Classifying texts from feminist and anti-feminist forums

Faye Fong
March 3, 2024

Introduction

Online forums are a convenient large database of social discourse.

We chose to examine two subreddits which represent opposing perspectives on societal expectations for gender norms:

- ‘TwoXChromosomes’
- ‘Men’s Rights’

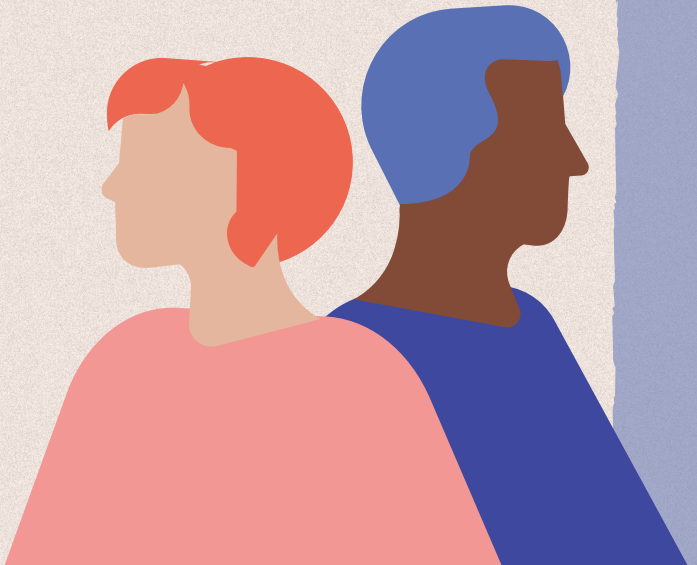


Table of contents

01 / 02 Data collection and cleaning

Playing by API rules

Filters: New, Hot, Top, Controversial

Best Features

Combined title and self text

Confidently drop null values and duplicates

Baseline model

0.504	**TwoXChromosomes**
0.4959	**Men's Rights**

03 Tuning transformers for a Naive Bayes model

CountVectorizer

Stop words

