Classifying Subreddits

Space, AskScience & AskScienceFiction

Problem statement

Correctly classify posts from the “AskScienceFiction” subreddit vs. (combined) “space” and “askscience”

Data problem

Collected posts for 10 days:

	count	percentage
AskScienceFiction	1,197	41.4%
space	957	33.1%
askscience	737	25.5%

Top-Level Results

TfidfVectorizer & MultinomialNB model was best:

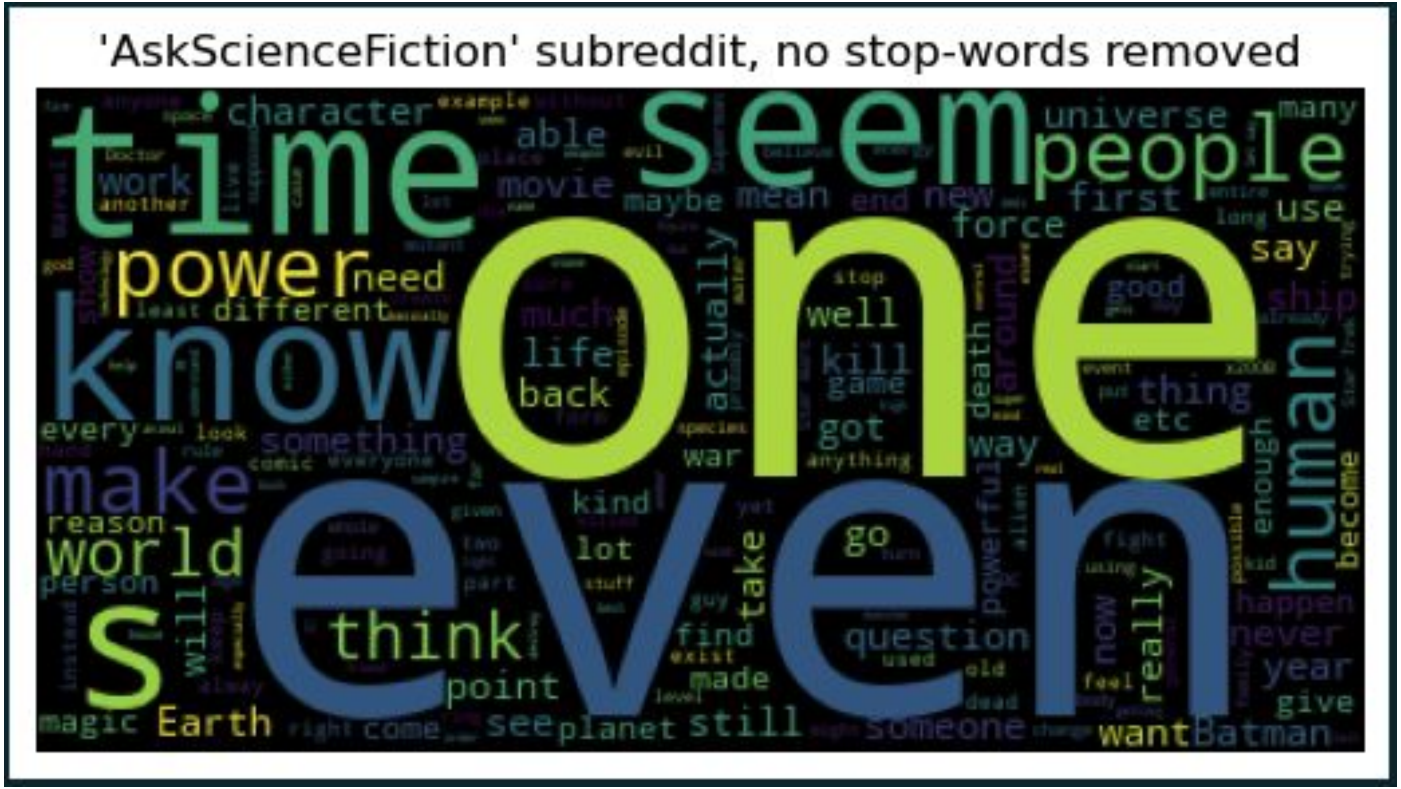
	score
Balanced accuracy score	0.904701
F-1 score	0.890323
recall	0.862500
precision	0.920000

