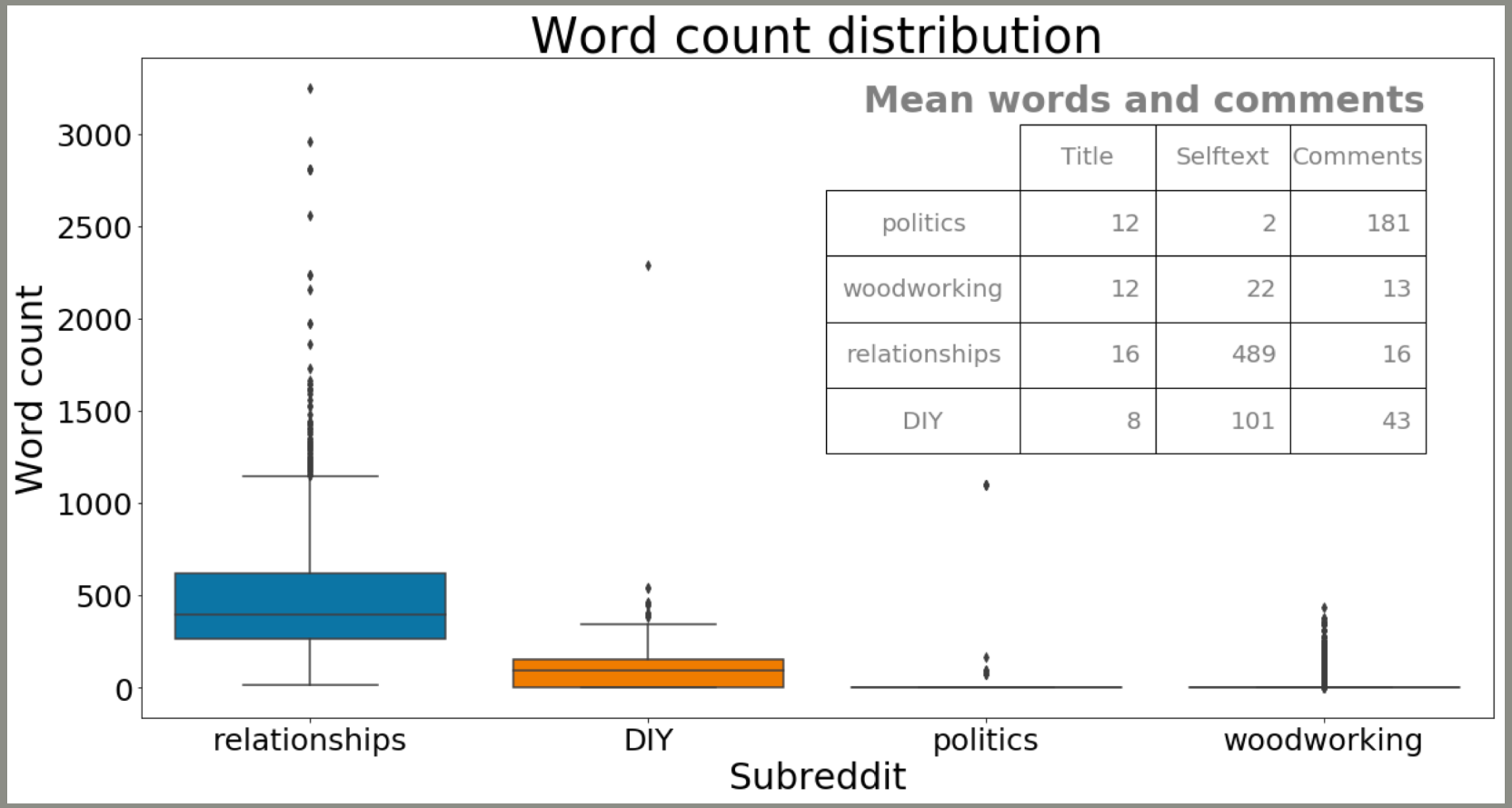
Logistic Regression (logr)

K Nearest Neighbors (knn)

Ensemble – logr & bayem (vote)

6 pairs of subreddits, 3 text fields, 2 vectorizers, 4 models = 144 models selected via grid search

