



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

The screenshot shows the landing page of the r/AskWomen subreddit. At the top, there's a purple header with the subreddit name and a small robot icon. Below it is a white 'About Community' section containing text about the subreddit's purpose and rules, along with subscriber and viewing statistics. To the right is a sidebar titled 'r/AskWomen Rules' with 11 listed rules, each with a collapse arrow.

About Community

AskWomen: A subreddit dedicated to asking women questions about their thoughts, lives, and experiences; providing a place where all women can comfortably and candidly share their responses in a non-judgmental space. As part of our commitment to that mission, the AskWomen subreddit is curated to promote respectful and on-topic discussions, and not serve as a debate subreddit.

1.6m Subscribers 3.5k Viewing

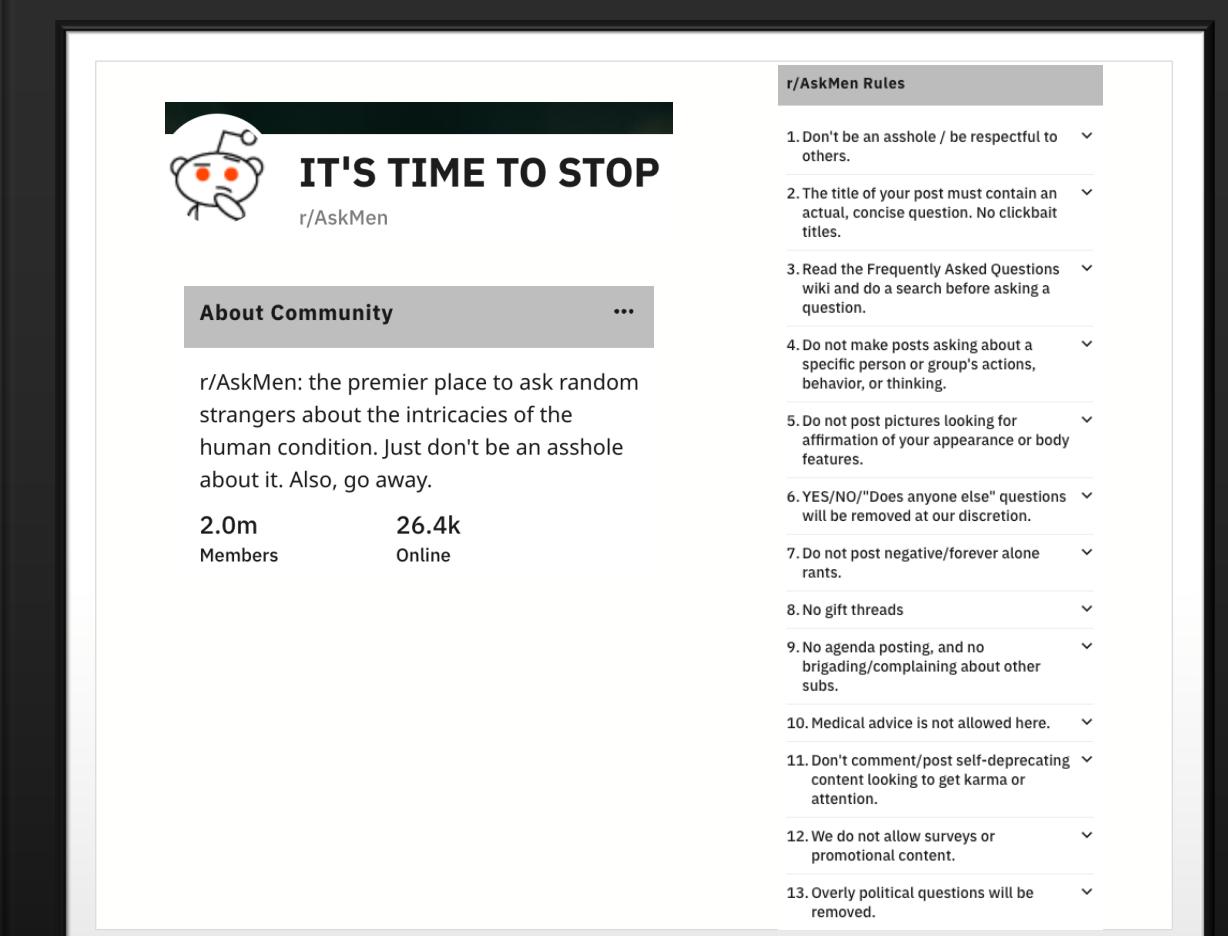
r/AskWomen Rules

1. Basic Question Posting Requirements. ▾
2. Questions posted must be descriptive & open-ended. ▾
3. No posts about specific individuals or situations ▾
4. No specifying majority/excluding minority demographic ▾
5. No invalidation of others' experiences ▾
6. No derailing ▾
7. No Graceless Generalizations ▾
8. No disrespectful or hateful commentary ▾
9. No gendered slurs ▾
