

# PROJECT 3: SUBREDDIT NLP ANALYSIS

Noah Antle



## PROBLEM STATEMENT

- Using two subreddits based on U.S. politics, this project aims to use NLP to predict whether or not a news headline was published by a conservative-leaning or liberal-leaning news source. The two subreddits selected represent both sides of the political spectrum, but cover (roughly) the same news topics. This analysis hopes to understand the relationship between the words used in a headline and the political leaning of its source.

# THE SUBREDDITS



u/ConeCANDY

 **Politics** Joined

[Posts](#) [Rules](#) [Wikipages](#) ▾

 Create Post

 Hot  New  Top ...

**24.9k** Posted by u/DrAnthonyFauci 5 hours ago  10  13  4 & 12 More

John Lewis memorial to replace Confederate monument in Georgia  
[thehill.com/homenews](http://thehill.com/homenews)



283 Comments  Share  Save ...

**8.3k** Posted by u/oranjemania 3 hours ago   

Lawyer for ‘Guy with the Horns and Fur’ Offers to Bring Down Trump by Having ‘QAnon Shaman’ Testify at Impeachment Trial  
[lawandcrime.com/u-s-ca...](http://lawandcrime.com/u-s-ca...)



**About Community** ...

/r/Politics is for news and discussion about U.S. politics.

7.2m Members  61.0k Online 

Created Aug 6, 2007

**Create Post**

Community options

**Welcome!**

Welcome to [r/politics](#)! Please read the wiki before participating.

# THE SUBREDDITS

The screenshot shows the homepage of the **r/Conservative** subreddit. The header features a stylized American flag background with stars. The subreddit name "Conservative" is displayed in large white letters, with "r/Conservative" below it. A "Join" button is visible. The navigation bar includes links for "Posts" (which is underlined), "Official Discord", "User Flair Policy", "What r/Con is not", and "Full Rules".

On the left, there's a "Create Post" button and a sorting menu with "Hot" (selected), "New", "Top", and "...". Below this, a pinned post by u/Jibrish titled "/r/Conservative Official Political Discord" has 2.3k upvotes and 78 comments. It was posted 3 months ago and crossposted from /r/wall... with 762 upvotes and 501 comments.

Another post by u/TimbitGaming has 34.7k upvotes and is titled "Satire - Flaired Users Only New SEC Rule: Wall Street Will Now Only Allow Traders Who Wear A Top Hat And Monocle And Carry Around Giant Pads Of Money".

The right sidebar contains an "About Community" section with 667k members and 8.6k online users, created on Jan 25, 2008. It also includes a "Create Post" button, "Community options" dropdown, and a "Filter by flair" section with "Satire - Flaired Users Only" and "Satire" buttons. There are also small thumbnail images of posts at the bottom.

# THE DATA

## r/Politics

3. Post titles must be the exact headline ^  
from the article.

[Full rule here.](#)

A title must be comprised only of the  
copied and pasted headline of the  
article.

Do not add, remove, or change words.

## r/Conservative

6. Title not from article ^

Submission headlines must match the  
article headline or quote the article.

