CLASSIFYING SUBREDDITS

STEPHEN GODFREY

PROBLEM STATEMENT

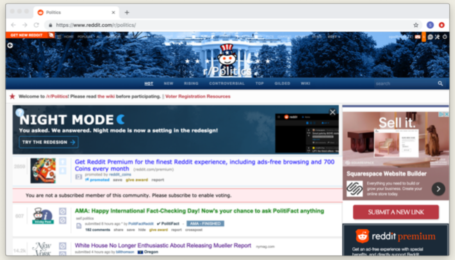
Given a post, identify the source subreddit

Data

- 6,200 posts from four subreddit categories
- Collected Saturday March 30, 2019 from 8:30 am - 10:30 am and 6:15 pm – 8:15 pm
- High degree of duplication

subreddit	Posts
politics	2,129
woodworking	2,020
relationships	1,424
DIY	627
Total	6,200

Process



Politics

https://www.reddit.com/r/politics/

r/Politics

HOT NEW RISING CONTROVERSIAL TOP GILDED WIKI

Welcome to /r/Politics! Please read the wiki before participating. | Voter Registration Resources

NIGHT MODE

You asked. We answered. Night mode is now a setting in the redesign!

TRY THE REDESIGN

Get Reddit Premium for the finest Reddit experience, including ads-free browsing and 700 Coins every month (reddit.com/premium)

You are not a subscribed member of this community. Please subscribe to enable voting.

AMA: Happy International Fact-Checking Day! Now's your chance to ask PolitiFact anything

182 comments share save hide give award report crosspost

White House No Longer Enthusiastic About Releasing Mueller Report nymag.com

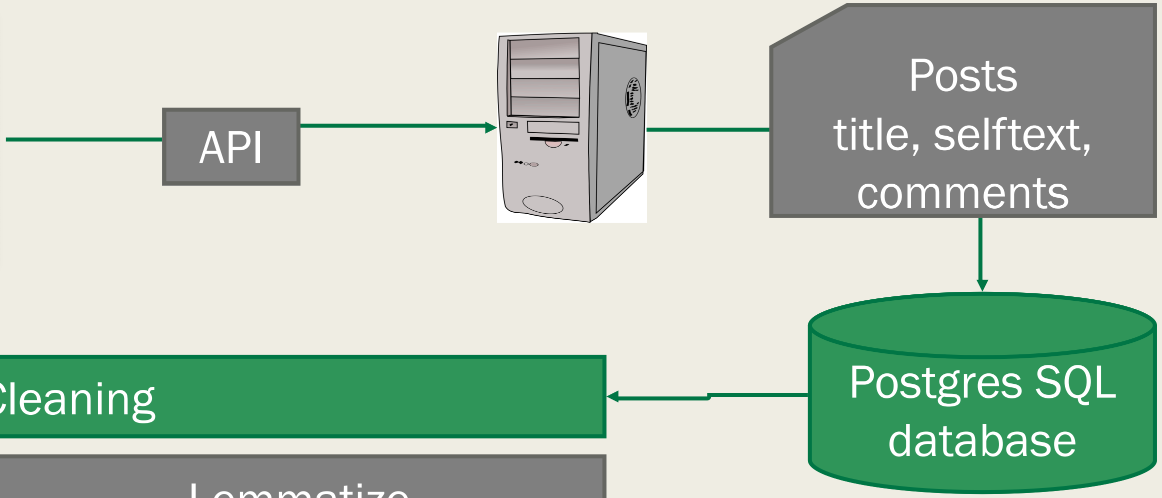
Sell it.

Squarespace Website Builder

Everything you need to build or grow your business. Create your online store today.

