Project 3: Web APIs & NLP

By: Zaid Aziz

Problem Statement

I am a data scientist working at an online dating platform. We asked Reddit for all the posts for the subreddits “Dating over 30”, and “Dating over 40” to identify key words associated with each demographic.

Reddit gave us the posts, but says we have to pay extra to know which posts belong to which subreddit. I am going to save my company money and classify in house.

Background: Dating over 30 and 40 Subreddits

- Discussion boards to exchange ideas about the art of dating.
- The intention of each board is to cater to the particularities of each groups demographics.
- They really are a place for people to complain and argue about the difficulties of dating in the information age.

NLP Classification

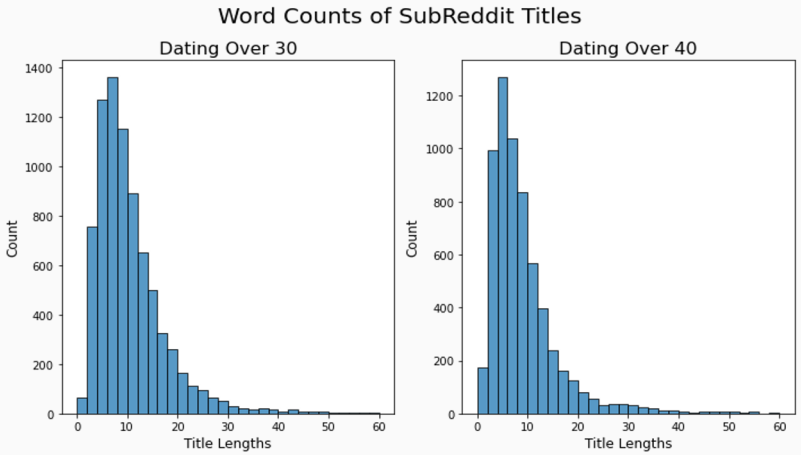
- We will use NLP to classify whether a particular post belongs to the subreddit “Dating Over 30” or “Dating over 40”.
- Using Random Forest Classifier, SVM, and Multinomial Naive Bayes.
- We will be focusing on accuracy. Want to maximize correct classification.

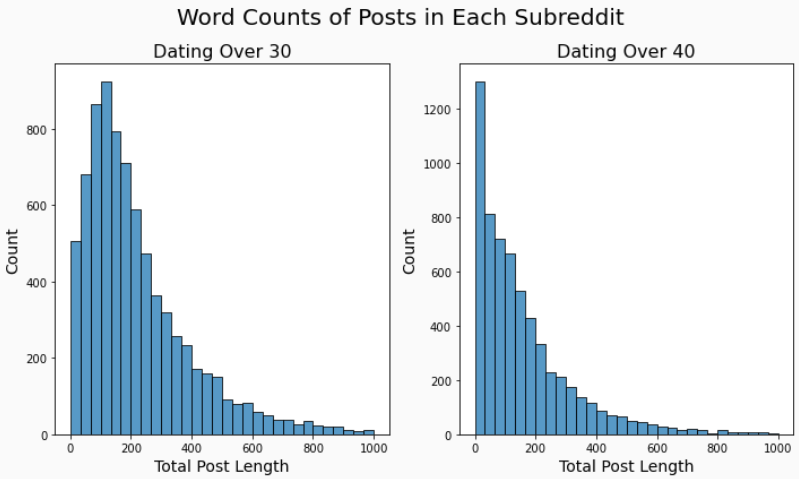
Gathering Data: Pushshift API

- Created a function to pull Reddit posts from both subreddits.
- Included a sleep timer of 2 seconds, to not overload the Reddit server with requests.
- 10:1 Imbalance of posts from “Dating over 30” vs. “Dating over 40”.
- Pulled 30K posts from DOT, and 8K posts from DOF.

Explore the Data:

- Word Counts Subreddit Titles:
  - Date over 30 Mean: 10 words
  - Date over 30 Mode: 6 words
  - Date over 40 Mean: 8.5 words
  - Date over 40 Mode: 5 words
- Word Counts Subreddit Total words:
  - Date over 30 Mean: 230 words
  - Date over 30 Mode: 44 words
  - Date over 40 Mode: 5 words
  - Date over 40 Mean: 163 words