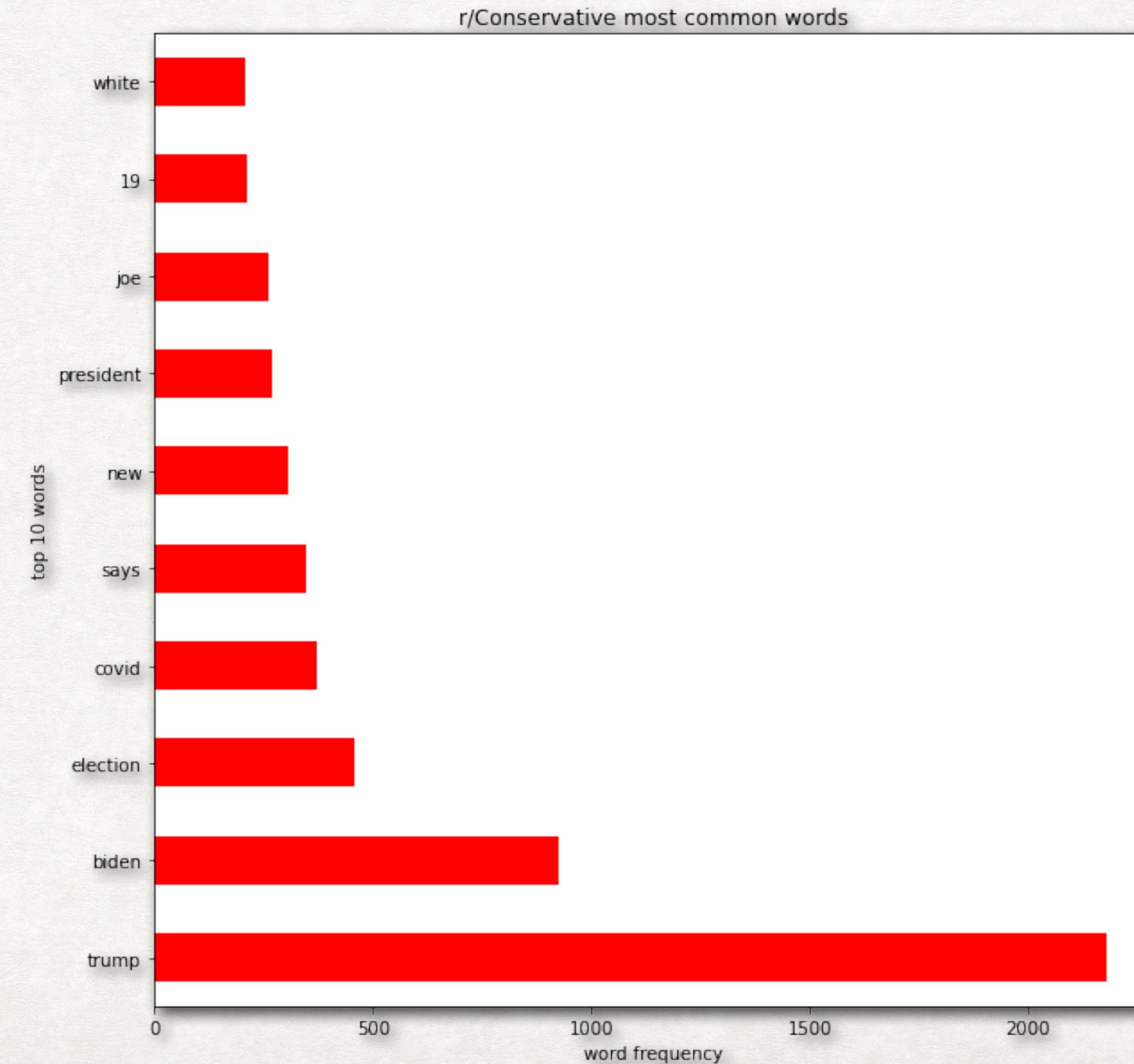