SUBMIT A NEW LINK

reddit premium



API

Posts
title, selftext, comments

Postgres SQL database

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)

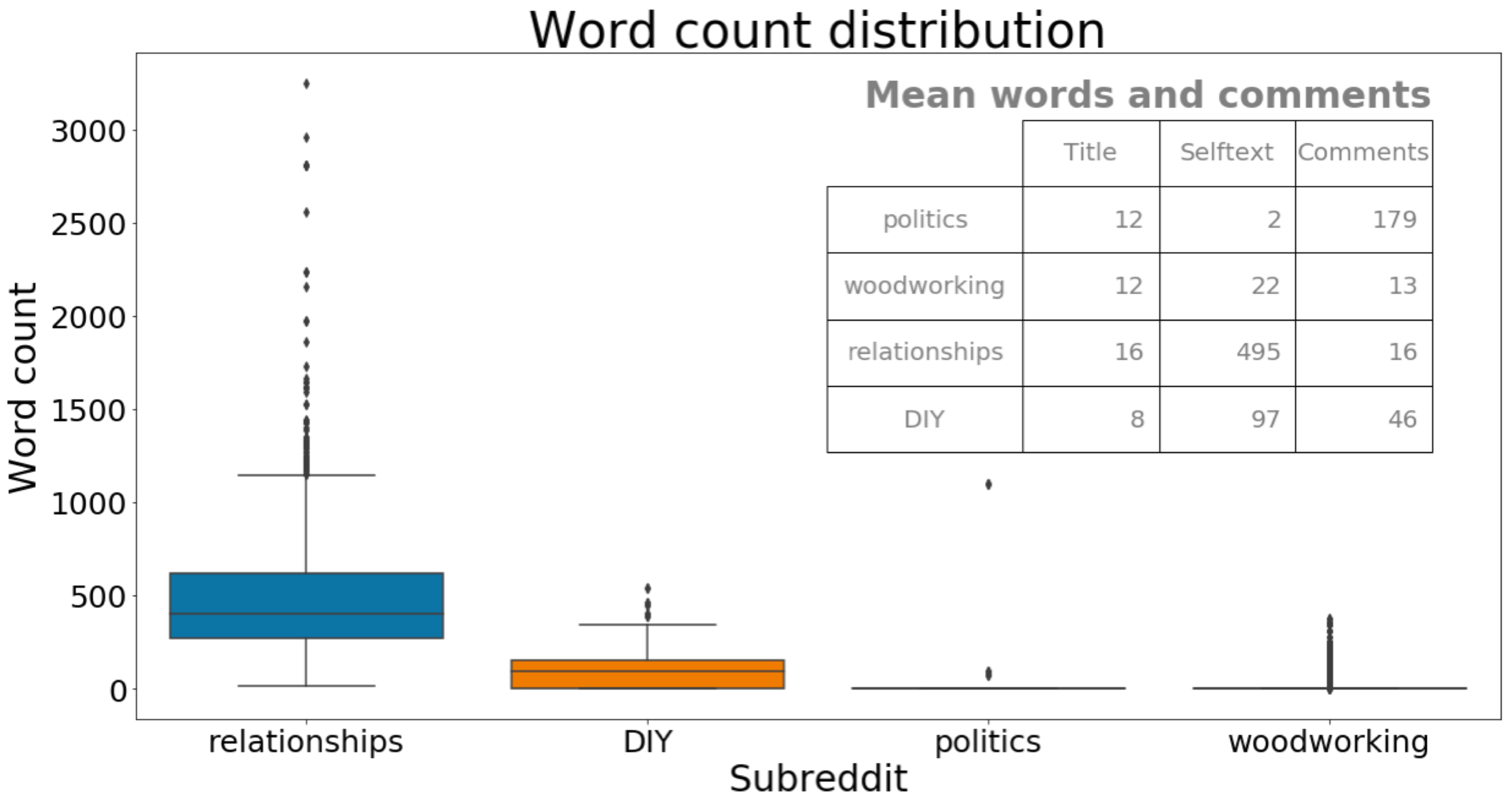
Logistic Regression
(logr)

K Nearest Neighbors
(knn)

Ensemble – logr & bayem
(vote)

For four subreddits with and without comments, 96 models selected via grid search