Post characteristics

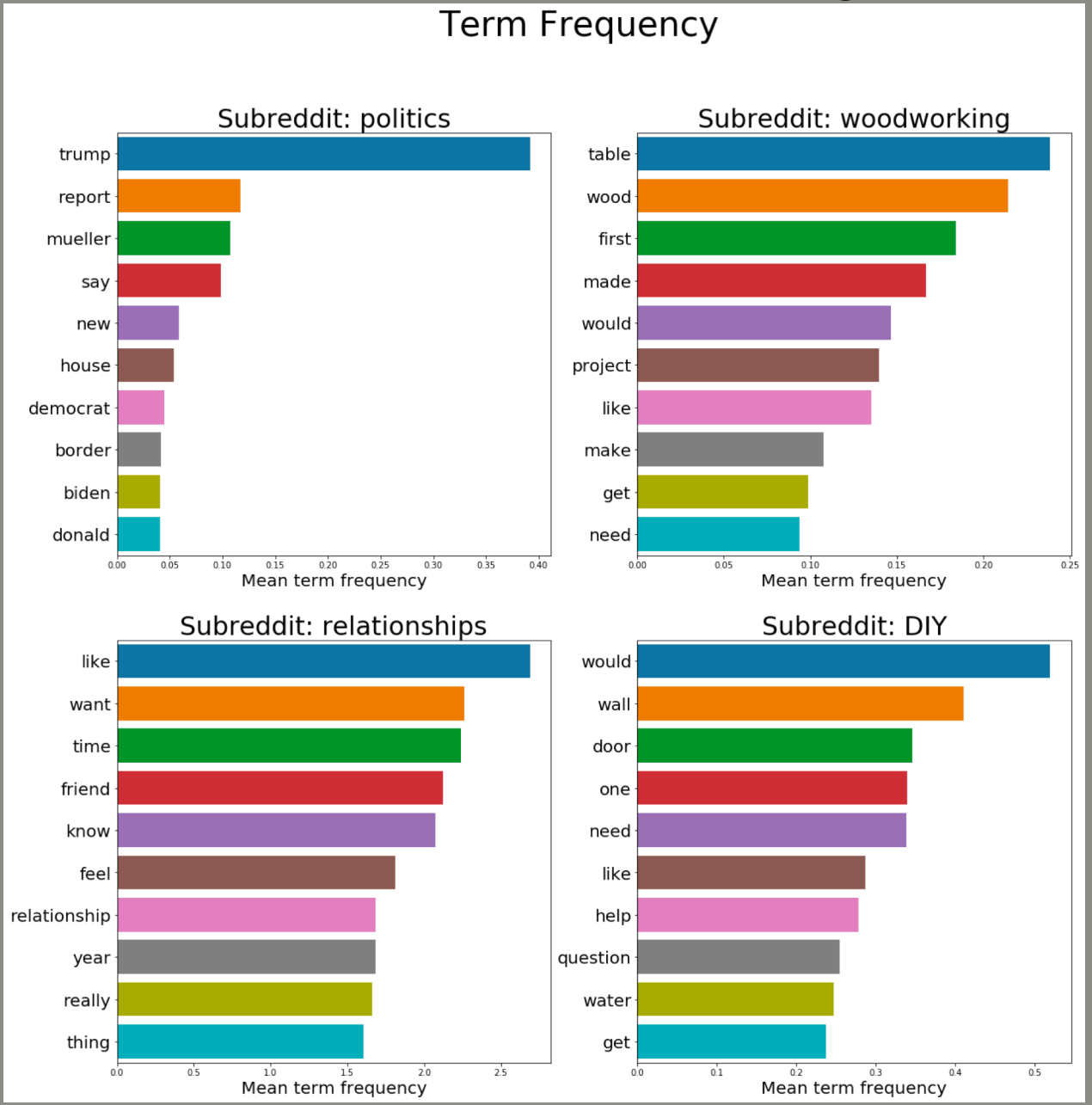


Mean words and comments

	Title	Selftext	Comments
politics	12	2	181
woodworking	12	22	13
relationships	16	489	16
DIY	8	101	43

- Observations
  - *Relationship posts have far more words than other categories*
  - *Relationship post lengths are more diverse*
  - *Politics selftext posts are short but they generate many comments*
  - *Woodworking posts are the shortest with fewest comments*

Term frequency



- Observations
  - *Some expected results and notable differences*
  - *Politics*
    - Trump, Mueller
  - *Woodworking*
    - Table, wood
  - *Relationships*
    - Like, want
  - *DIY*
    - Wall, door