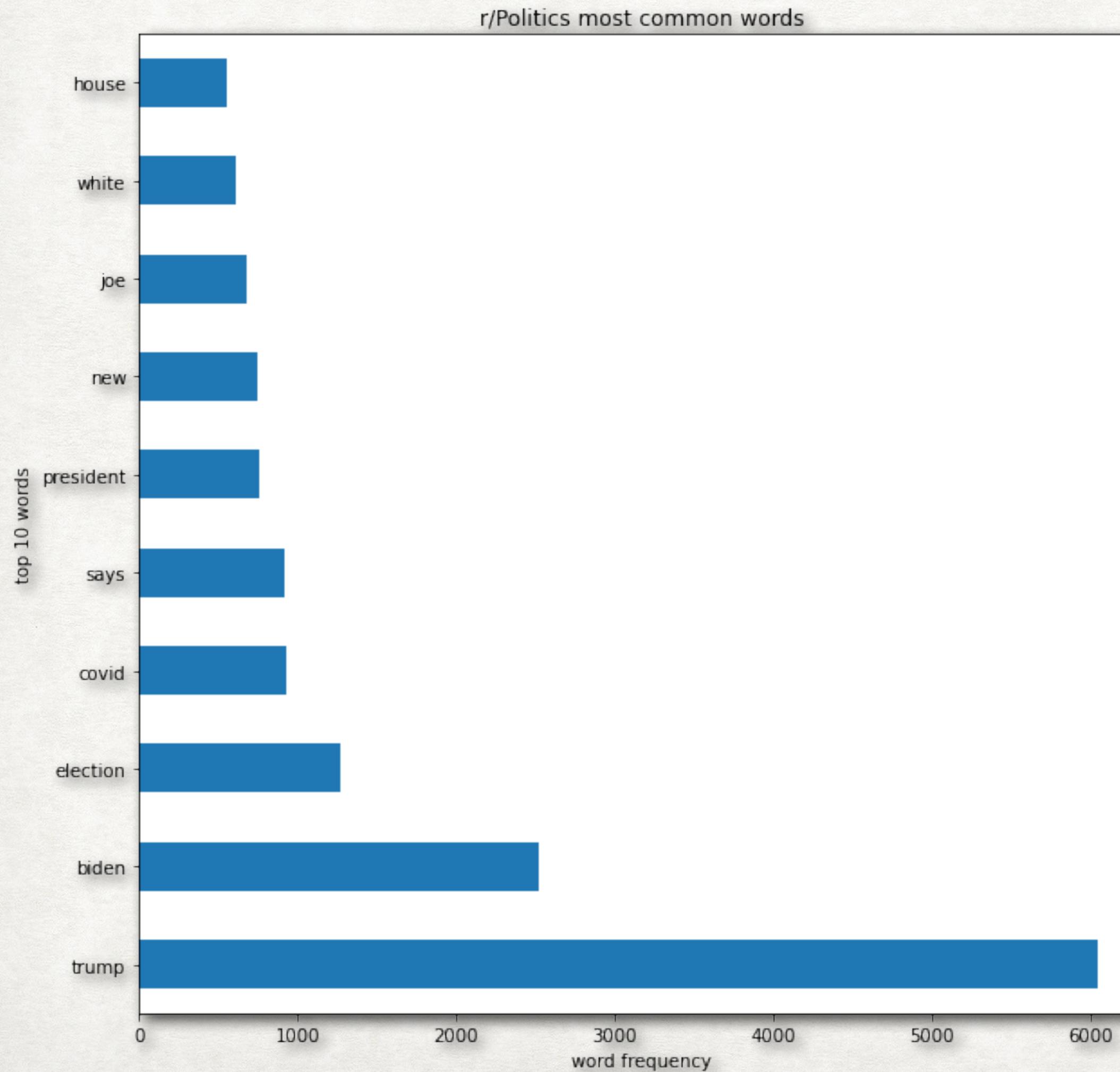
# THE DATA