Post characteristics

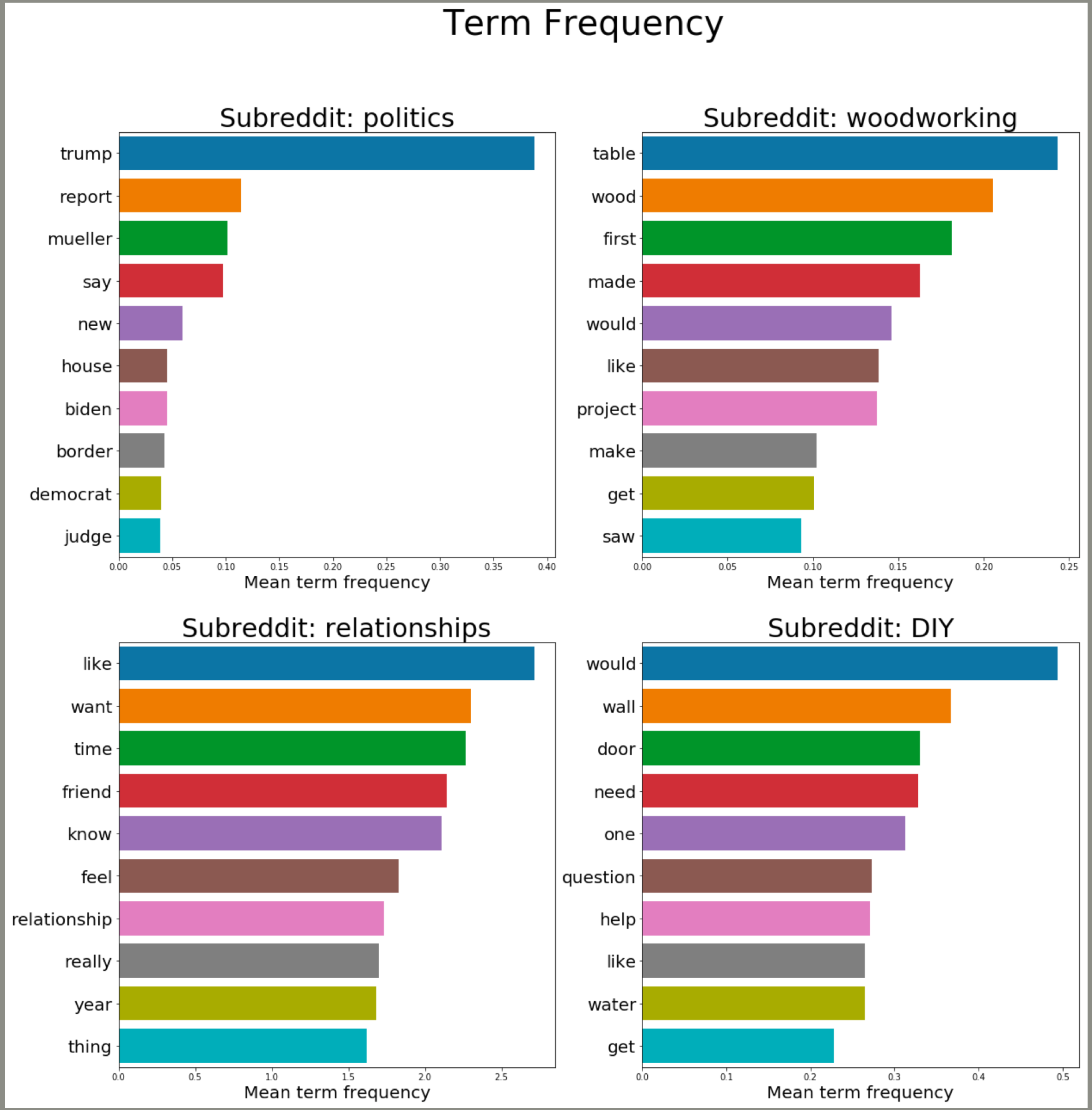


Mean words and comments

	Title	Selftext	Comments
politics	12	2	179
woodworking	12	22	13
relationships	16	495	16
DIY	8	97	46

