Modeling with NLP

Using Text Data from Reddit

Matt Williams

July 24th, 2020

Problem Statement

A linguistics researcher is interested in studying how people communicate in online forums. They have decided to look to Reddit – "the front page of the internet" – as a source for their investigation. They've identified two subreddits for analysis: r/AskMen and r/AskWomen. However, they lack the skills necessary to parse through the data quickly. They've hired a consulting firm to help with the project, with a goal of determining whether the language used in posts from each subreddit is sufficient to predict which subreddit a post comes from.

Disclaimer: observations are posts from users of r/AskMen and r/AskWomen; we are not making assumptions or generalizations about genders.

Collecting the Data

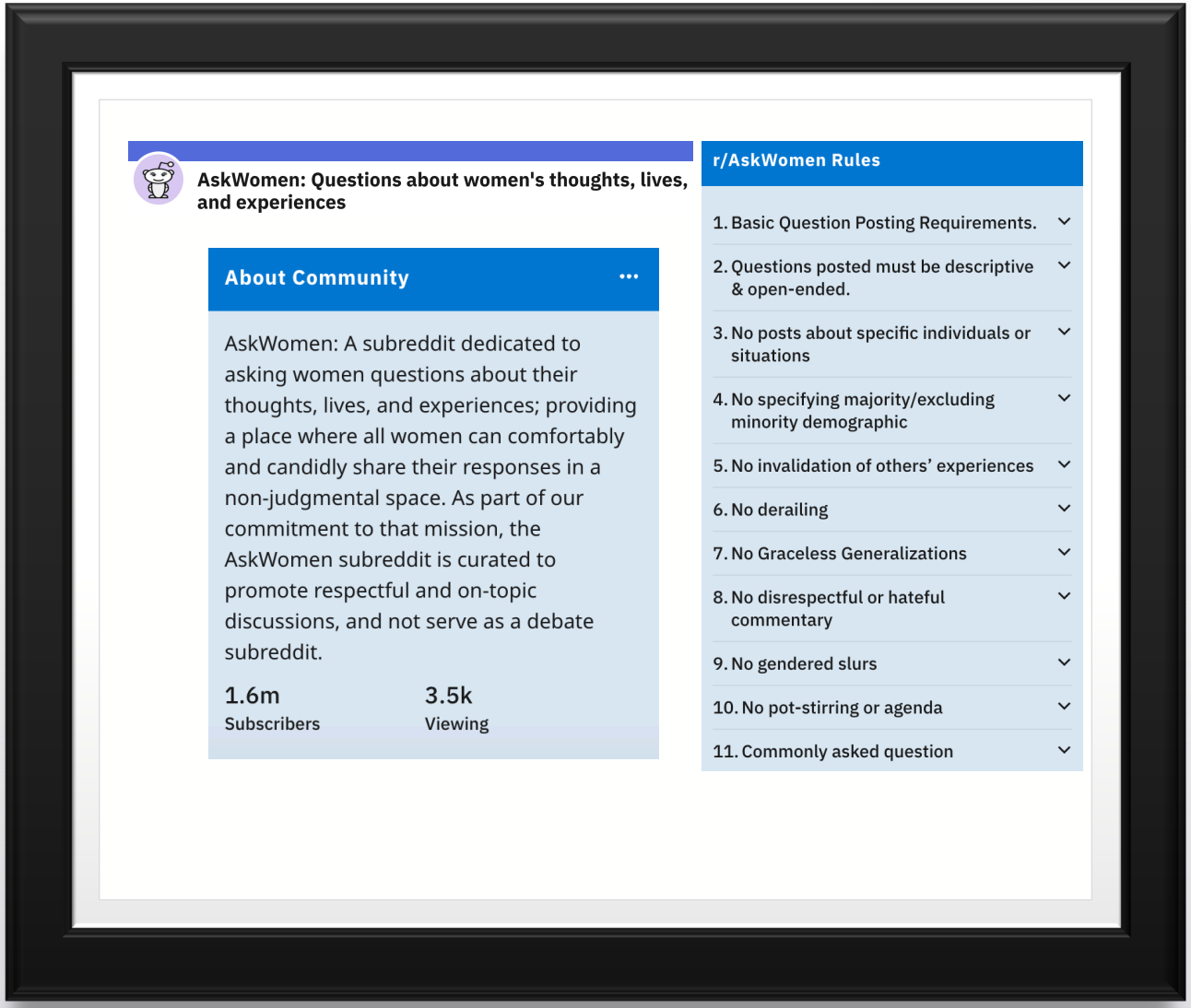
- Python Reddit API Wrapper (PRAW)
- Nearly 1,900 Observations from r/AskWomen and r/AskMen