Data cleaning & prep

- Making `subreddit` consistent
- Missing `selftext`

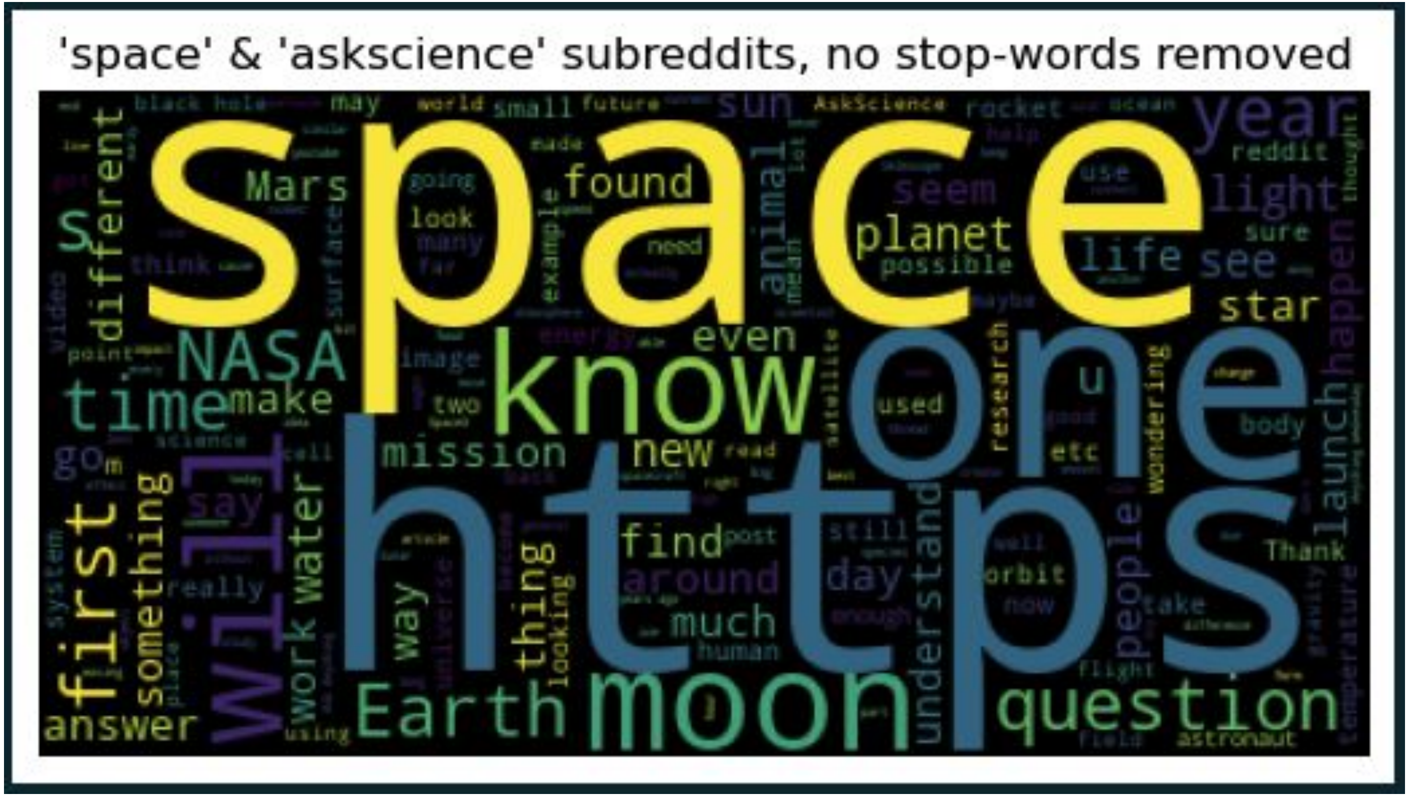
askscience	15.6%
AskScienceFiction	24.8%
space	81.9%