- Data was collected with up to 100 posts from each of the last 200 days from each subreddit:
  - r/Politics = 19,807 posts
  - r/Conservative = 16,164 posts
- ALL data collected was strictly post titles.
  - Why is this good?
    - More uniform text, generally following newspaper publishing standards (No emojis, few typos, no urls, etc.)
  - Why is this bad?
    - Less text than a full-body text submission, so I need to acquire a huge number of samples.

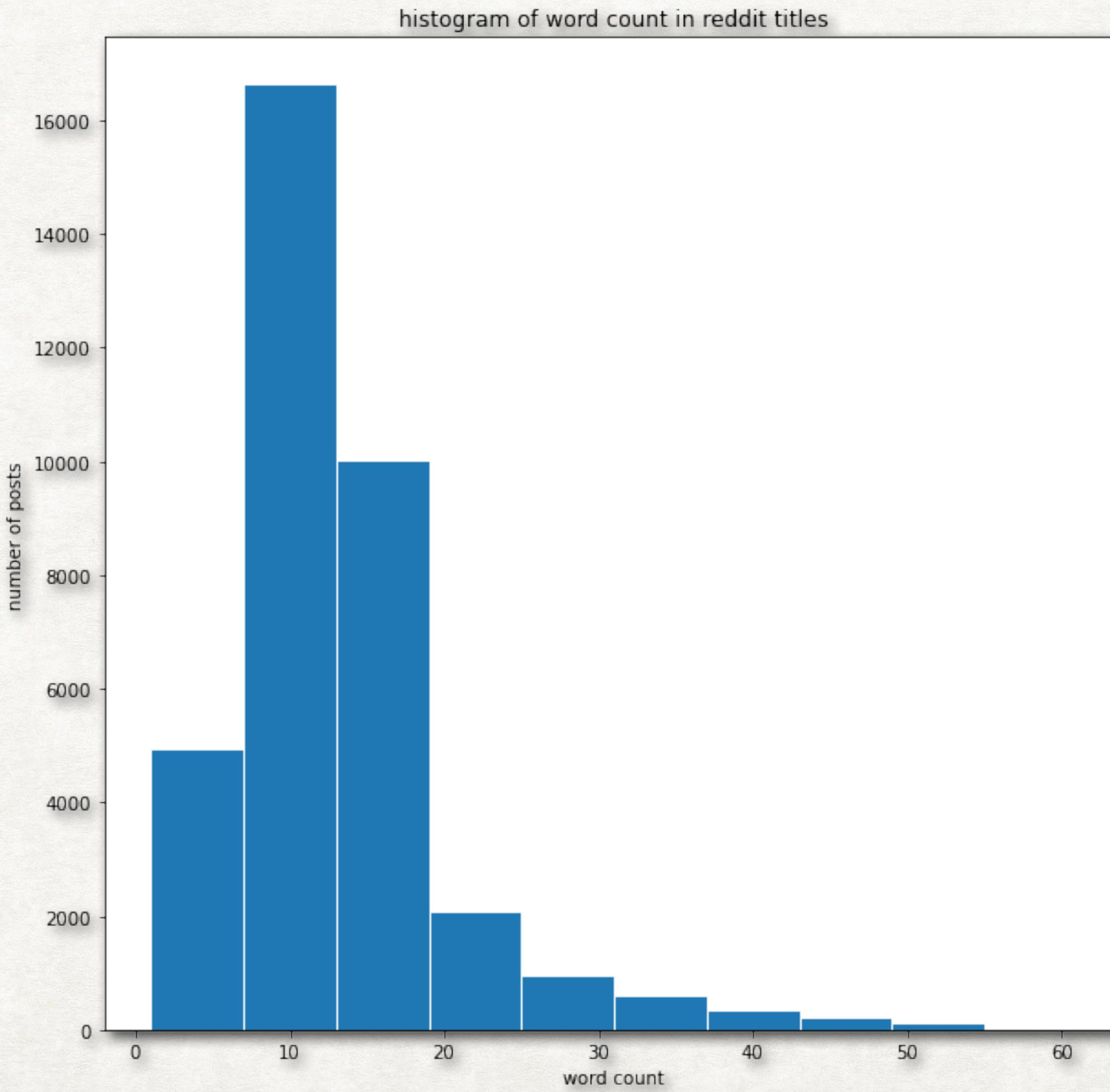
# THE DATA

## THE BIG PROBLEM

- There is significant overlap between the content of the two subreddits.



# THE DATA



# THE MODELS

- 5 models were fit:
  1. Logistic Regression
  2. k-Nearest Neighbors
  3. Random Forest
  4. Multinomial Naive Bayes
  5. GradientBoosting
- Each model was tuned through GridSearch and the text data was vectorized through TfIdfVectorizer.
- For clarity, the “0” class corresponds to r/Conservative, and the “1” class corresponds to r/Politics.