Feature	Data Type	Summary
title	Object	Text for title of post
score	Integer	# Upvotes - # Downvotes
url	Object	Url for specific post
number of comments	Integer	How many comments on a post received
created	Float	Timestamp for when item was posted
body	Object	Supporting text for post (60% of posts were null)
subreddit	Integer	0 = r/AskMen; 1 = r/AskWomen

Image Source

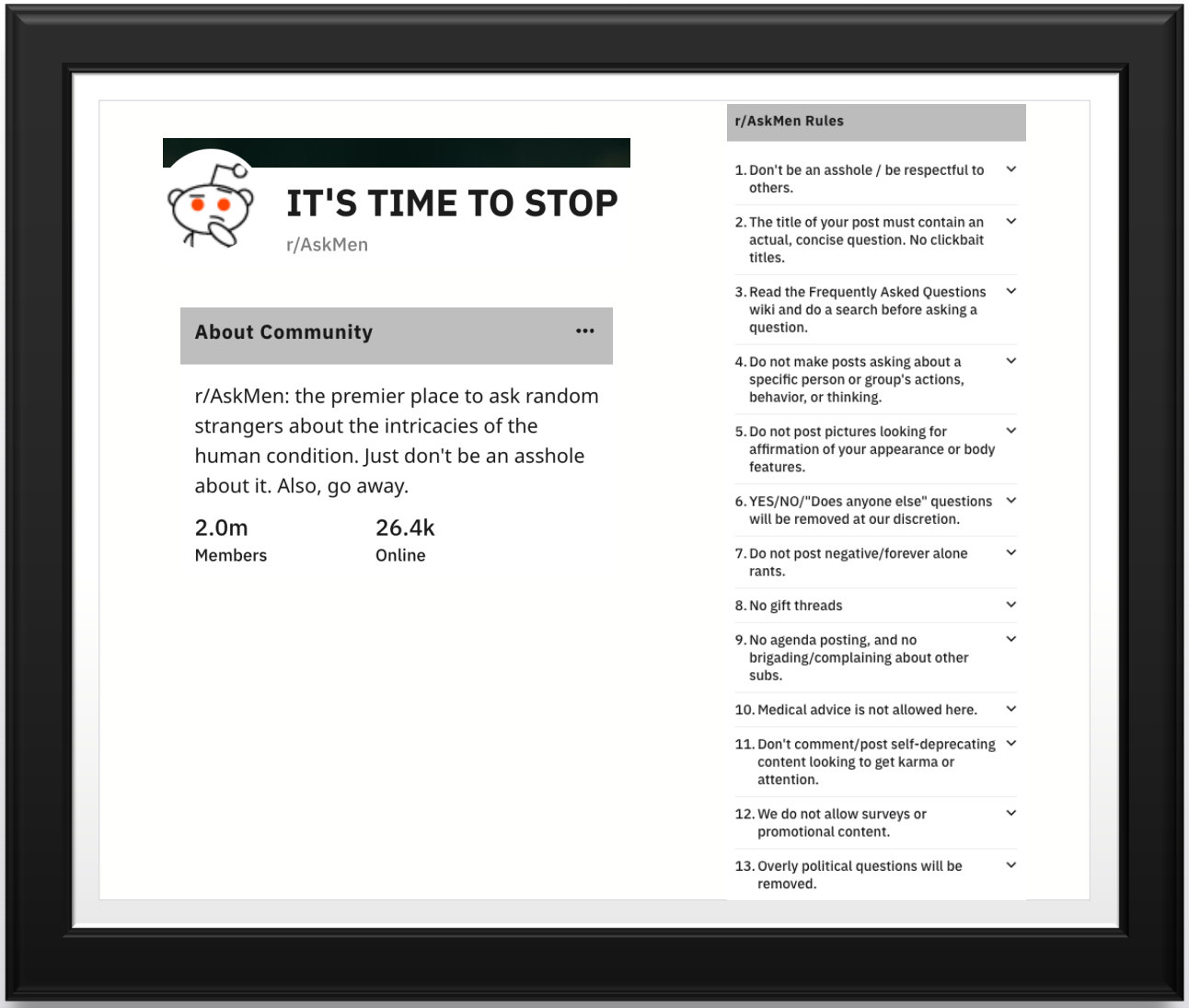
r/AskWomen

- Source: https://www.reddit.com/r/AskWomen/
- Created July 17th, 2010
- Data based on 980 newest posts



