



Relationship
advice

&

Petty revenge



NLP Classification Demo

April 2021

Problem

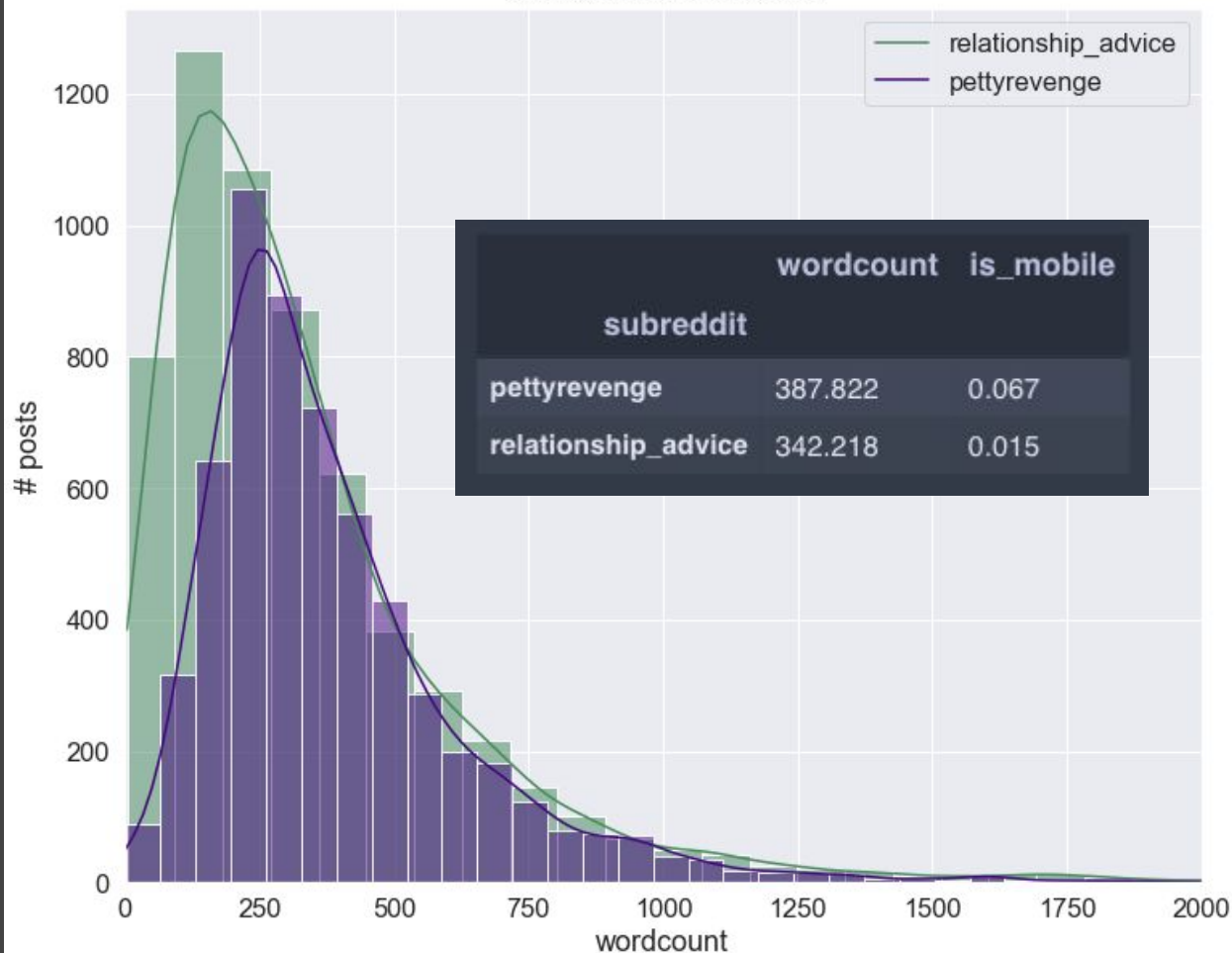
A disgruntled moderator left his forums in ruins as you wrested the moderator badge from his chest.

We want to come up with methods for classifying text blocks from these two subreddits back into their proper homes. These methods can be applied to other NLP tasks in the future.

