AskWomen AskMen Classification

Subreddit comparison

Sean Bjork

Problem Statement:

To build an accurate (over 70%) predictive model for classifying Subreddit posts as from either AskMen or AskWomen.

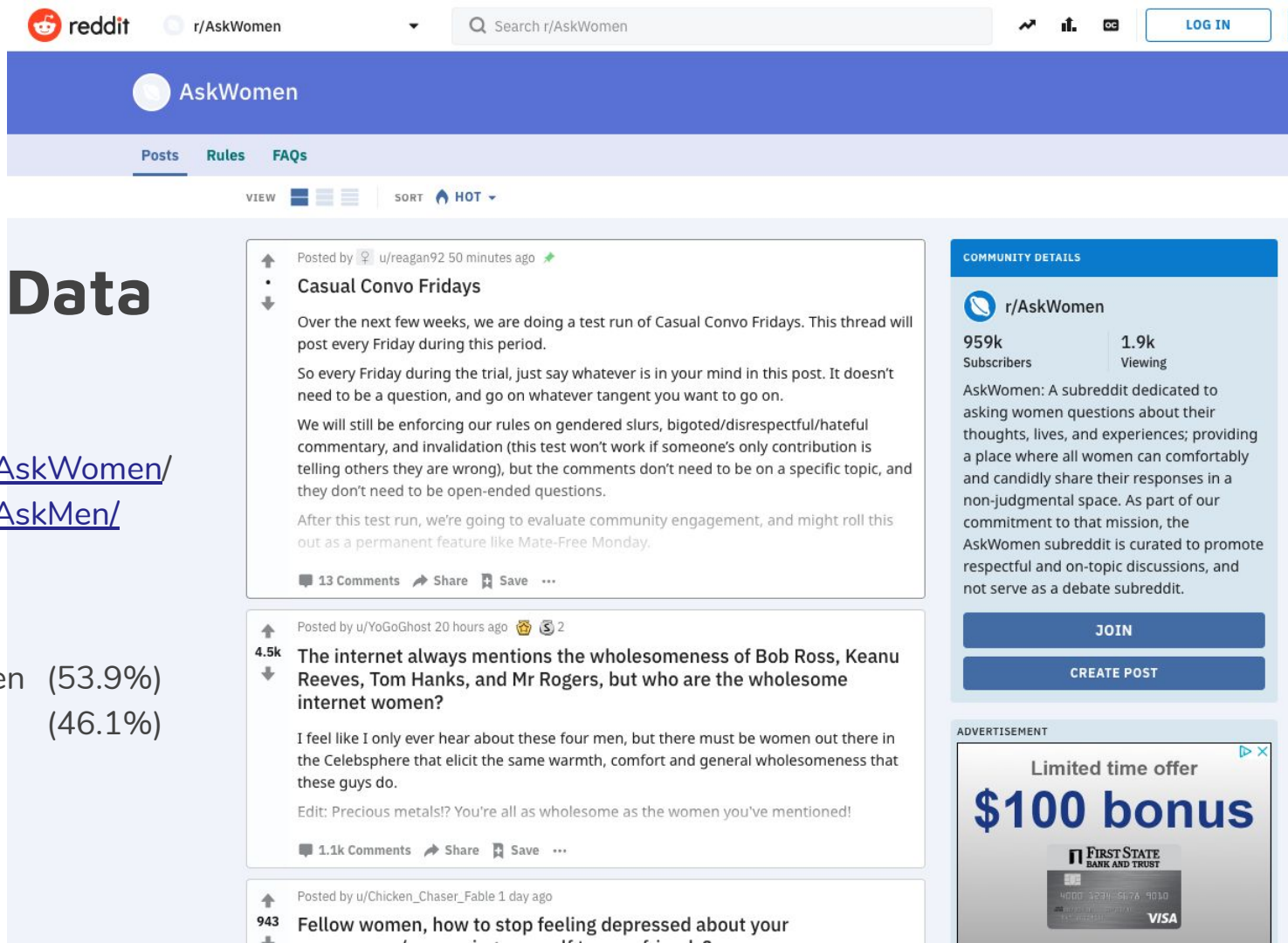
About the Data

Collected questions from:

- https://reddit.com/r/AskWomen/
- https://reddit.com/r/AskMen/

Number of questions:

- 684 from AskWomen (53.9%)
- 585 from AskMen (46.1%)
  - 1269 total



Sample Questions with *Signal*

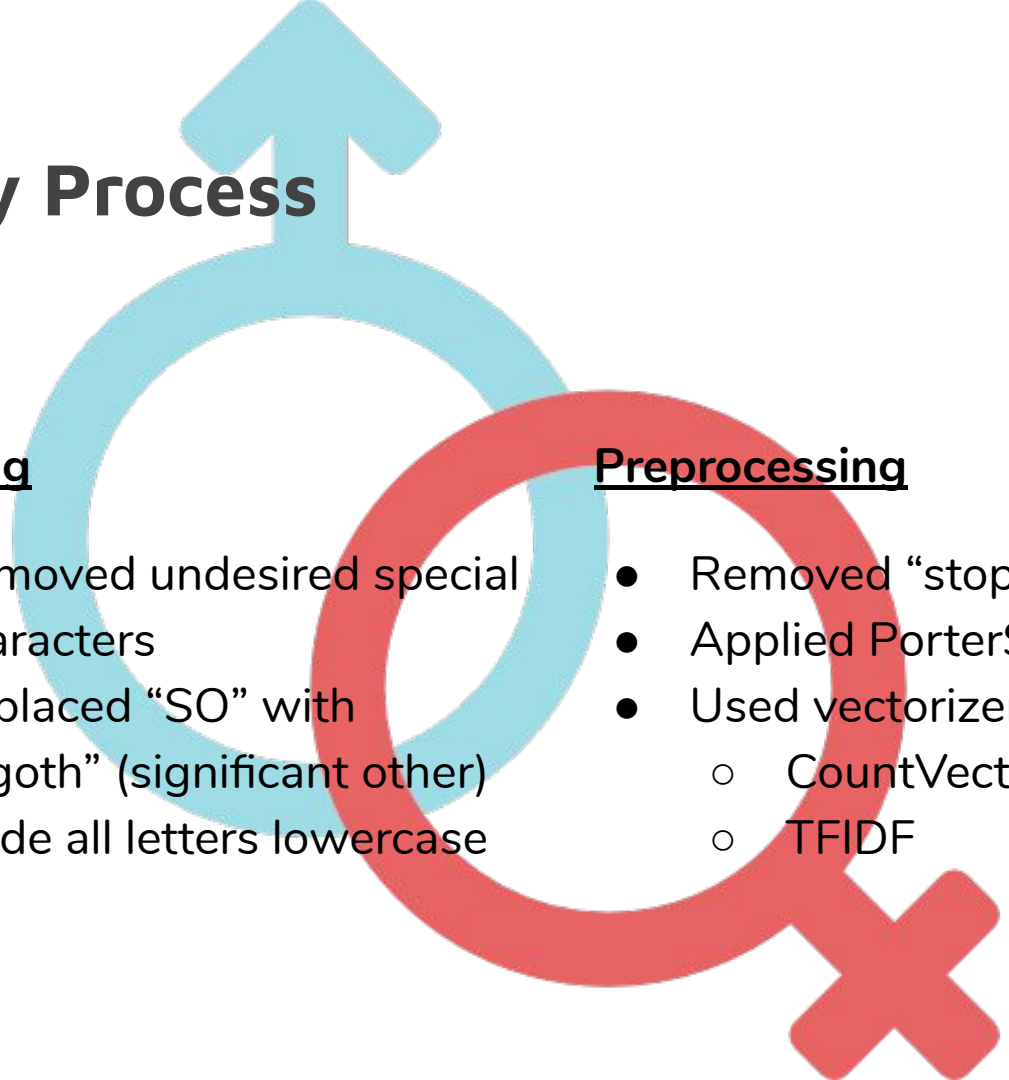
"What weird food craving did you have while pregnant? And what kind of food couldn't you stand anymore?"

- AskWomen

'What makes a strong woman in your opinion? As in how can you tell she is strong?'