r/AskMen

- Source: https://www.reddit.com/r/AskMen/
- Created August 30th, 2010
- Data based on 983 newest posts



25 Most Commonly Used Words by Subreddit

r/AskMen

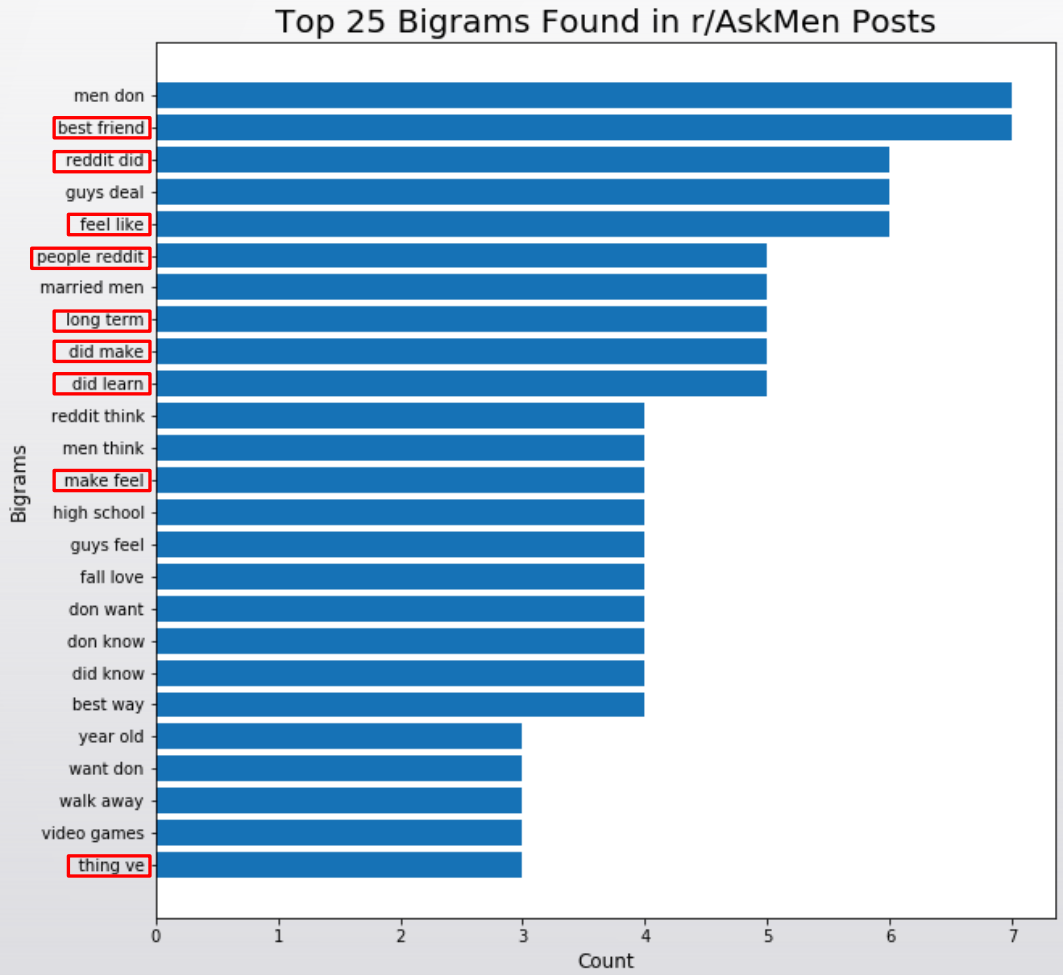
r/AskWomen

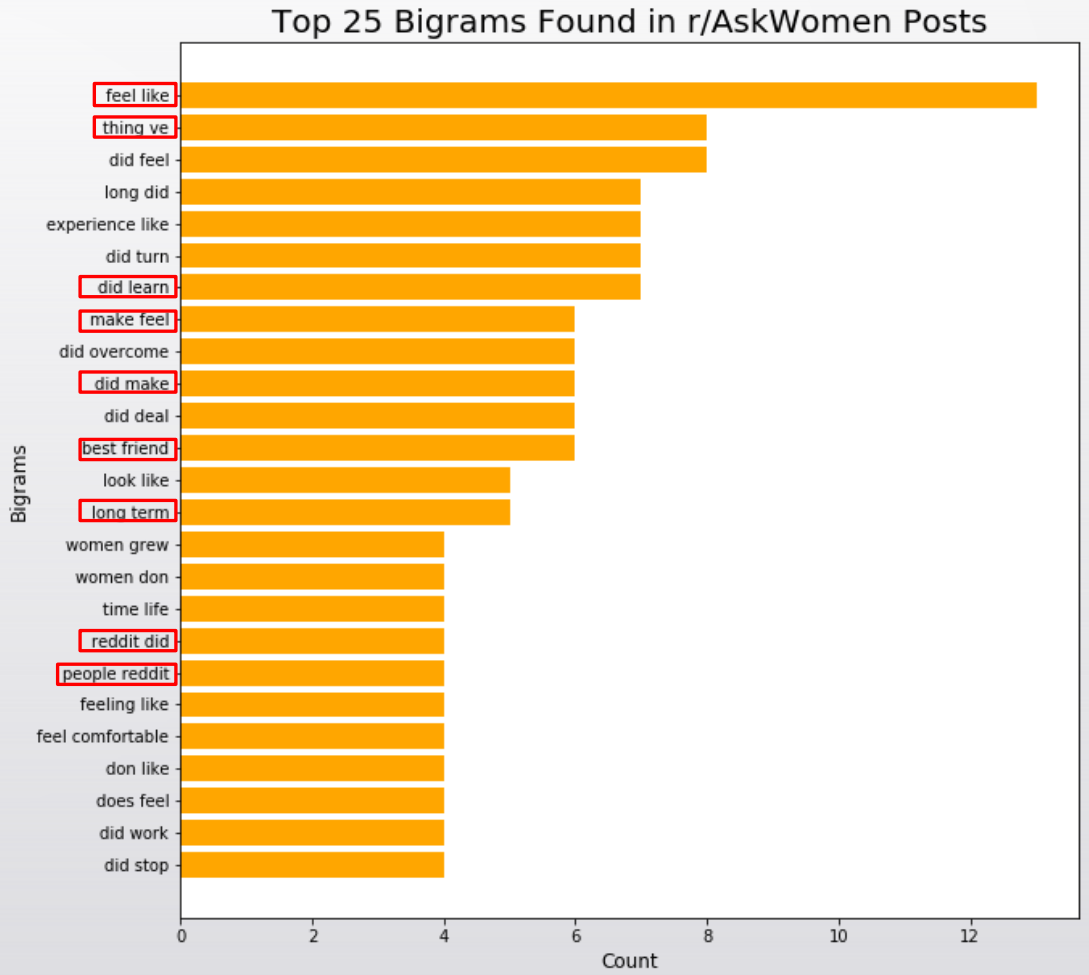
r/AskMen Only

r/AskWomen Only

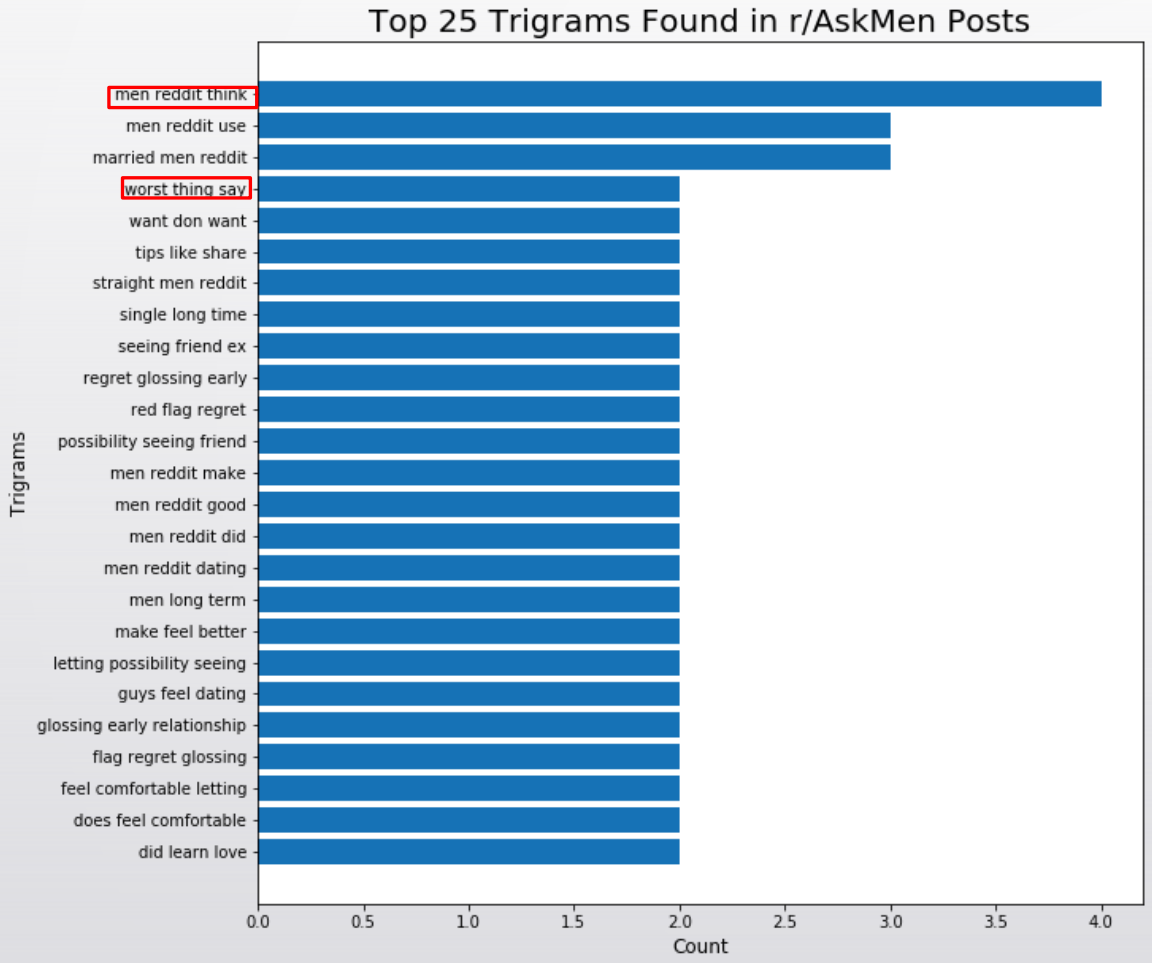
Both

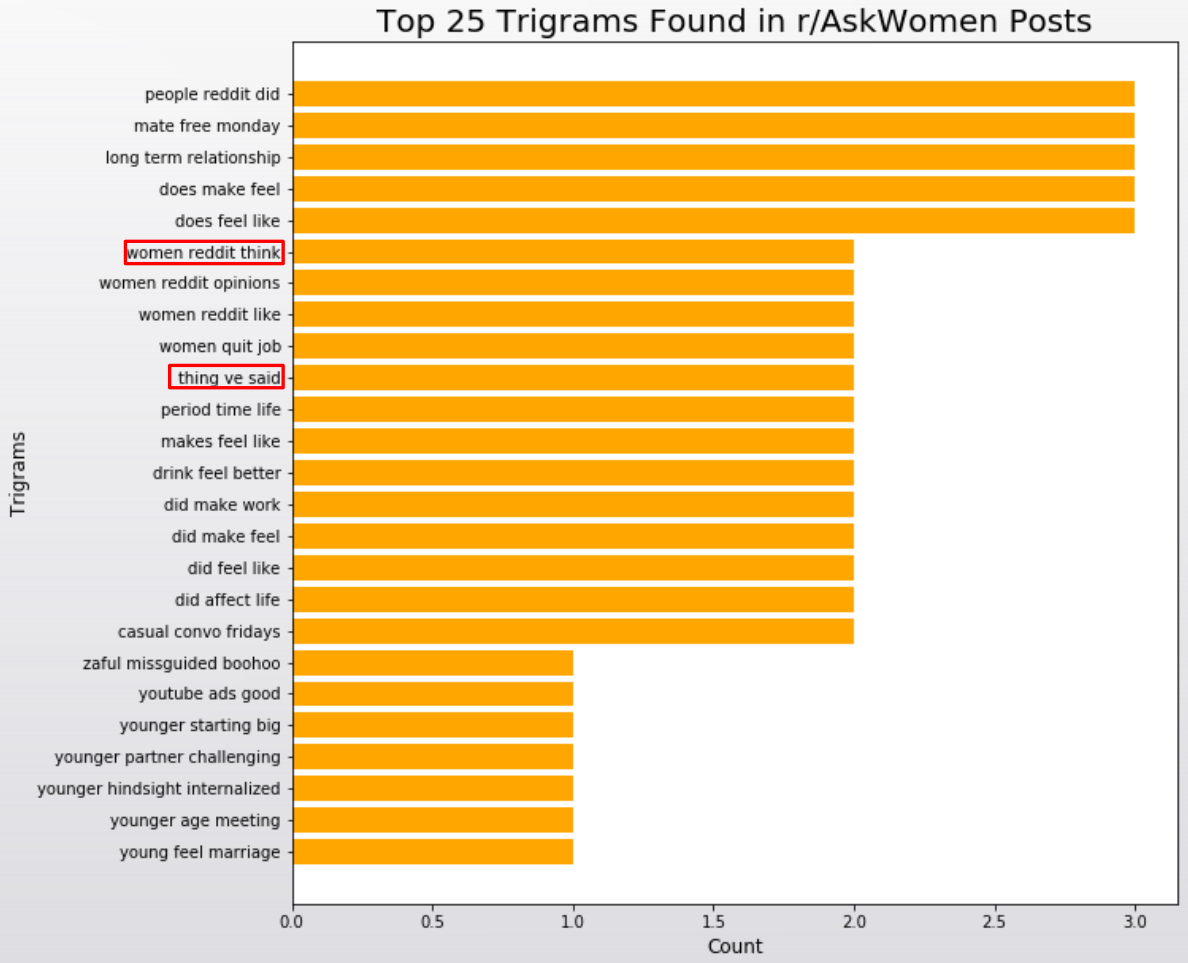
25 Most Common Bigrams





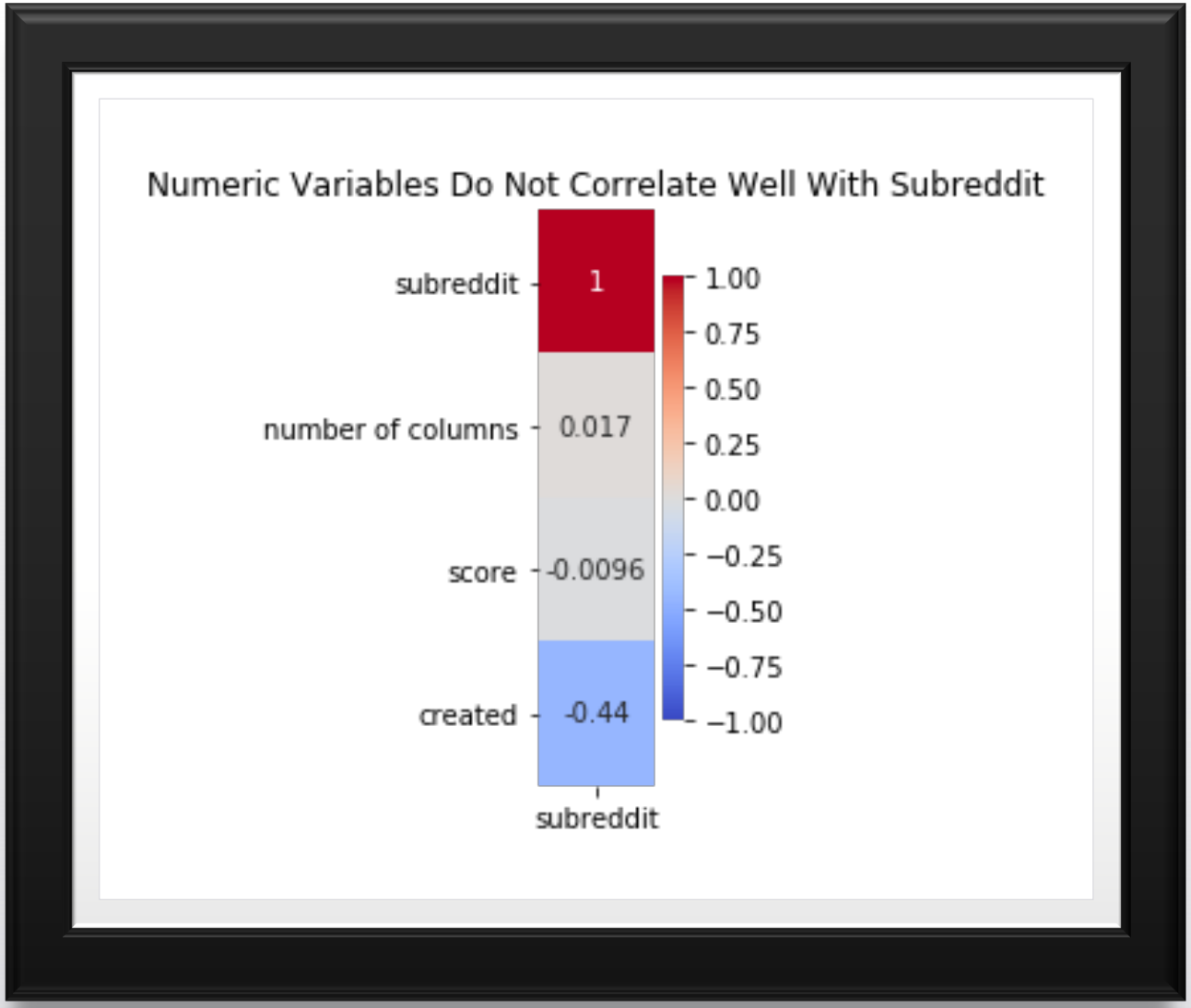
