Classification of Subreddits through NLP

Danielle Reycer, Data Scientist

The Task

- Improving Marketing campaigns through classification and understanding needs of parents (Part 1 of a larger scope project)

meredith

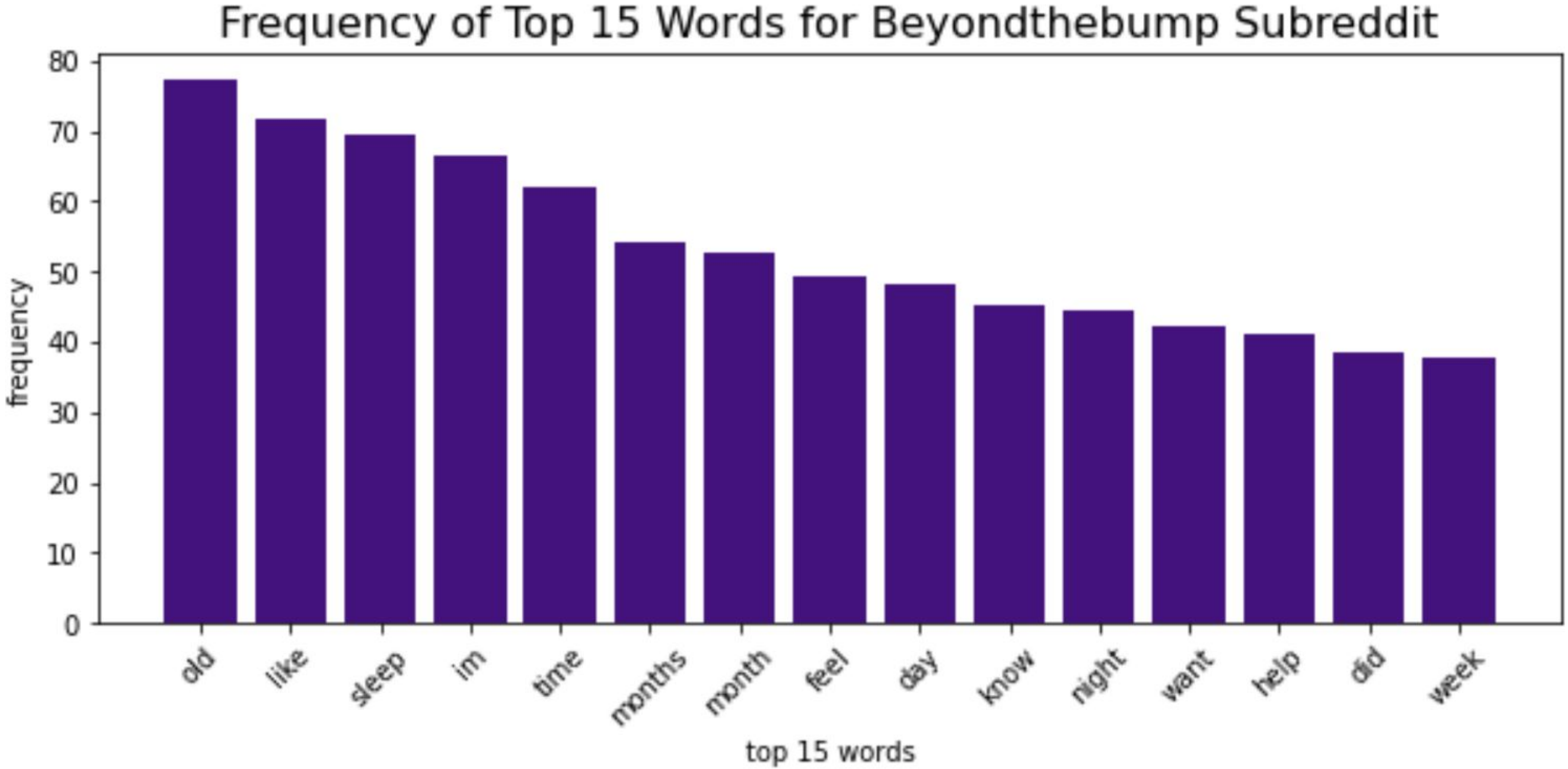
Bold. Together.