Minimum document frequency

N-gram

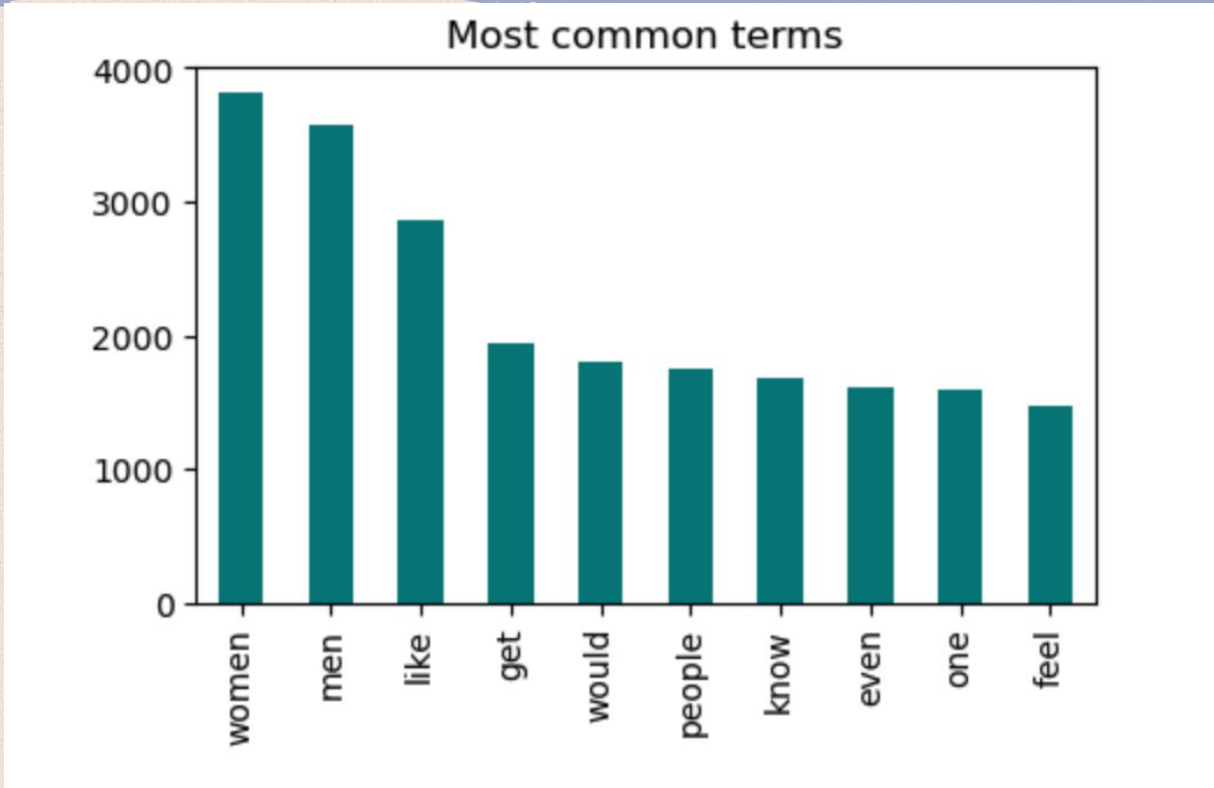
Maximum number of features

Multinomial Naive Bayes

Caveat: Assumes independence of terms

Fast
Simple
Makes good predictions

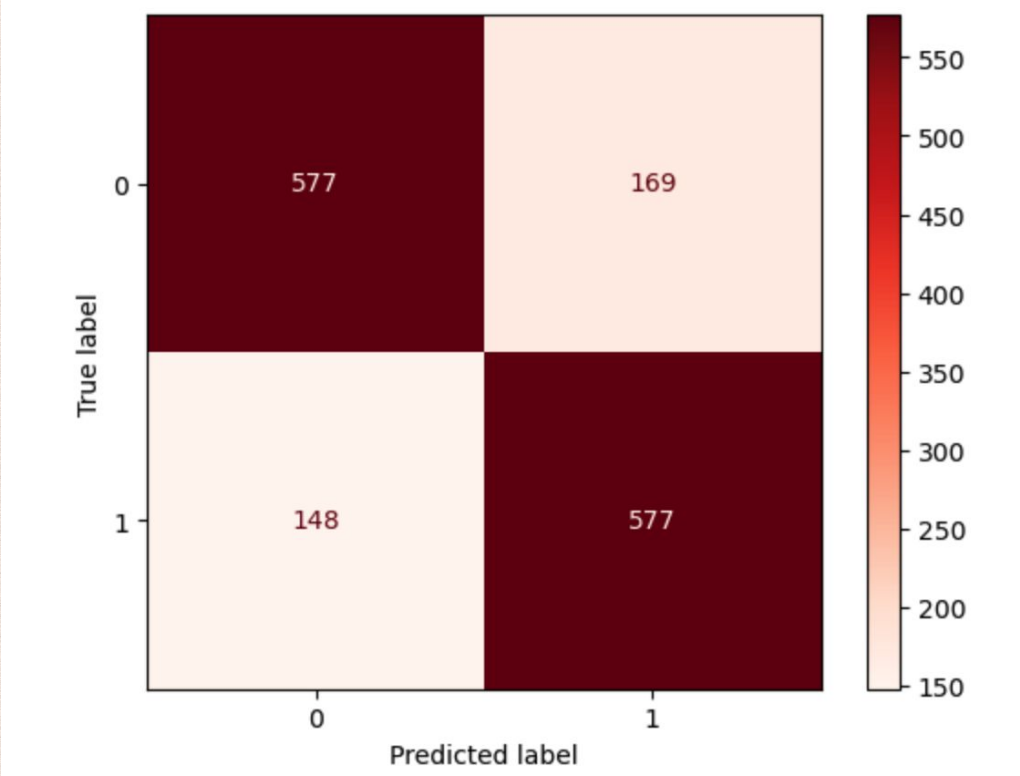
Most common terms found

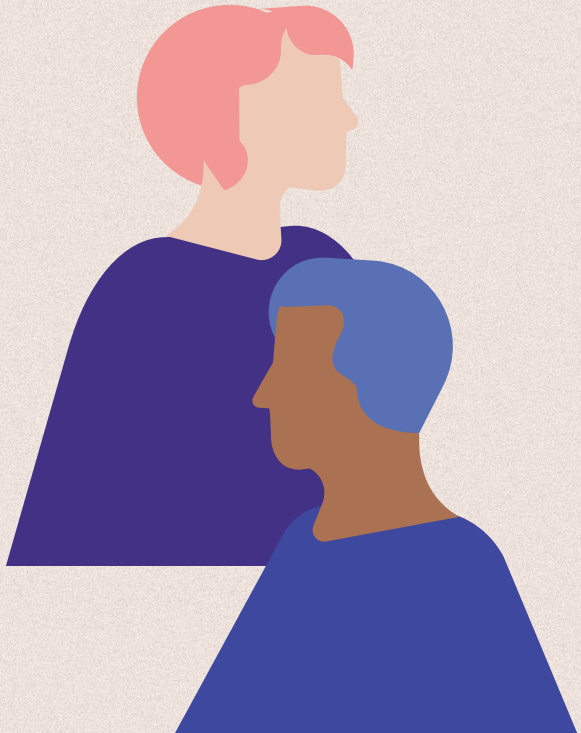


03 Evaluate the model

Best parameters for the Vectorizer	Scoring
• 'cvec__max_features': None, • 'cvec__min_df': 2, • 'cvec__ngram_range': (1, 2), • 'cvec__stop_words': from NLTK stopwords library	• Training: 0.876 • Testing: 0.783
Use these tuned parameters	Modest accuracy of predictions

03 Confusion matrix





03 Tuning a Term Frequency-Inverse Document Frequency Vectorizer

Why consider a different vectorizer?