Common words in the positive class



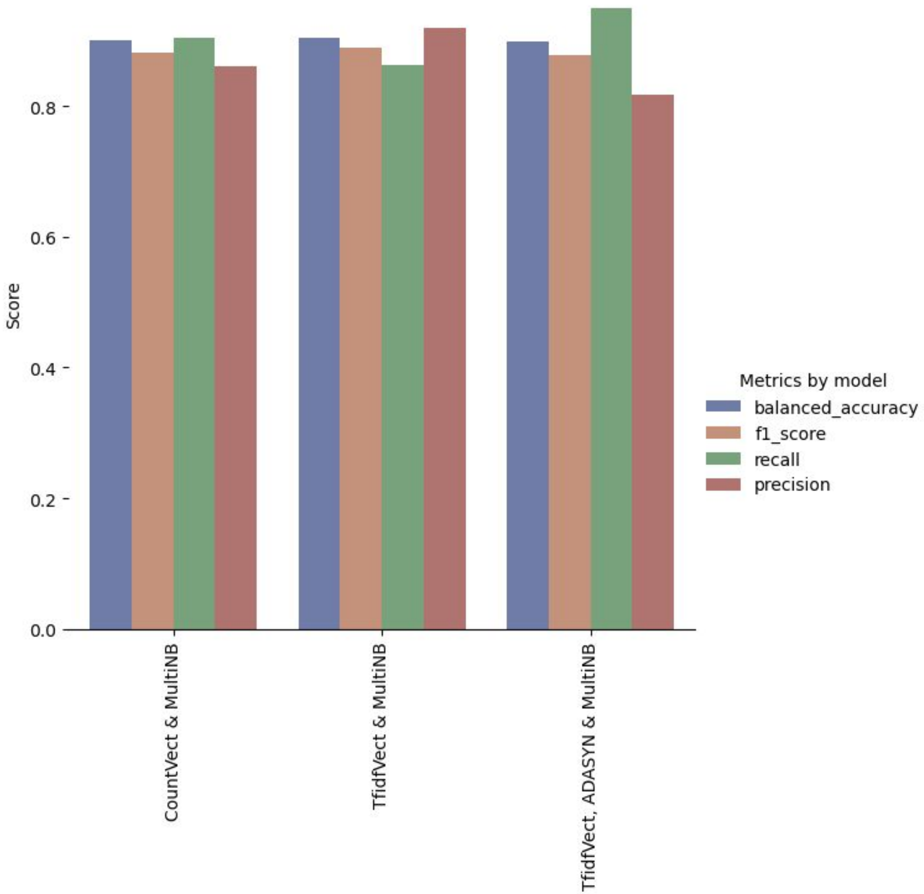
'AskScienceFiction' subreddit, no stop-words removed