25 Most Common Trigrams





Numeric Variables

- Since these do not have a strong relationship with our target, we will only be using text data for modeling



Modeling

- Measure: Accuracy
- Baseline Accuracy: .5
- Default hyperparameters
- Extreme overfitting
- TF-IDF vectorizer performs better
- Decision criteria: further explore models with the highest cross val scores

Model	Vectorizer	Accuracy		
		Best Score	Train	Test
Logistic Regression	TF-IDF	0.6997	0.8987	0.6904
Extra Trees	TF-IDF	0.6877	0.9905	0.6925
Multinomial Naïve Bayes	Count	0.6868	0.9171	0.6762
Logistic Regression	Count	0.6834	0.9552	0.7088
Random Forest	Count	0.6800	0.9905	0.6965
Random Forest	TF-IDF	0.6793	0.9905	0.6823
Multinomial Naïve Bayes	TF-IDF	0.6712	0.9300	0.6639
Extra Trees	Count	0.6692	0.9905	0.7006

Logistic Regression

Model	Vectorizer	Accuracy		
		Best Score	Train	Test
Logistic Regression	TF-IDF	0.695	0.8397	0.7026
		Hyperparameters		
		TF-IDF	Max Feature	1000
		TF-IDF	Max DF	0.9
		TF-IDF	Min Df	2
		TF-IDF	N-gram Range	(1,1)
		TF-IDF	Stop Words	None

Extremely Randomized Trees

Model	Vectorizer	Accuracy		
		Best Score	Train	Test
Extra Trees	TF-IDF	0.6787	0.8016	0.6965
		Hyperparameters		
		Extra Trees	Max Depth	4
		Extra Trees	Estimators	125
		TF-IDF	Max Features	500
		TF-IDF	N-gram Range	(1,1)
		TF-IDF	Stop Words	None

Conclusion & Recommendations

- r/AskMen or r/AskWomen? Don't ask me!
- Lots of common language and themes in these subreddits (more cleaning during pre-processing)
- Models used are prone to overfitting with this data
- Further fine tune hyperparameters
- Experiment with other models
- Using post titles is not enough - Incorporate more data (gather more observations; incorporate comments; use 'body' column)

Questions?