# THE NULL MODEL

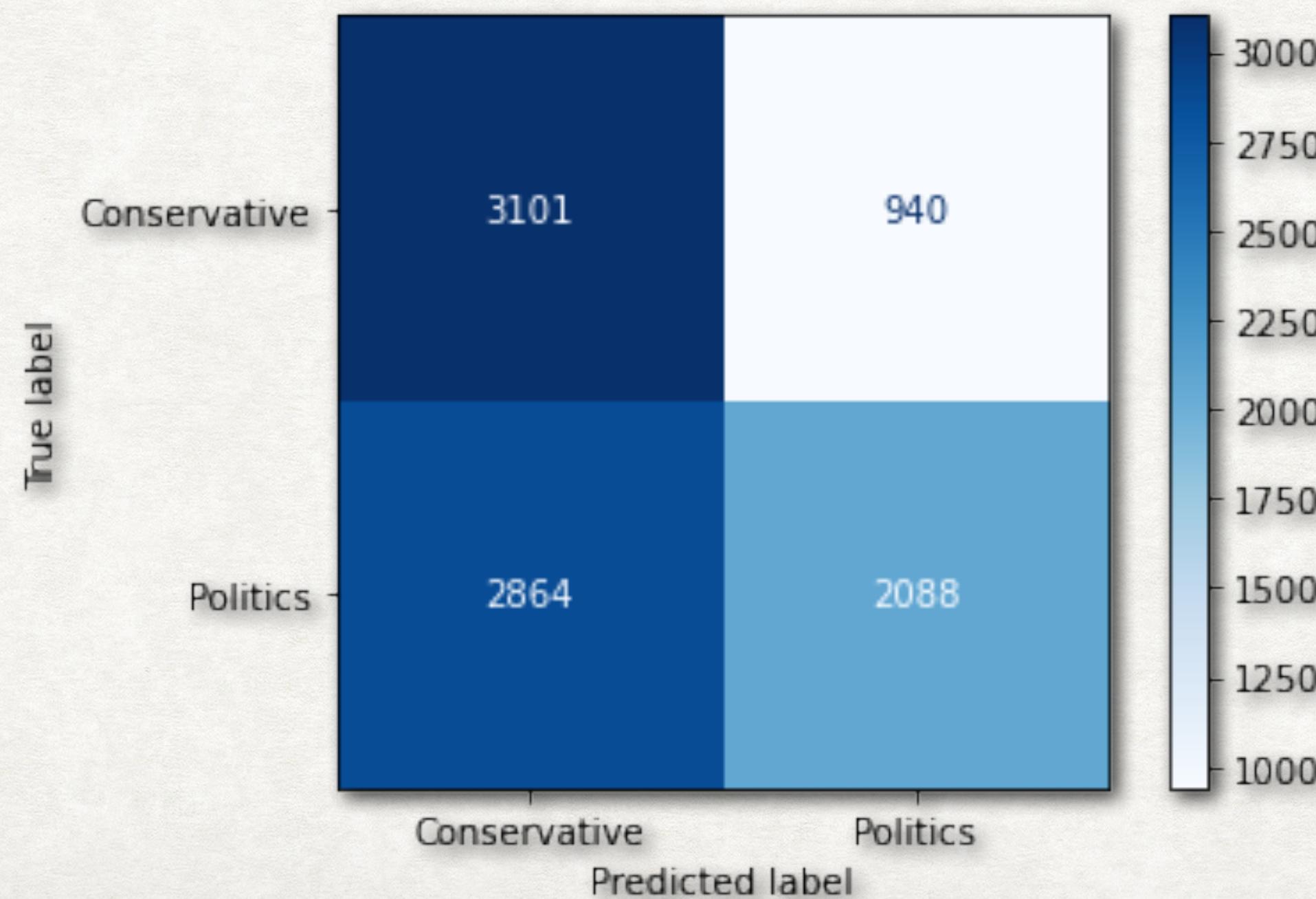
- r/Politics is the majority class in the data, comprising 55% of the observations.
- The baseline model would always predict that a post came from the Politics subreddit, and it would be correct 55% of the time.

```
In [31]: 1 y.value_counts(normalize = True)
```

```
1    0.550638
0    0.449362
Name: subreddit, dtype: float64
```