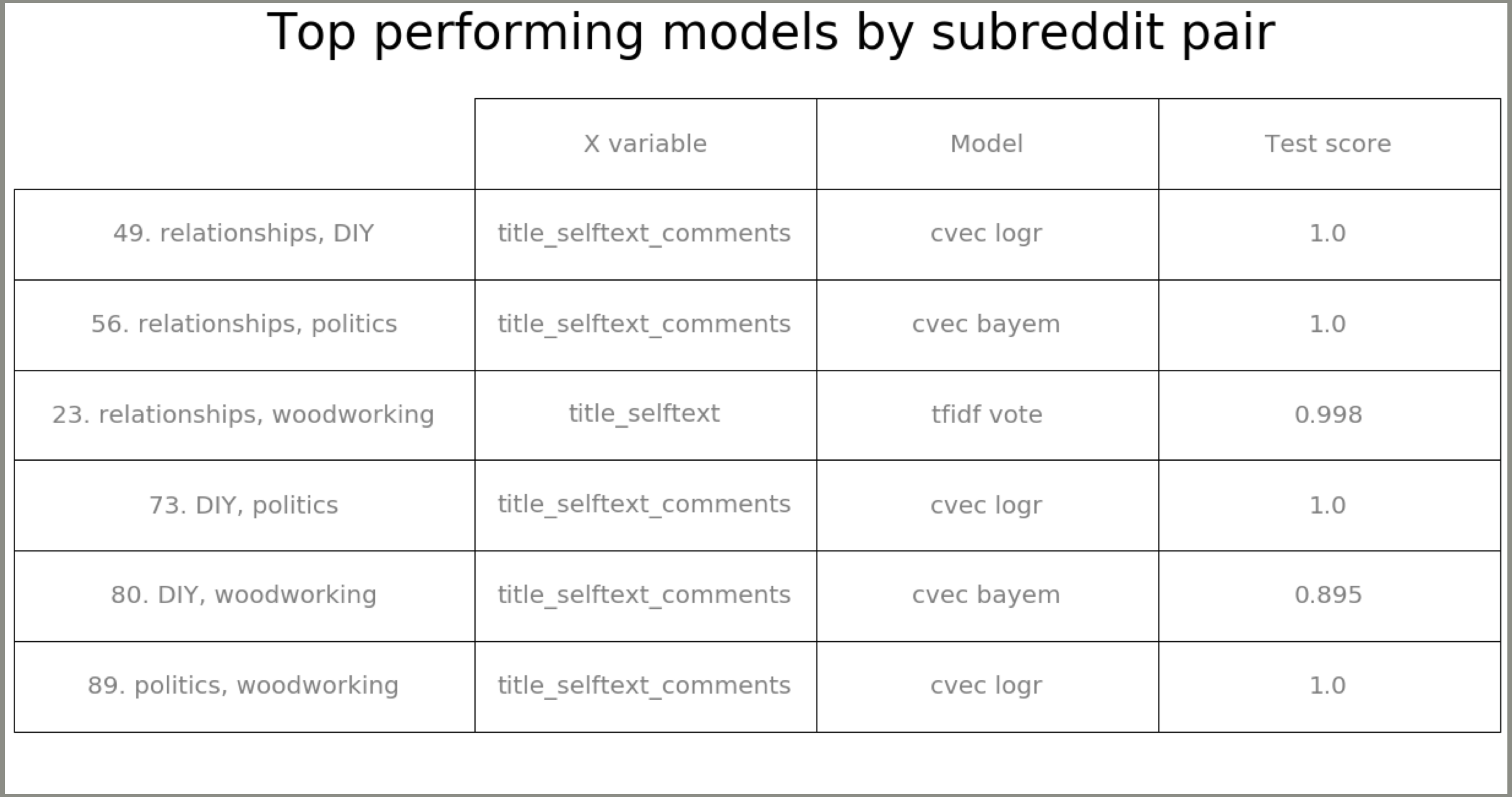
- Observations
  - *Relationship posts have far more words than other categories*
  - *Relationship post lengths are more diverse*
  - *Politics selftext posts are short but they generate many comments*
  - *Woodworking posts are the shortest with fewest comments*