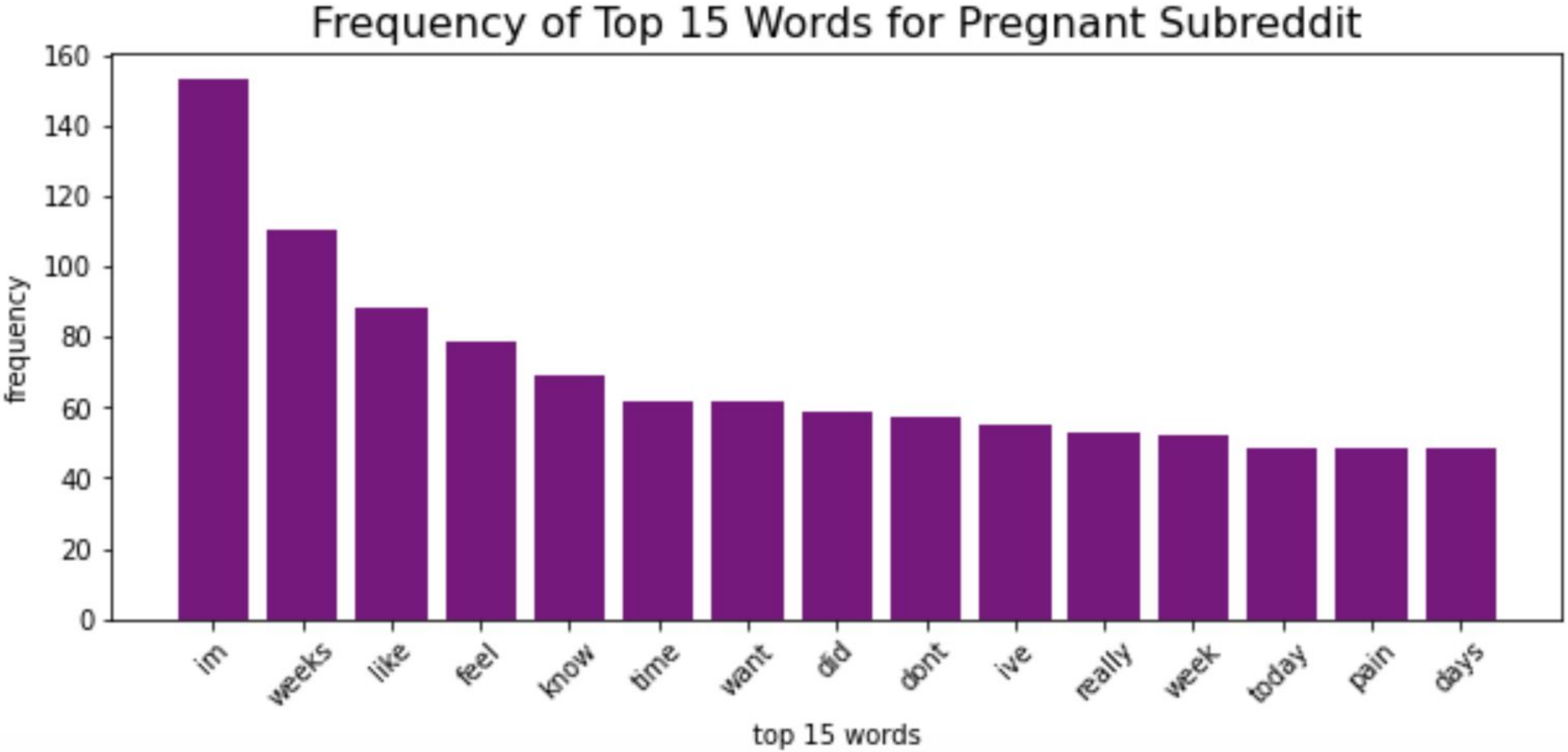
Process

- Used Reddit's built in API to collect 2,000 posts from two subreddits:
  - r/pregnancy
  - r/beyondthebump
- Cleaned and Analyzed Data
  - Deleted posts with very few words

- Looked at frequent words
  - Tfidf Vectorizer
- Word Count
- Question Marks

Top Words


Frequency of Top 15 Words for Beyondthebump Subreddit
frequency
0
10
20
30
40
50
60
70
80
old
like
sleep
im
time
months
month
feel
day
know
night
want
help
did
week
top 15 words