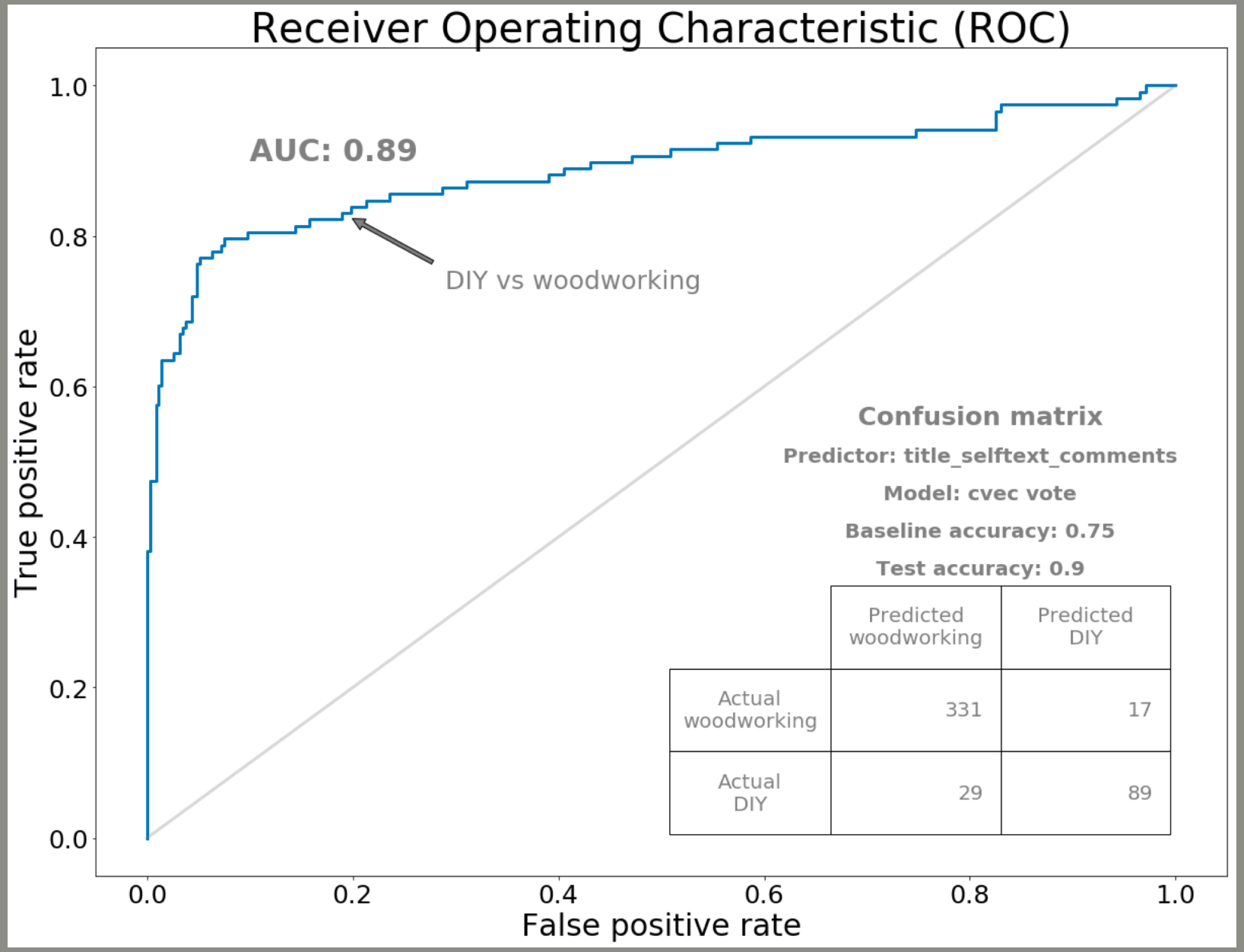
Best models

Pair	X_variable	Model	Test score	Sample
97. relationships, DIY	title_selftext_ comments	cvec logr	.998	1758
111. relationships, politics	title_selftext_ comments	tfidf vote	1.0	3026
113. relationships, woodworking	title_selftext_ comments	cvec logr	1.0	2675
126. DIY, politics	title_selftext_ comments	tfidf knn	1.0	2214
131. DIY, woodworking	title_selftext_ comments	cvec vote	.901	1863
141. politics, woodworking	title_selftext_ comments	tfidf logr	1.0	3131

- High testing accuracy observed across all pairs with some groups at 100%
- Lowest performance is between similar topics
- Using title, selftext and comments worked best
- Averages