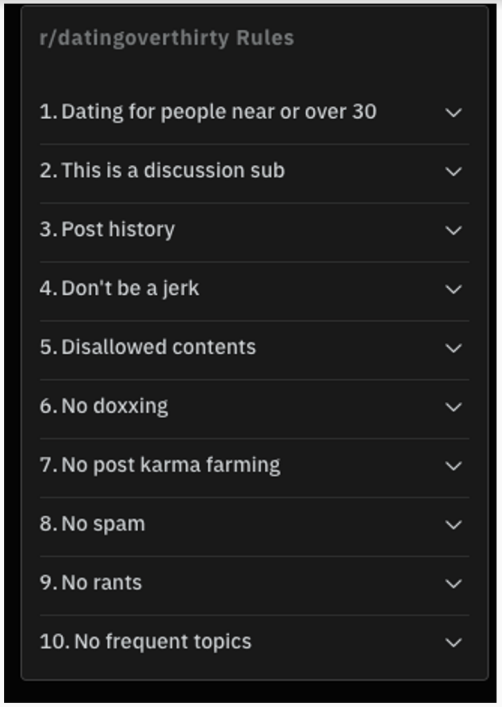




Pre Processing

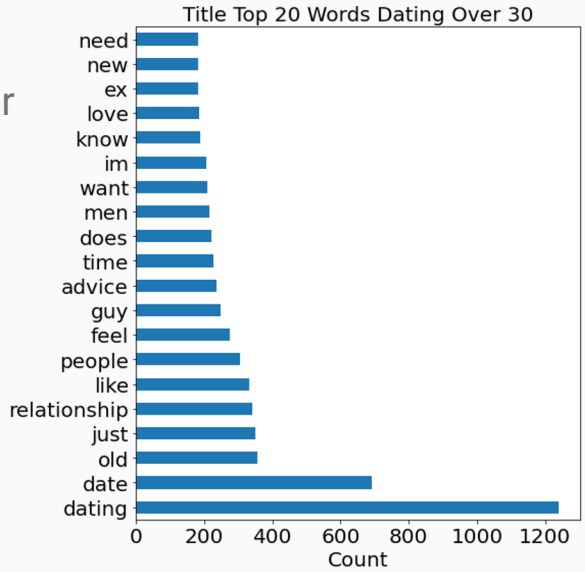
- Dropped the removed by moderator posts.
    - DOT: 60%
    - DOF 25%
- Combined the posts “selftext” (Body) feature, with the “Title”
- Pulled out all the “apostrophes”, in the corpus.
- Total of ~7100 DOT/ ~6000 DOF

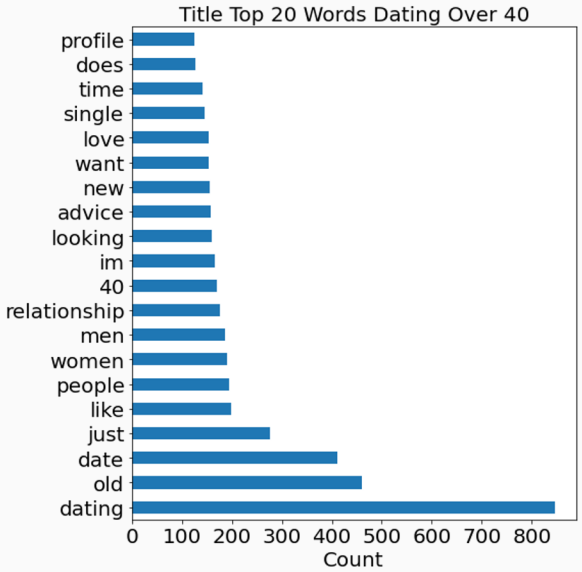
“Dating over 30” - Rules



Pre Processing Con’t

- 50% split of data into the model
- CountVectorizer / TF-IDF Vectorizer
- Stop Words
- Max_df
- Min_df
- Ngram_range
- Max_features.





Models Tested

SVM Class. w/Count Vectorizer:	Random Forest Classifer w/Count Vectorizer :	Multinomial Naïve Bayes w/ Count Vectorizer:	Multinomial Naïve Bayes w/ TF-IDF Vectorizer:
CV5 Score: 72%	CV5 Score: 71%	CV5 Score: 71%	CV5 Score: 70%
Train Score: 96%	Train Score: 99%	Train Score: 80%	Train Score: 80%
Test Score: 73%	Test Score: 71%	Test Score: 71%	Test Score: 71%
	Max Depth: None		

SVM Classifier: Difficulty

- Used 12k Reddit posts.
  - Specificity: 65 %
  - Precision: 69 %
  - Accuracy: 73 %

Classifications

FN: “Meeting People from the past”

FP: “Why does he still talk to them?”

TP: “Stood up! How should I react??”

TN: “I need some input”

Recommendations

Find another SubReddit
Or
Further Pre-Processing

Recommendations Cont.

- There is a possibility that the subreddits are too similar.
- Possibly that there is something in the lexicon and sentiment analysis that splits the two subreddits.
- Find another Subreddit that splits the two demographics 30 y/o vs. 40 y/o better.

Thank You

Any Questions?