Similar parameters:
Stop words,
Minimum document frequency,
N-gram,
Maximum number of features

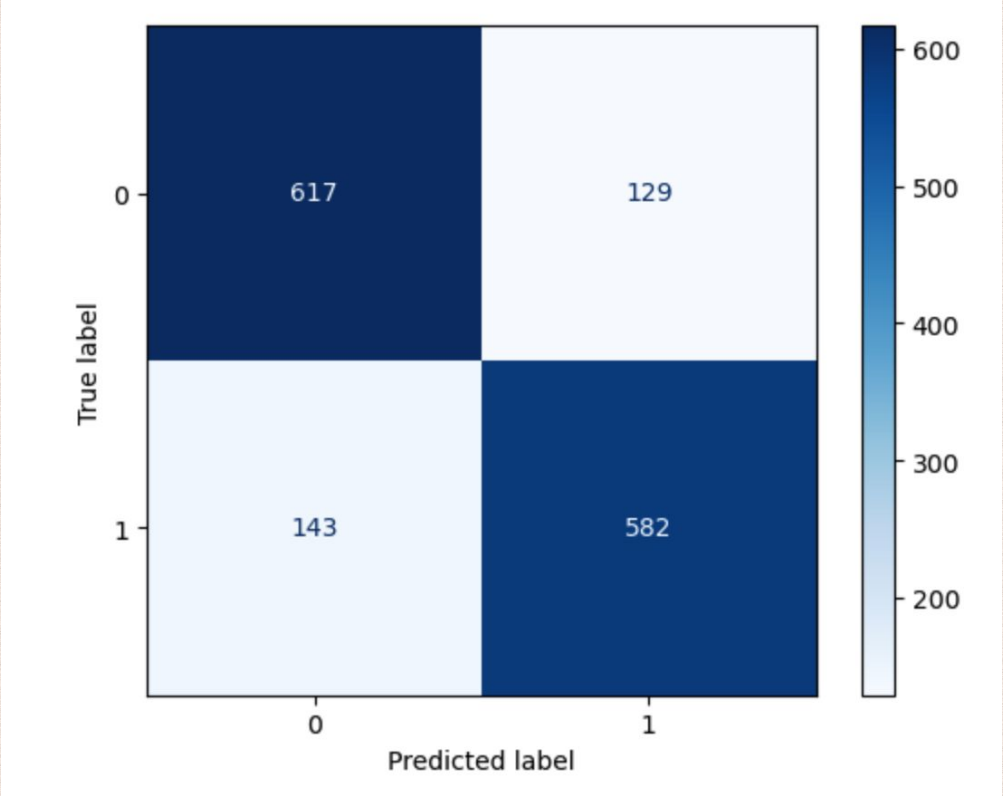
Score a term's importance relative to all documents

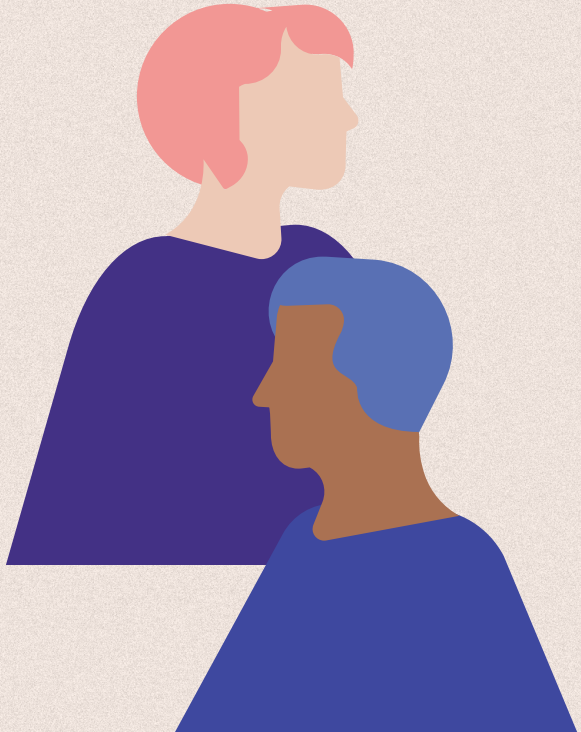
03

Build an estimator from Logistic Regression

Best parameters for the LogReg	Scoring
• logreg__C': 2.154434690031884, • 'logreg__penalty': 'l1', • 'logreg__solver': 'liblinear',	• Training: 0.889 • Testing: 0.815
Use these tuned parameters	Modest improvement

03 Confusion matrix





03 Decision Tree Classifier

Why consider a decision tree?