- AskMen

Early Process

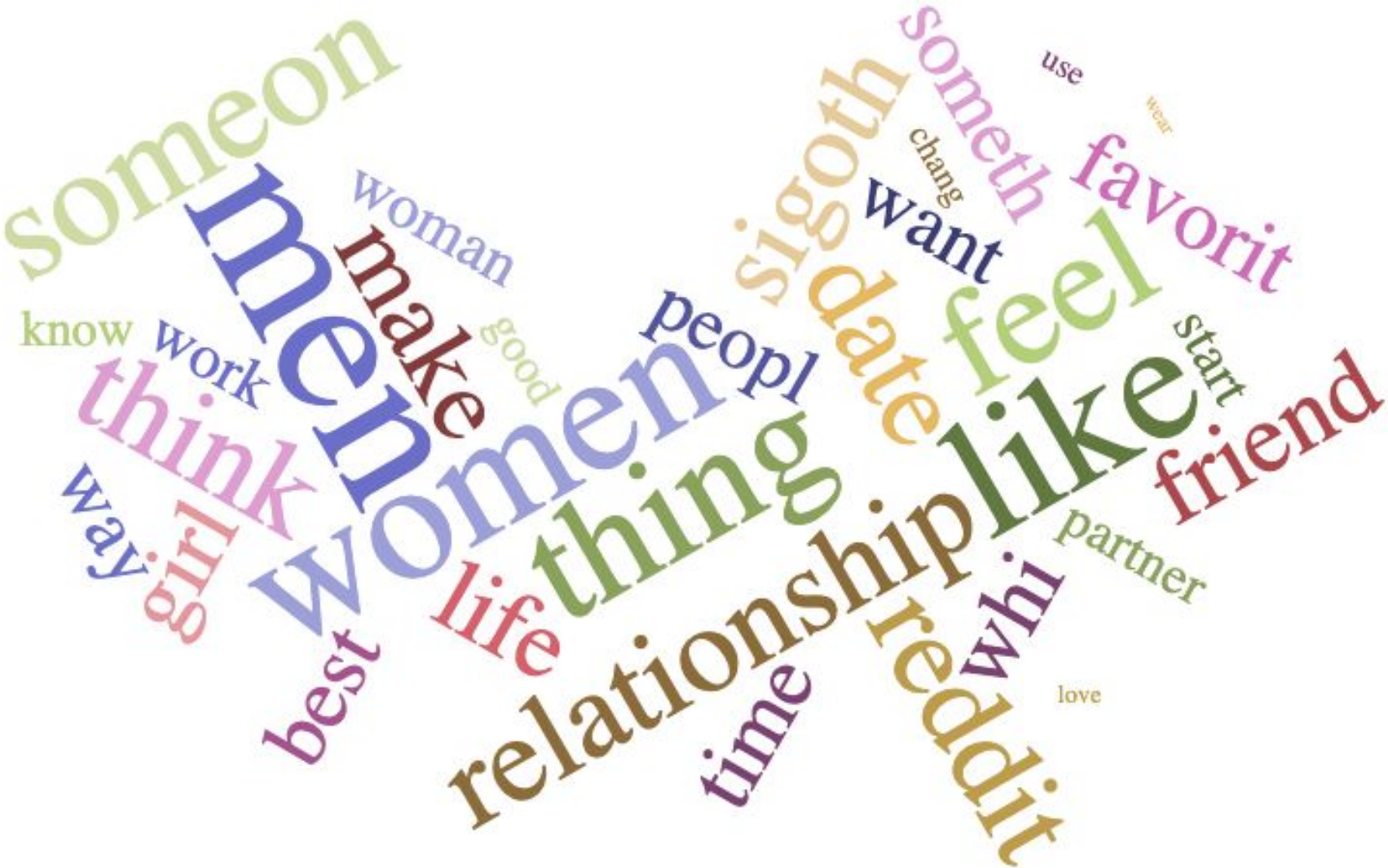


Cleaning

- Removed undesired special characters
- Replaced “SO” with “sigoth” (significant other)
- Made all letters lowercase

Preprocessing

- Removed “stop words”
- Applied PorterStemmer
- Used vectorizers:
  - CountVectorizer
  - TFIDF


someon
men
woman
make
know
work
good
think
women
way
girl
life
best
relationship
thing
peopl
date
sigoth
someth
use
wear
chang
want
favorit
feel
start
like
friend
partner
whi
reddit
time
love

Top words:
Women Only

Top words:
Men Only

Modeling

Logistic Regression
Naive Bayes

with CountVectorizer and TFIDF

We used GridSearchCV to find best hyperparameters.

Best model:

Multinomial Naive Bayes

TFIDF & alpha = ~2.2

Best accuracy: **66.4%**

Conclusions / Limitations

- There signal coming from our predictive features is significant.
- We failed to achieve accuracy of greater than 70%.
- Our model may perform better with more training data.

- We recommend modeling with Multinomial Naive Bayes
  - TFIDF vectorizer.

- More data → Better model
- Tough to discern short questions with few / no distinguishing keywords
  - Both ask about “men” / ”women”