Term frequency



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

Best models

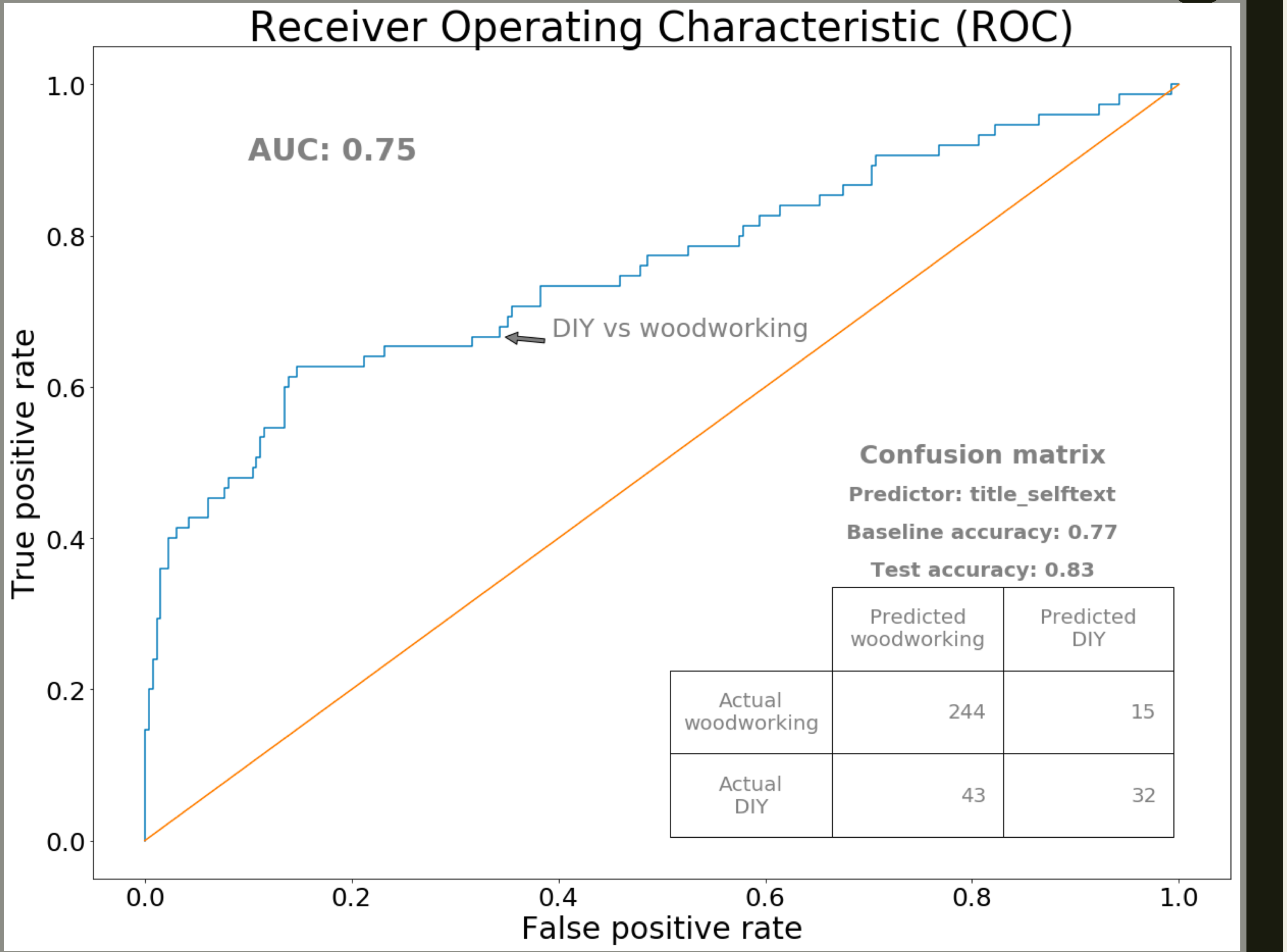


Top performing models by subreddit pair

	X variable	Model	Test score
49. relationships, DIY	title_selftext_comments	cvec logr	1.0
56. relationships, politics	title_selftext_comments	cvec bayem	1.0
23. relationships, woodworking	title_selftext	tfidf vote	0.998
73. DIY, politics	title_selftext_comments	cvec logr	1.0
80. DIY, woodworking	title_selftext_comments	cvec bayem	0.895
89. politics, woodworking	title_selftext_comments	cvec logr	1.0

- High testing accuracy observed across all pairs with some groups at 100%
- Lowest performance is between DIY and woodworking – two similar topics
- Using title, selftext and comments worked best in most cases
- Count Vectorizing and Multinomial Naïve Bayes performed best with DIY and woodworking