Data description

- **Source: Reddit API pushshift.io**
- **Relationship_advice: 5M members → 6,093 (51%)**
- **Pettyrevenge: 1M members → 5,907 (49%)**
- **Sampled over same timeframe → 2016 to present**

Wordcount distribution

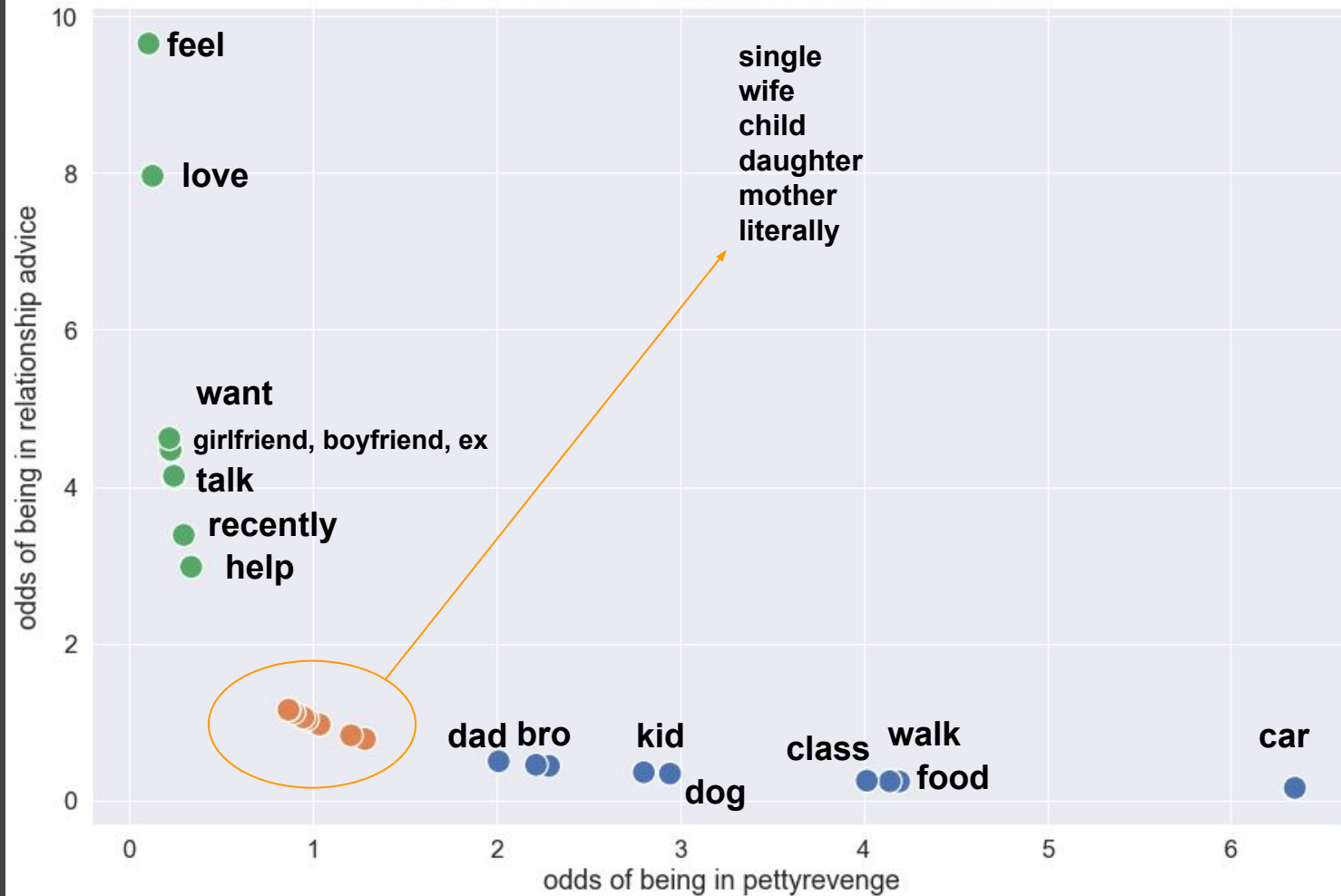


**PR average
wordcount: 388**

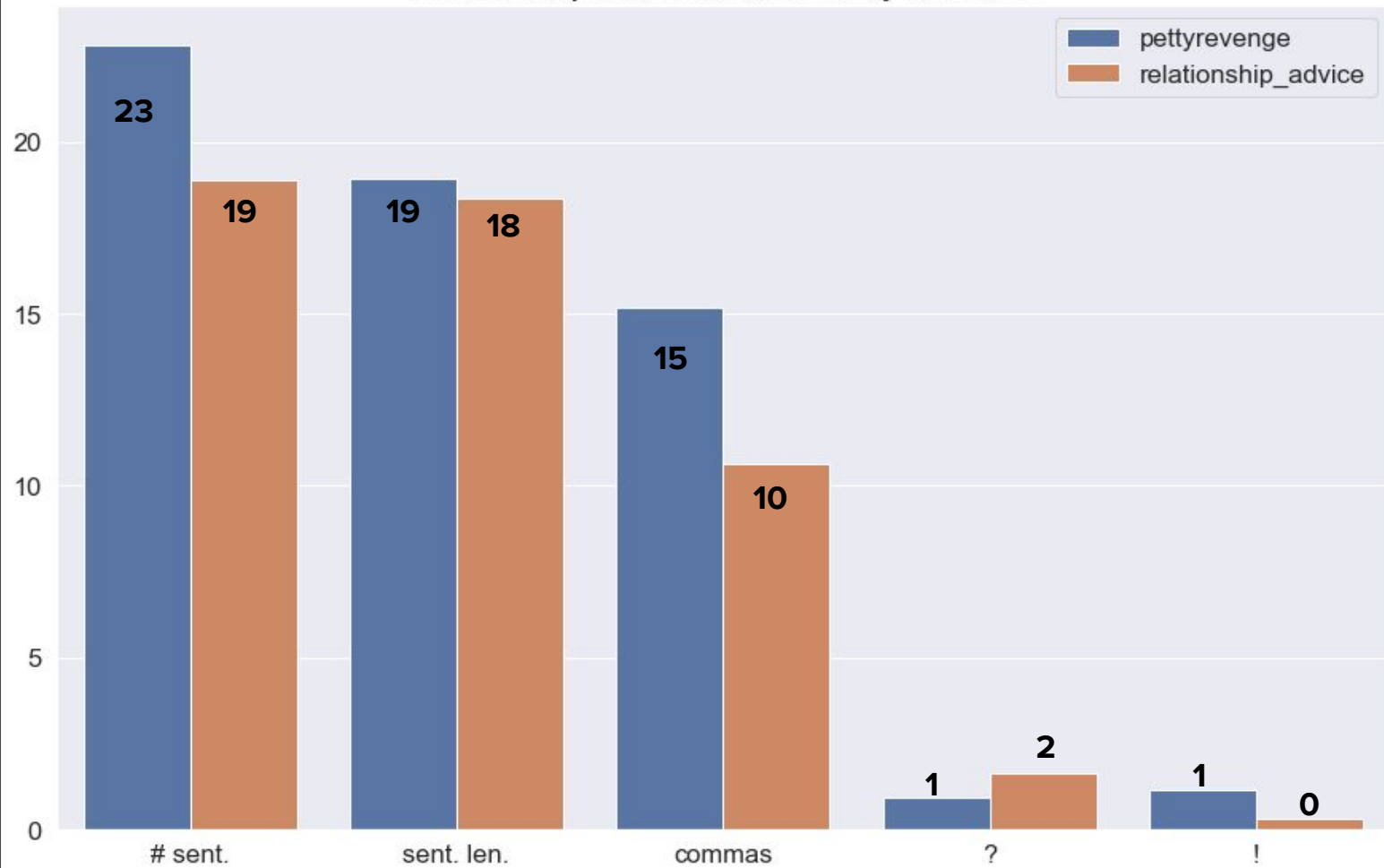
**RA average
wordcount: 342**

**A mobile post is 4.5x
more likely to be from
PR than RA**

Ratio of common words across both subreddits



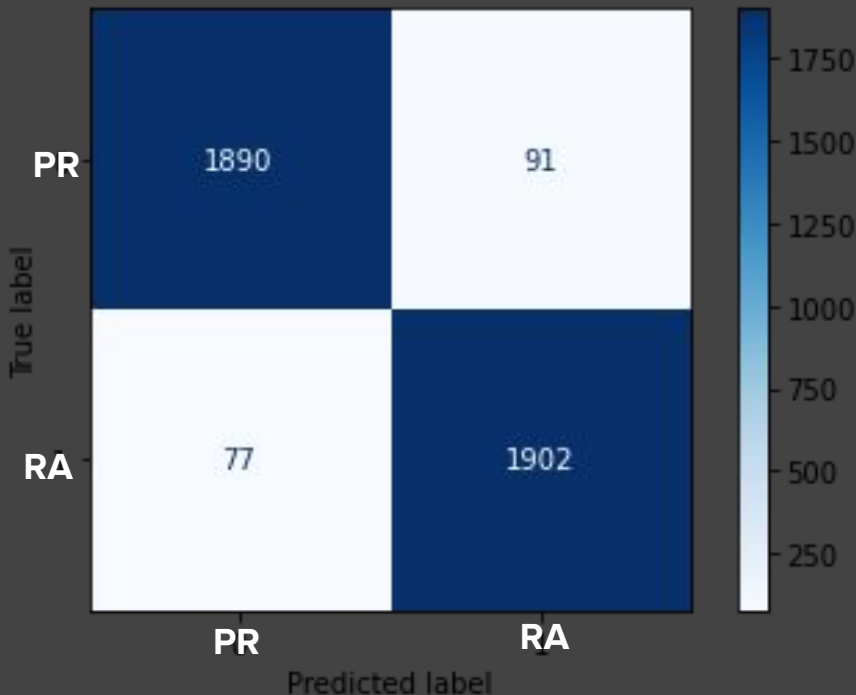
Post format: punctuation normalized by wordcount



Pre-Processing

1. Remove all URLs
2. Remove all punctuation
3. Remove all Chinese characters which is a thing that passes `isalpha()`...
4. Lemmatize the text
5. Custom stop_words
 - a. 'English'
 - b. Lemmatized 'english'
 - c. 'Relationship', 'advice', 'petty', 'revenge'

Model 1: Multinomial Bayes with CountVectorizer



- RandomSearchCV
- Best parameters:
 - max_df=100
 - Max_features = 6112
 - Ngram_range (1,3)
 - Mnb_alpha = .3077