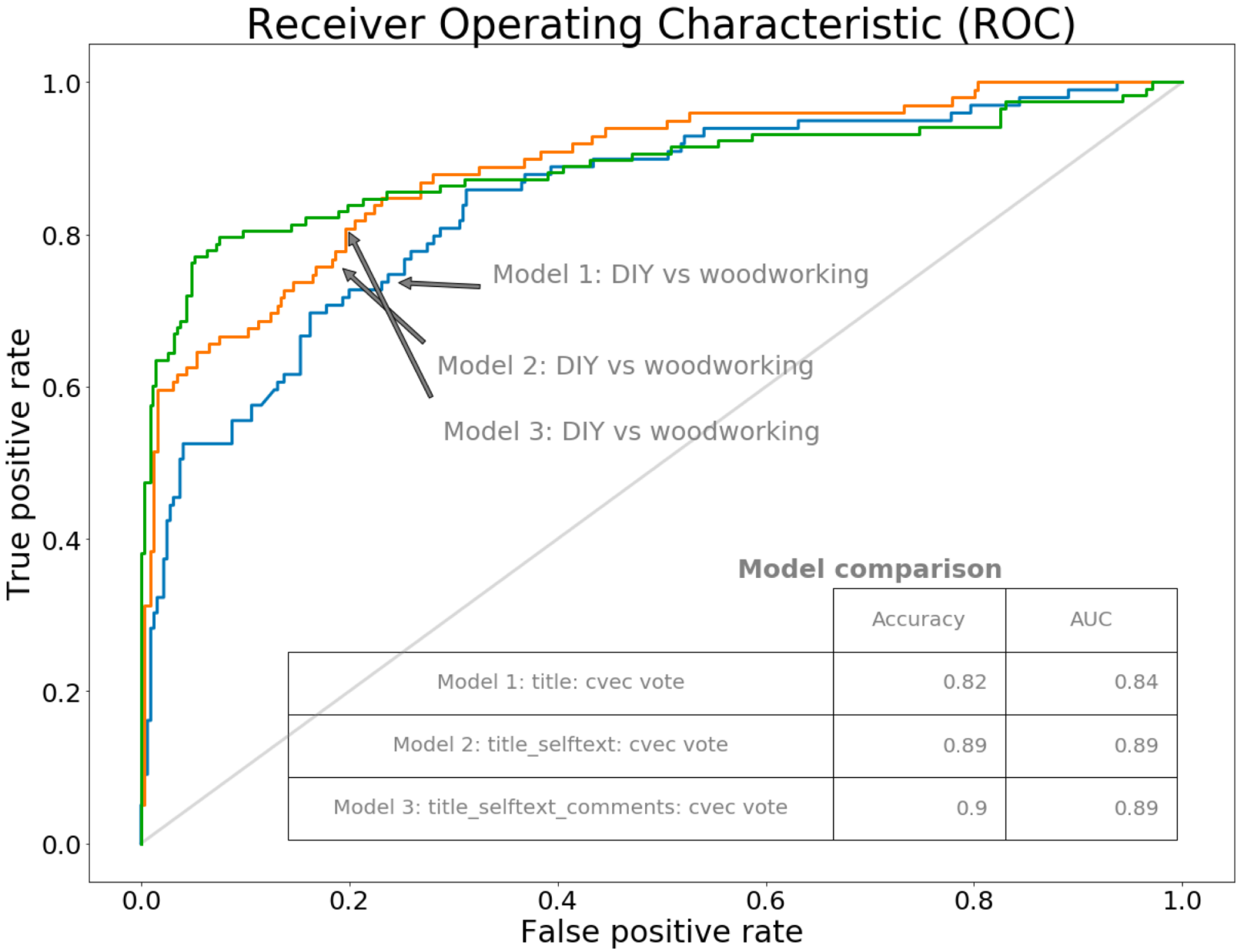
Model	Average test score
cvec bayem	0.973
cvec knn	0.822
cvec logr	0.982
cvec vote	0.979
tfidf bayem	0.887
tfidf knn	0.968
tfidf logr	0.971
tfidf vote	0.921

Similar topics



- The most difficult classification challenge
- Accuracy
  - *90% versus 75% baseline*
- Better at identifying woodworking
  - *Sensitivity = 75%*
  - *Specificity = 95%*