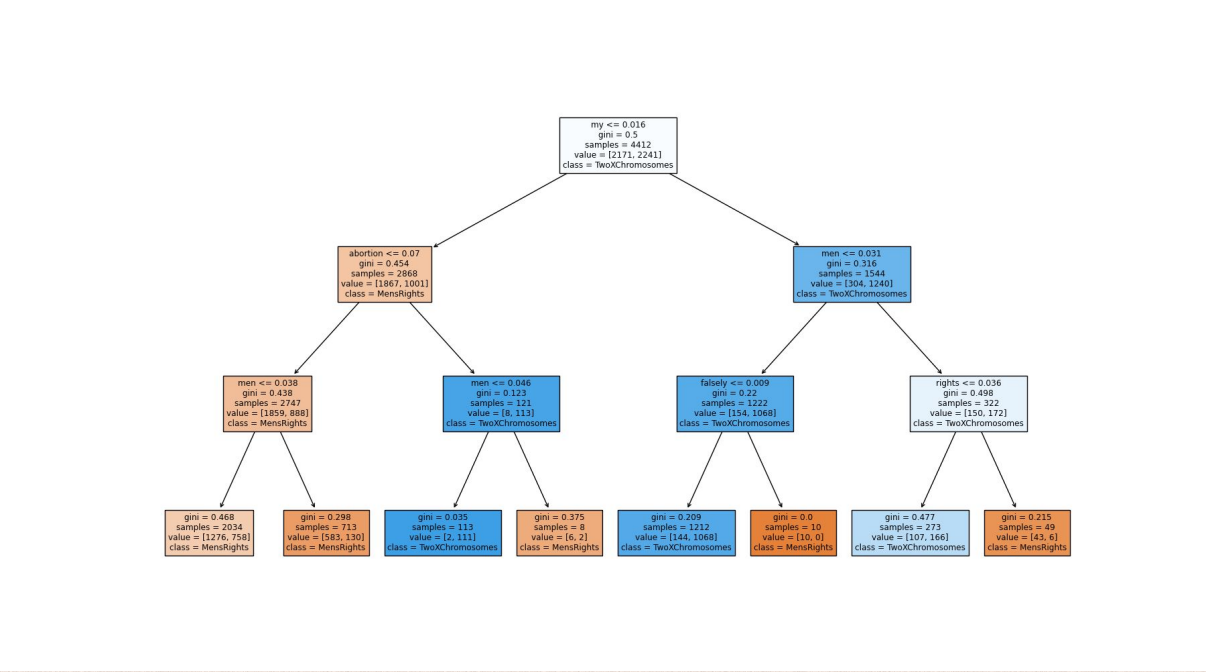
Different paradigm:
Data may not be linearly separable

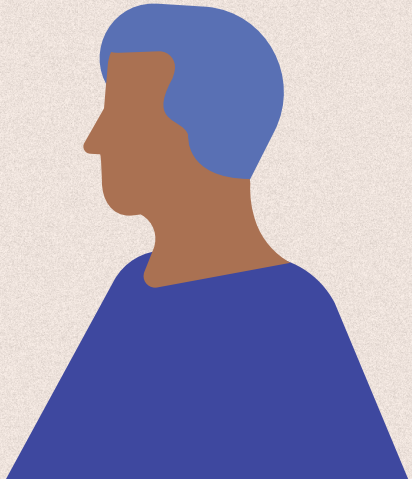
Tree models are robust and easy to tune

Best Parameters:

'tree__max_depth': 3,
'tree__min_samples_split': 3,

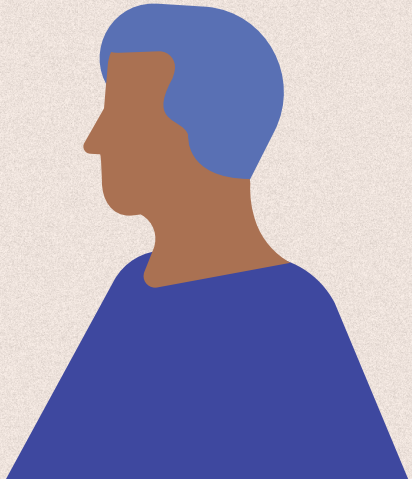
03 Decision Tree





03 Decision Tree

variable		importance
2665	my	0.646271
2546	men	0.164139
49	abortion	0.134946
3357	rights	0.030593
1353	falsely	0.024050
3331	republicans	0.000000



04

Future study

Compare to other non-linear estimators

May not be linearly separable

Add sentiment analysis

A tool for identifying hate speech
Proliferation of toxic beliefs about gender