Frequency of Top 15 Words for Pregnant Subreddit
frequency
0
20
40
60
80
100
120
140
160
im
weeks
like
feel
know
time
want
did
dont
ive
really
week
today
pain
days
top 15 words

Overlap:

time, did, know, want, like, feel, week, I’m (im)

53% of the top 15 words in each subreddit appeared in both

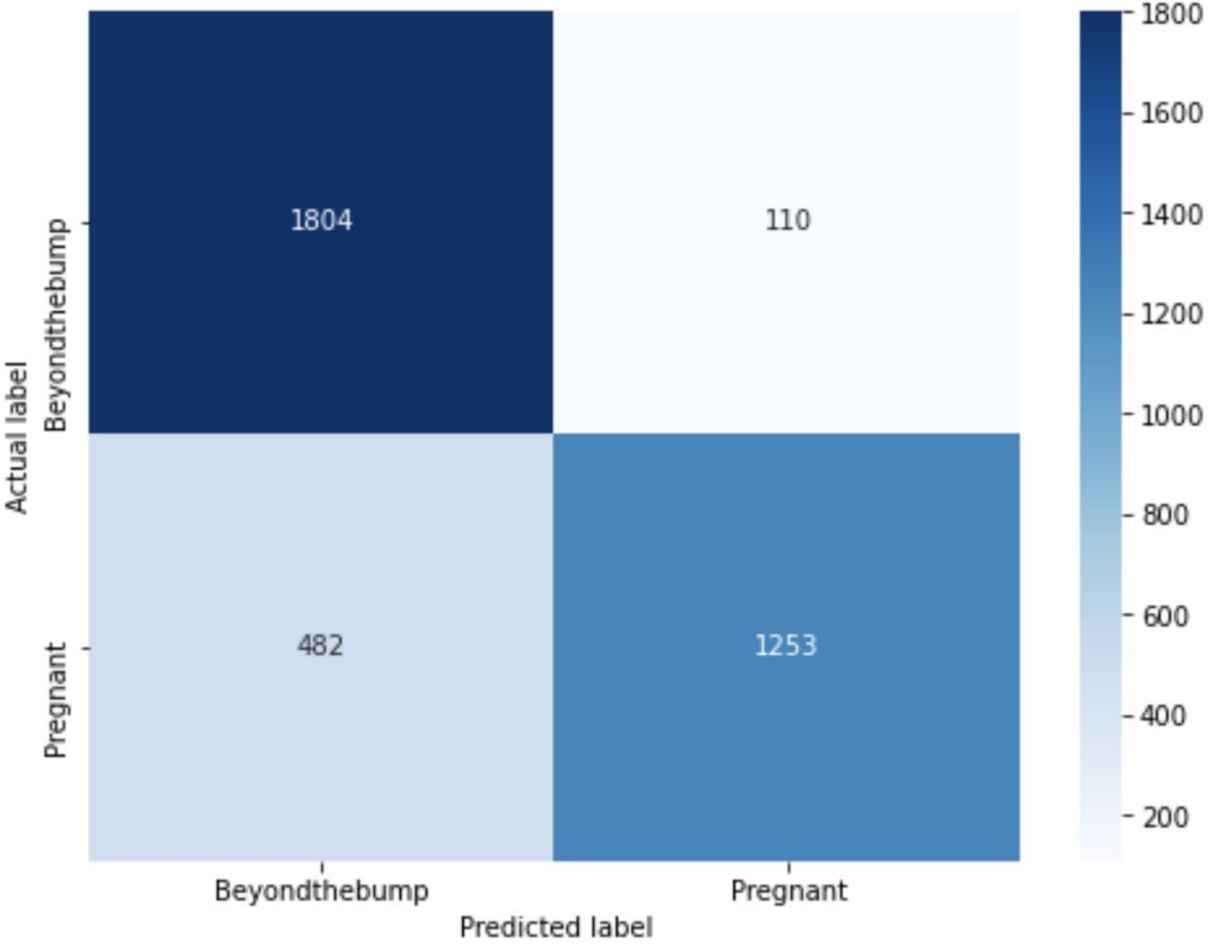
word
count

Best Model

Random Forest:

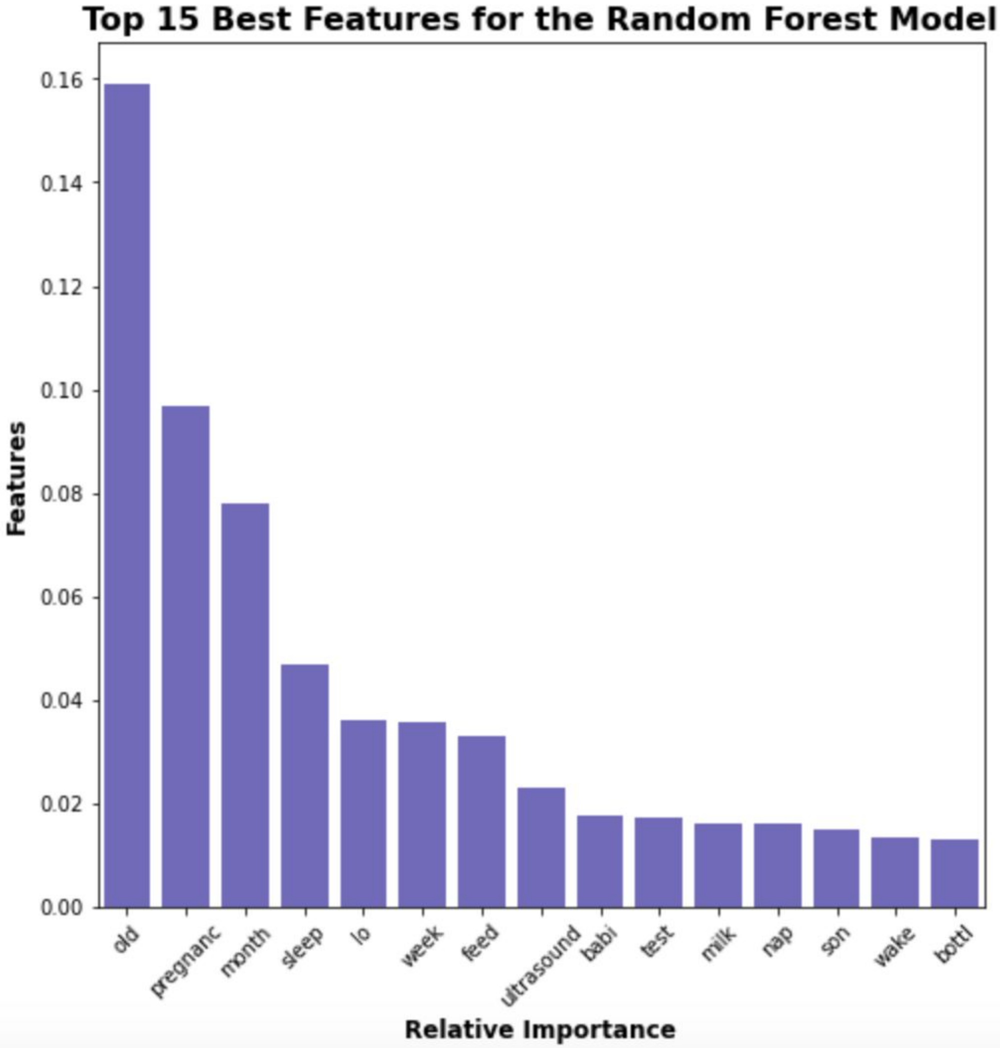
Accuracy of 80% ± 2% (95% confidence interval)

Chosen since it had less variance than other models and is easier to interpret than an ensemble model


1804
110
482
1253
Beyondthebump
Pregnant
Actual label
Beyondthebump
Pregnant
Predicted label
1800
1600
1400
1200
1000
800
600
400
200

Top Features

The models performed best with stemmed text - hence the shortened words. Here they are shown in order of importance in the model.



Conclusions and Recommendations

- Phase 1b:
  - Work to identify which types of posts are being misclassified - this may improve our initial model.

- Phase 2:
  - Analyzing classified posts for products/ services needed in order to create targeted marketing campaigns