Adding selftext & comments

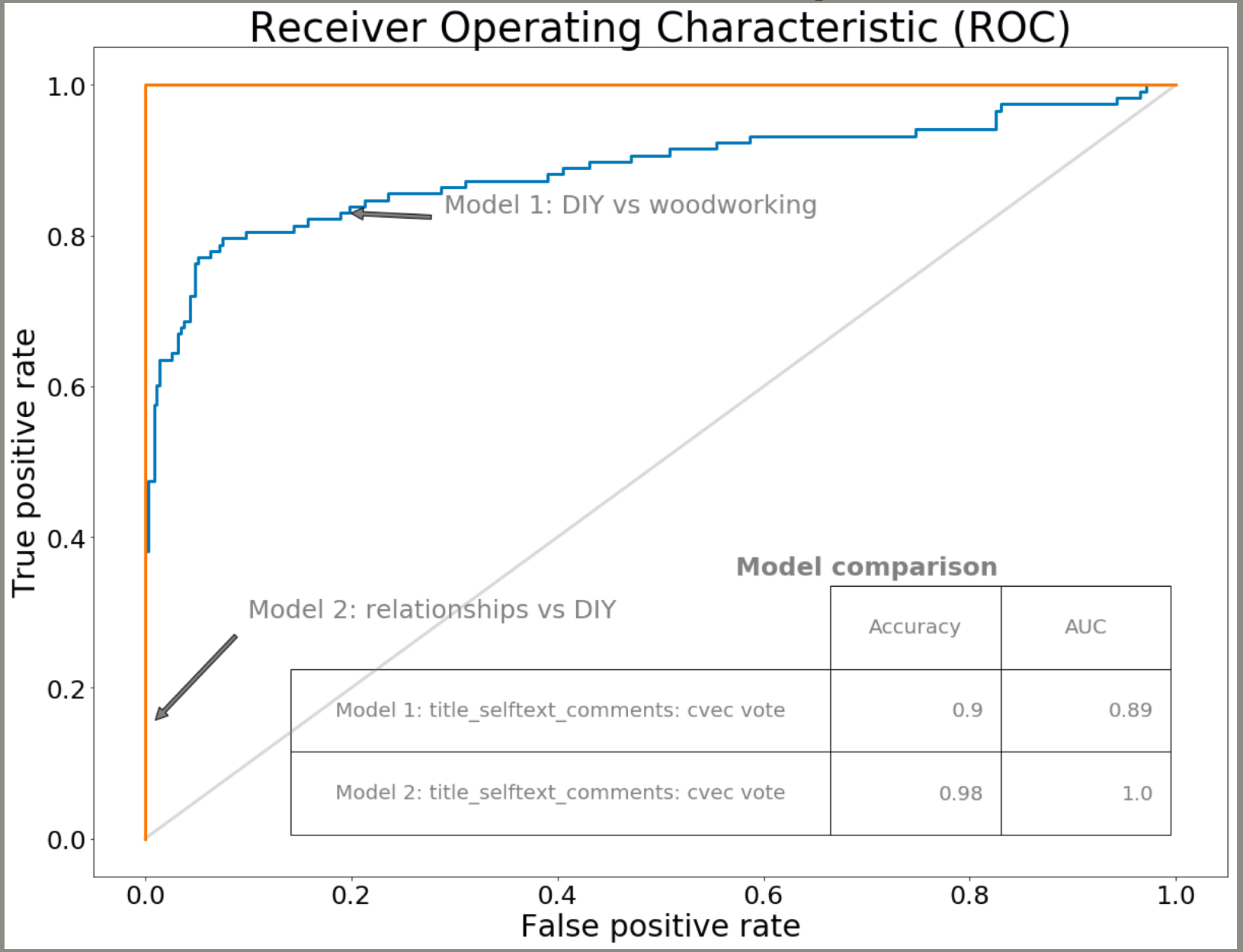


Model comparison

	Accuracy	AUC
Model 1: title: cvec vote	0.82	0.84
Model 2: title_selftext: cvec vote	0.89	0.89
Model 3: title_selftext_comments: cvec vote	0.9	0.89

- Comparing three models
  - *Title*
  - *Selftext*
  - *Comments*
- Most of the gain comes from selftext
  - *Accuracy improved to 89% from 82%*
- Modeling techniques remained constant

Dissimilar topics



Model comparison

	Accuracy	AUC
Model 1: title_selftext_comments: cvec vote	0.9	0.89
Model 2: title_selftext_comments: cvec vote	0.98	1.0

- The model works well with diverse topics
- Compare relationships and DIY to woodworking and DIY using the same model-variable paramreters

CONCLUSION

1. Models work well between dissimilar topics

2. Best models vary but Count Vectorization and Logistic Regression performed best on average

3. Adding comments provides limited benefit and is computationally intensive