# WHICH MODEL DID POORLY?

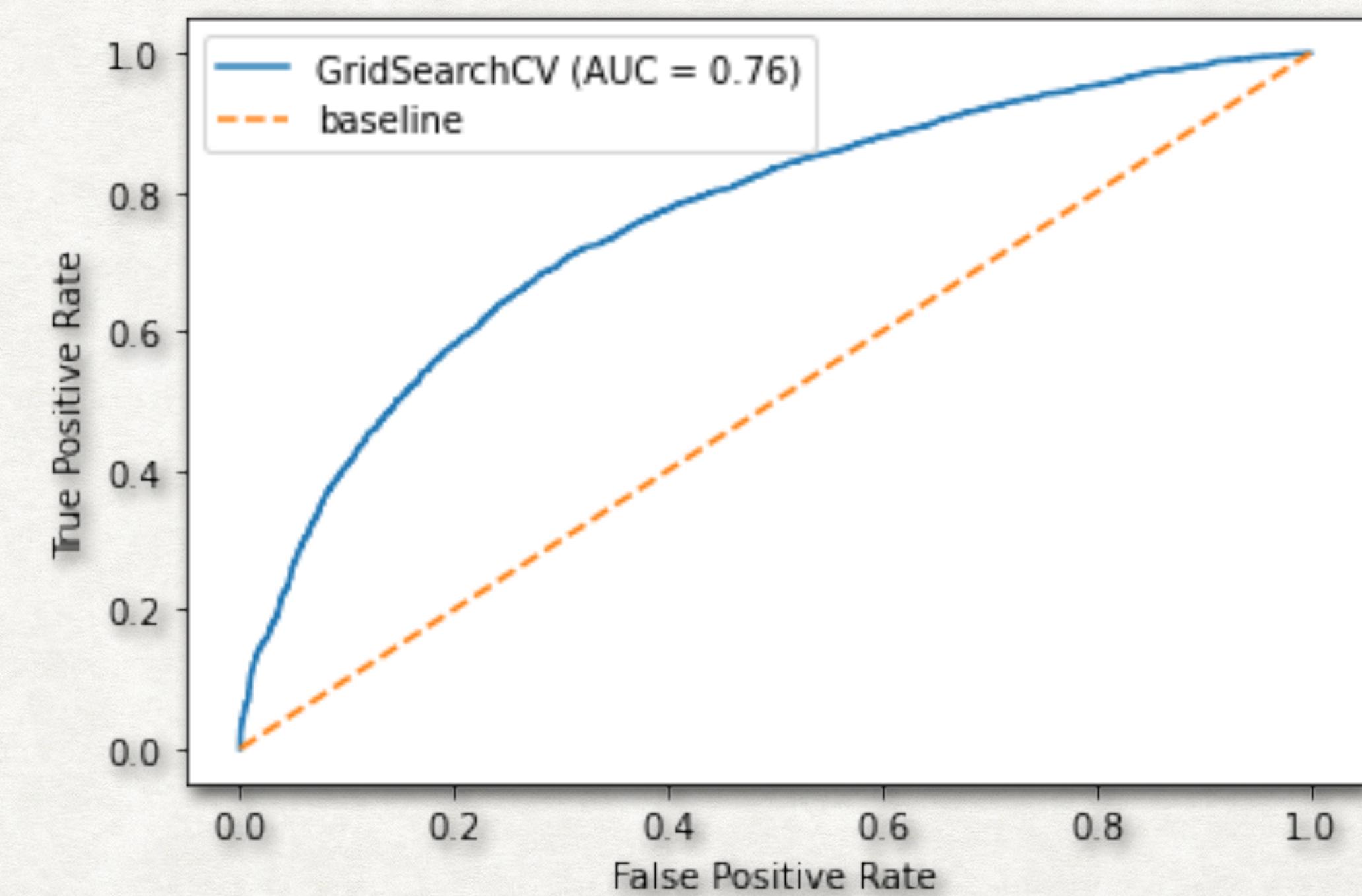
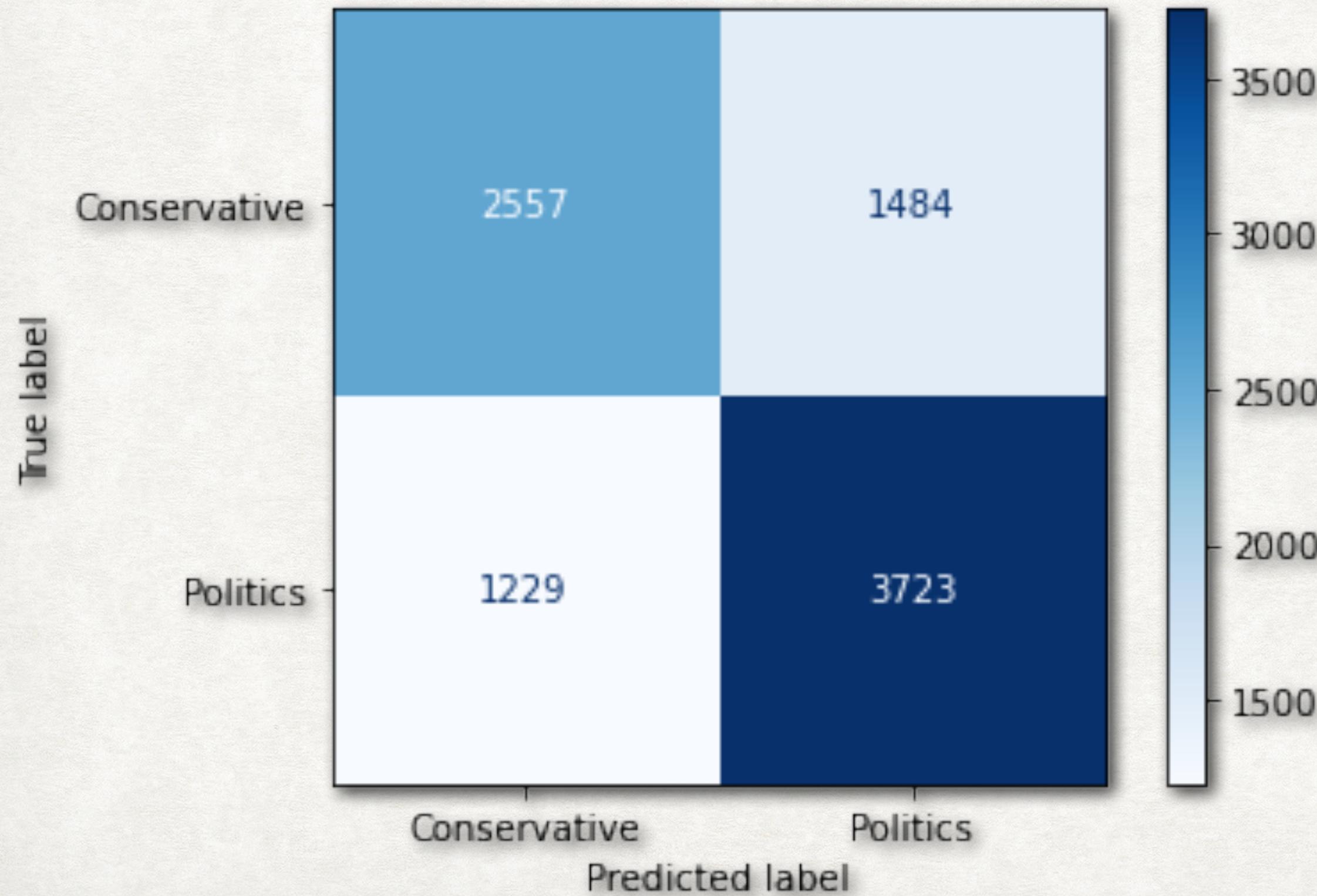
- kNN: The best trained kNN model only achieved 57.7% accuracy (only 2.7% higher than the null model).
- My speculation is that the poor performance of the kNN model is due to the extreme overlap between the two subreddits (even with stop-words accounted for).



```
3 recall = recall_score(y_test, knn_preds)
4 recall
0.42164781906300486
```