Scores

F1: 0.9587

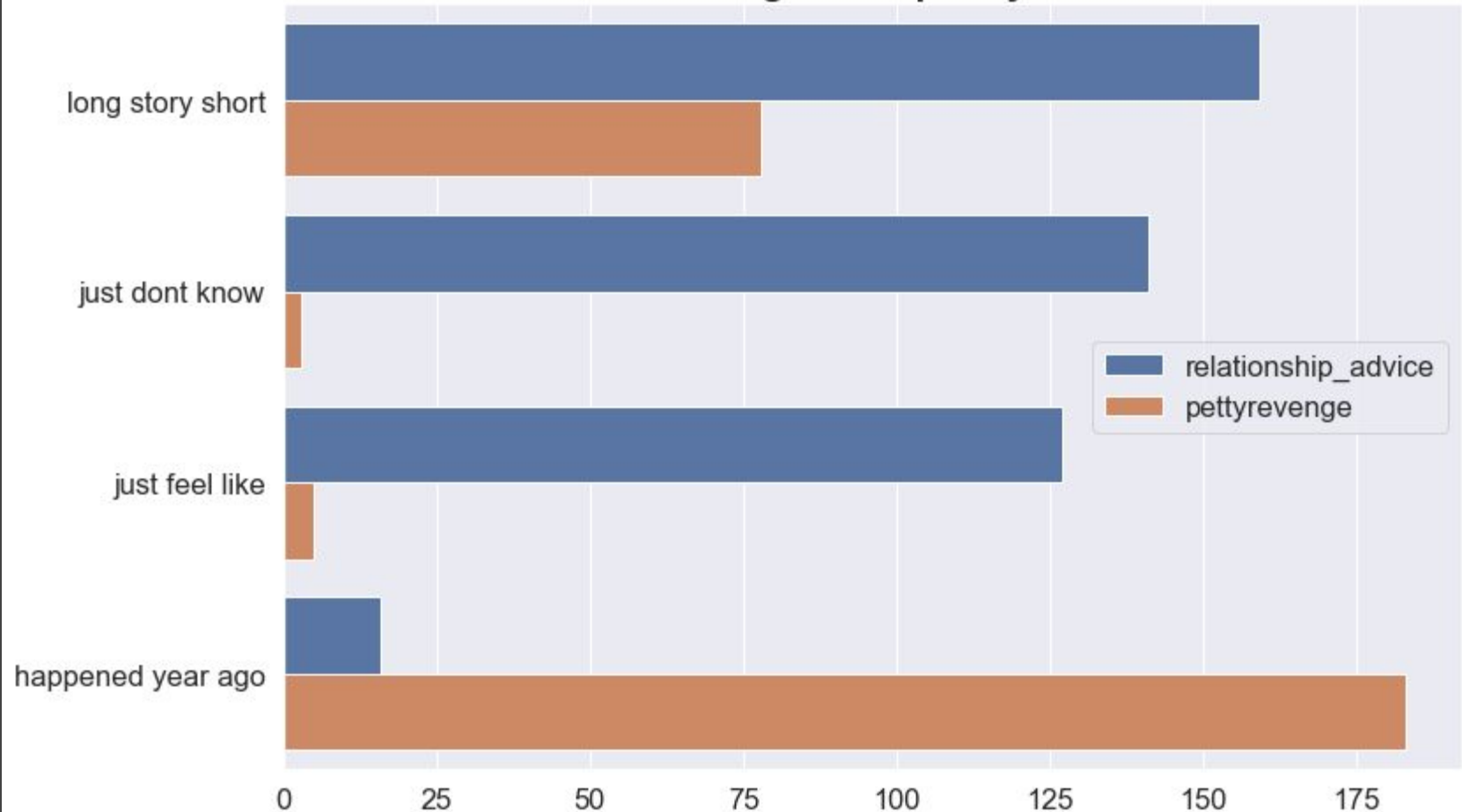
Accuracy: 0.9576

Sensitivity: 0.9611

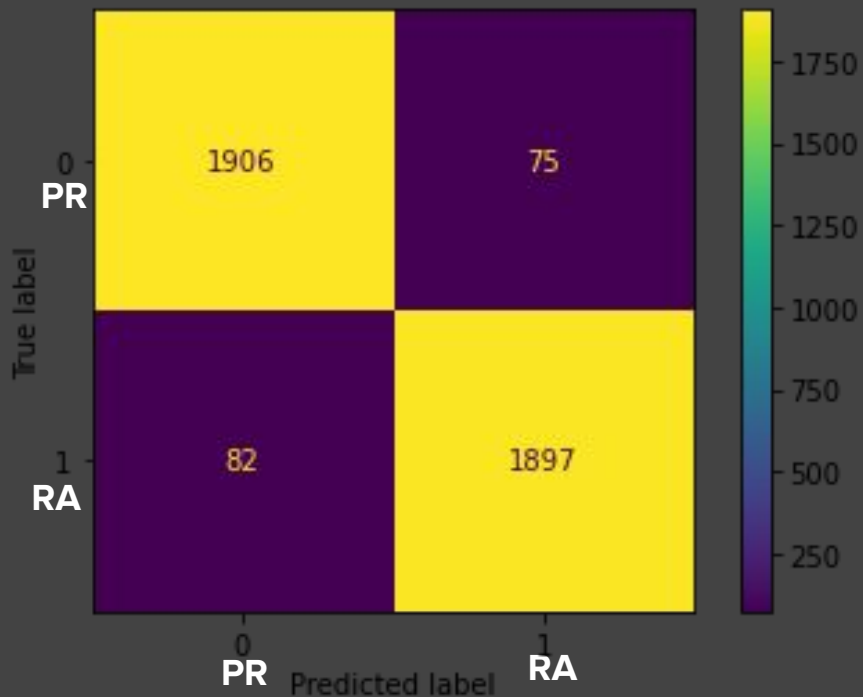
Precision: 0.9543

Specificity: 0.9541

Trigram frequency



Model 2: Support Vector Classification with Tf-Idf



- RandomSearchCV
- Best parameters:
 - max_df=0.93
 - Max_features = 4739
 - Ngram_range (1,3)
 - C = 0.63

Scores

Accuracy: 0.9604

Sensitivity: 0.9586

Precision: 0.9620

Specificity: 0.9621

F1: 0.9603

Conclusions



- Many 'most common words' were the shared across subreddits, but the relative frequencies among these words could still guide classification
- There are minor structural differences (wordcount, comma usage), but not with enough differentiation to base decisions on it
- SVC and MNB both yield competitive models, with SVC having the superior F1 score (and specificity and sensitivity)
- Coefficients for the best model (SVC) are not interpretable. No idea what's going on in there. But our word frequency EDA can give us an idea of feature importance.