Common words in the negative class



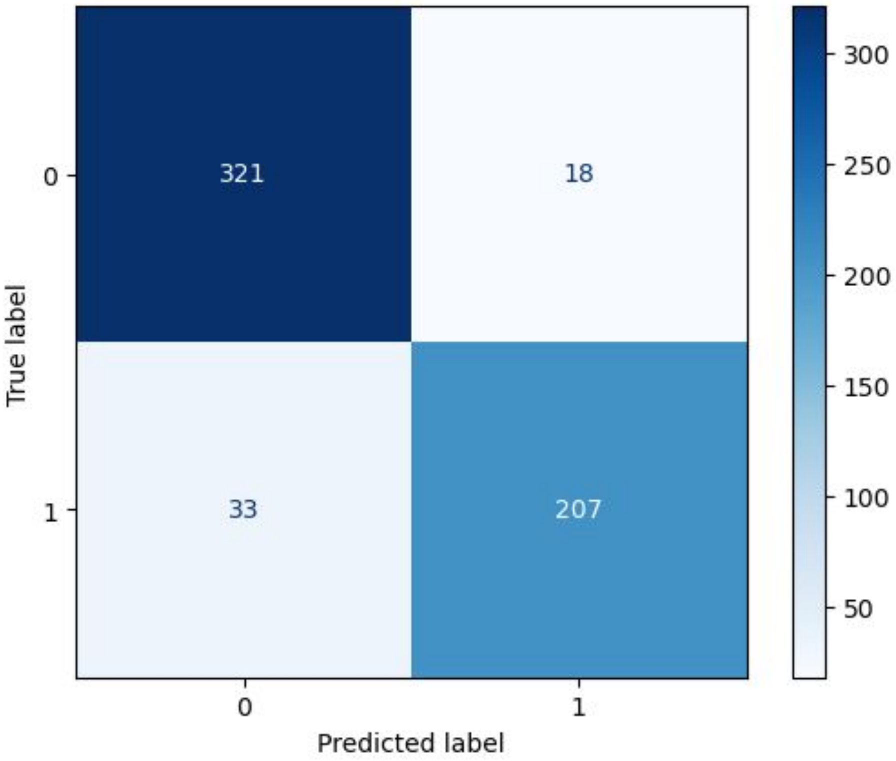
'space' & 'askscience' subreddits, no stop-words removed

Nine models



Score
0.0
0.2
0.4
0.6
0.8
CountVect & MultiNB
TfidfVect & MultiNB
TfidfVect, ADASYN & MultiNB
Metrics by model
balanced_accuracy
f1_score
recall
precision

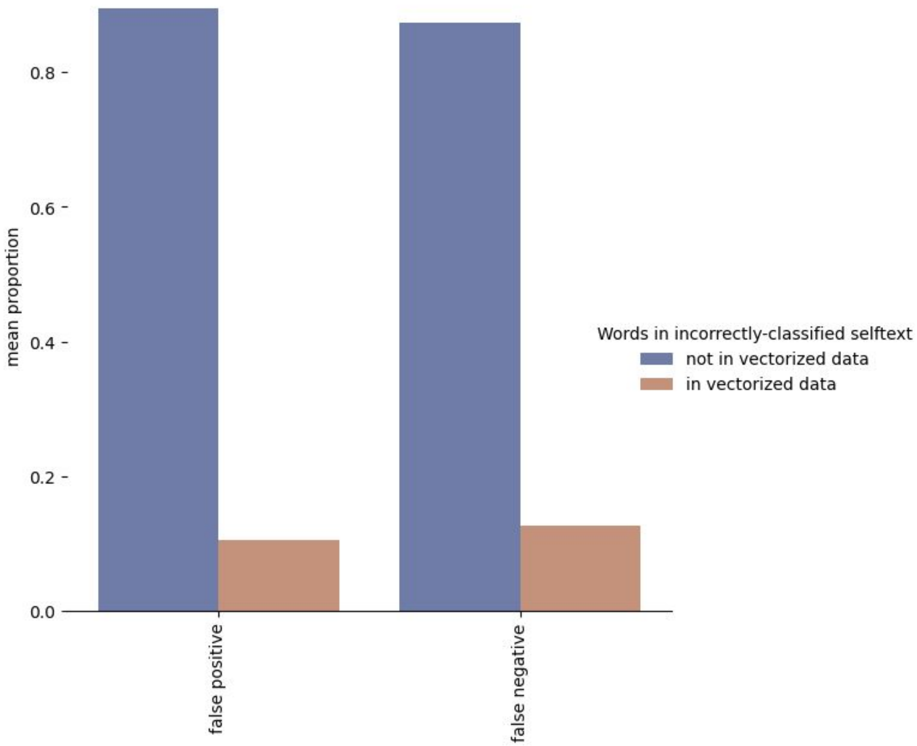
Confusion matrix



Breakdown of false positives

	Proportion of negative class	Proportion of false positives
space	56.5%	11.1%
askscience	43.5%	88.9%

Lots of words don’t make it into the model



Future work

- Re-run models using `title` instead of `selftext` to see if there’s any difference
- Try models that ended up with higher params