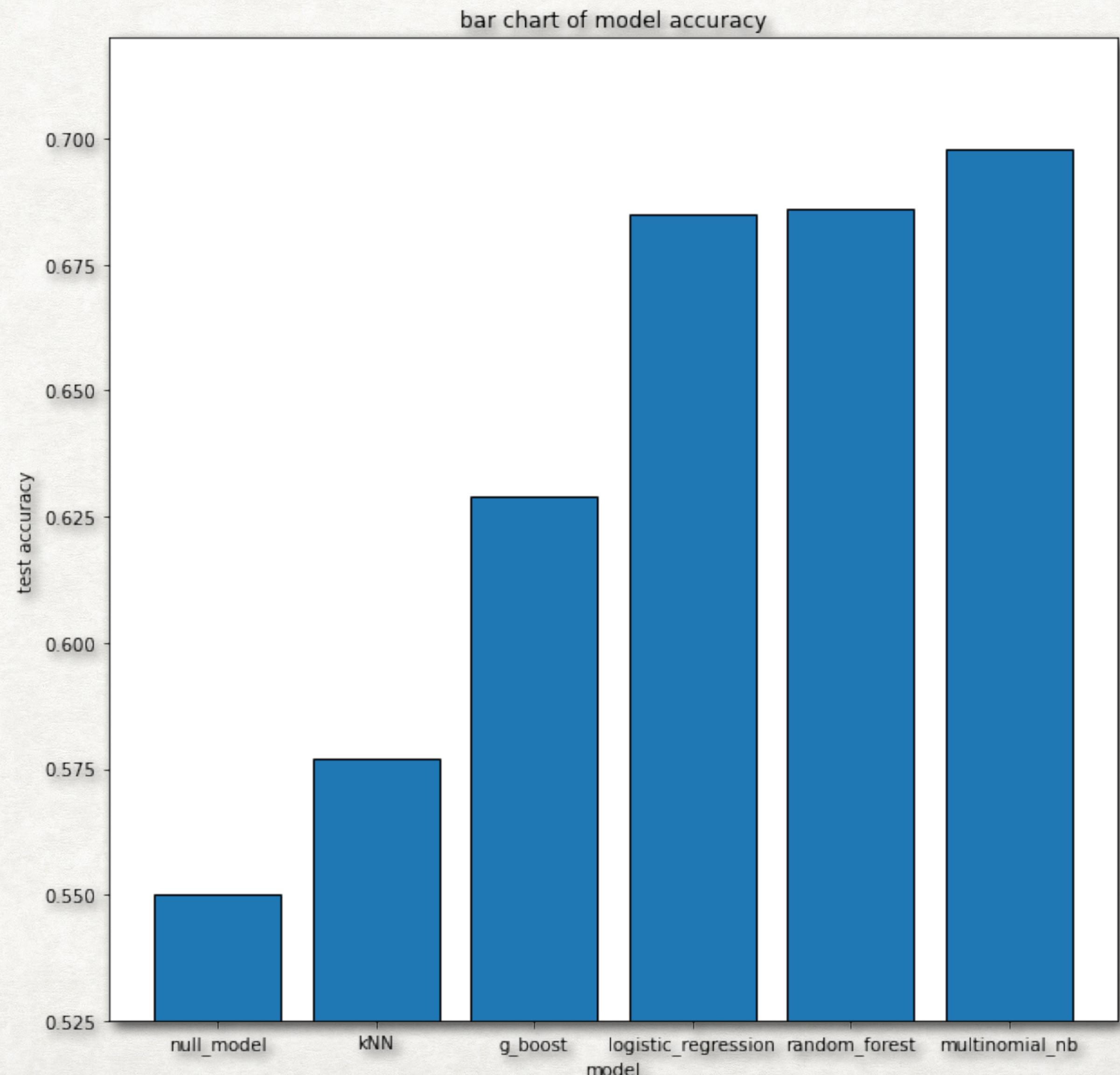
# WHICH MODEL DID WELL?

- Multinomial Naive Bayes: The best trained MultinomialNB model achieved 69.8% accuracy on the test data (14.8% higher than the null model).



# ALL THE MODELS!

	model	test_accuracy
0	null_model	0.550
1	kNN	0.577
2	g_boost	0.629
3	logistic_regression	0.685
4	random_forest	0.686
5	multinomial_nb	0.698



# MISSION ACCOMPLISHED?

- Sort of. Overall, I'm pretty happy with 70% accuracy.
- More could be done to tune hyper-parameters (particularly in the boost models).
- Feature engineering and incorporating non-text data.