Focus on DIY & woodworking



Confusion matrix

Predictor: title_selftext

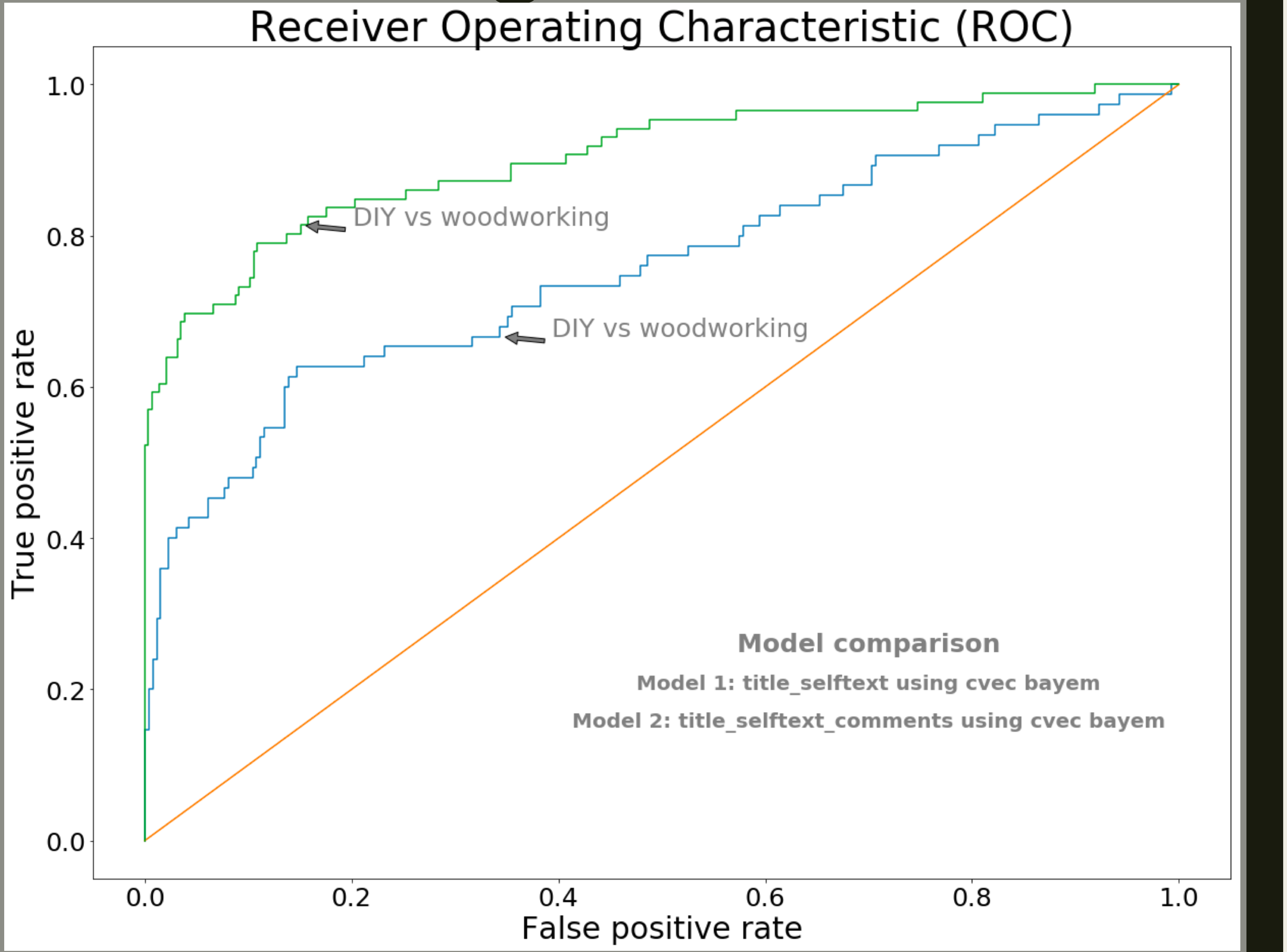
Baseline accuracy: 0.77

Test accuracy: 0.83

	Predicted woodworking	Predicted DIY
Actual woodworking	244	15
Actual DIY	43	32

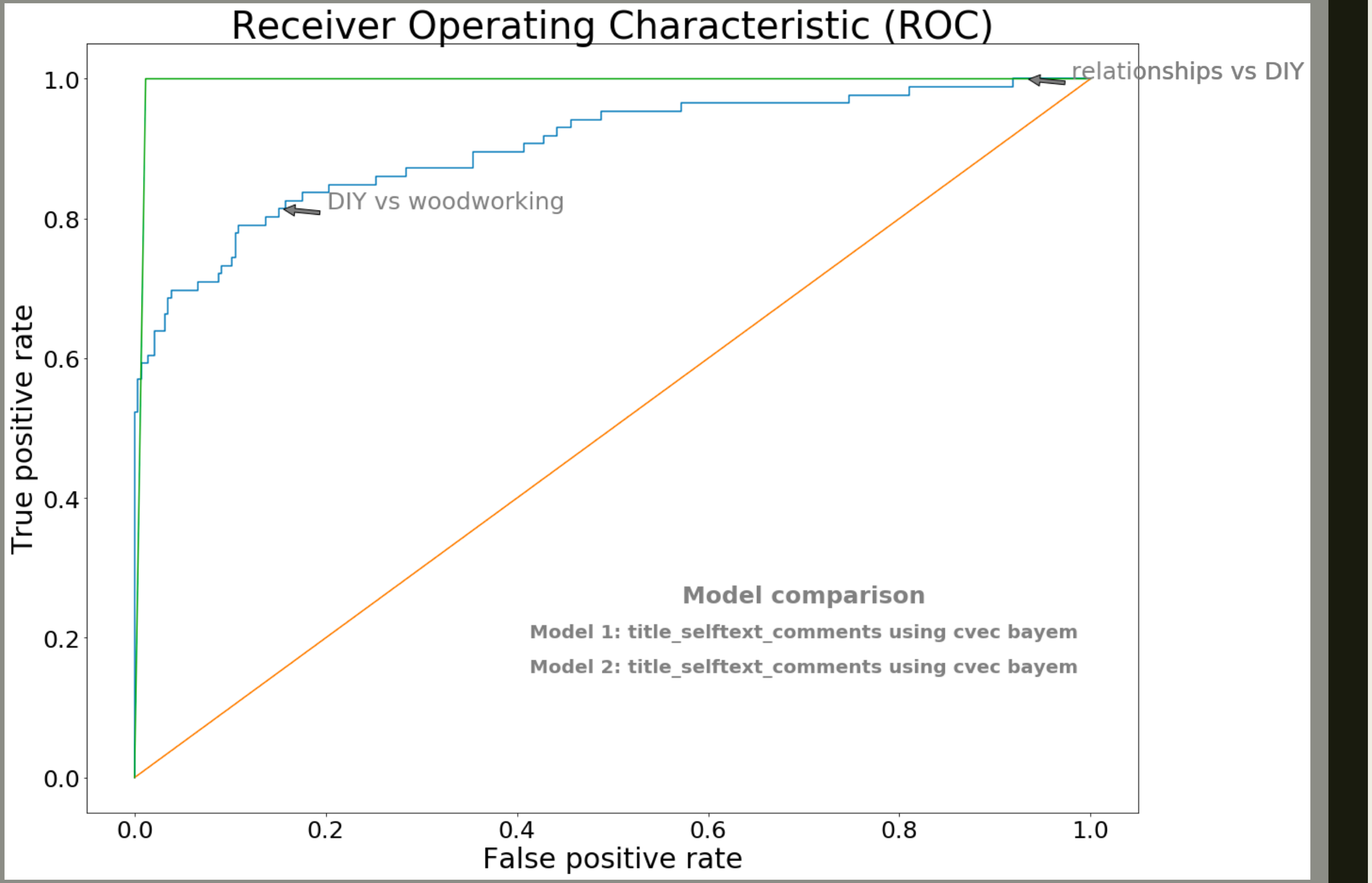
- The most difficult classification challenge

Adding comments



- Adding comments improved the model
- Accuracy improved to 89% from 83%
- Modeling techniques remained constant

Diverse topics



- The model works well with diverse topics
- Compare relationships and DIY to woodworking and DIY
- Model techniques remained the same but a slight improvement can be found with logistic regression

CONCLUSION

This classification approach works well between diverse subreddit topics

Performance is not as good between similar topics