10. No pot-stirring or agenda ▾
11. Commonly asked question ▾



r/AskMen

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



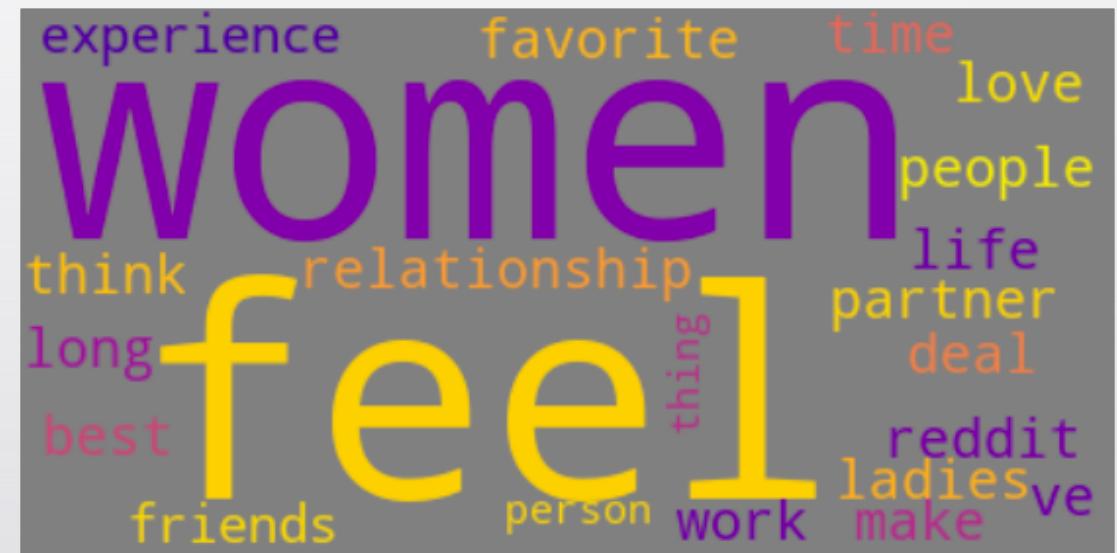
The screenshot shows the r/AskMen subreddit homepage. At the top, there's a dark banner with the text "IT'S TIME TO STOP" in white. Below it is the r/AskMen logo, which is a white cartoon character with red eyes and a small body. To the right of the logo, the text "r/AskMen" is displayed. Underneath the banner, there's a grey header bar with the text "About Community" and three dots on the right. The main content area contains a paragraph about the subreddit: "r/AskMen: the premier place to ask random strangers about the intricacies of the human condition. Just don't be an asshole about it. Also, go away." Below this text, there are two statistics: "2.0m Members" and "26.4k Online". On the right side of the page, there's a sidebar titled "r/AskMen Rules" which lists 13 rules, each with a dropdown arrow. The rules include: 1. Don't be an asshole / be respectful to others. 2. The title of your post must contain an actual, concise question. No clickbait titles. 3. Read the Frequently Asked Questions wiki and do a search before asking a question. 4. Do not make posts asking about a specific person or group's actions, behavior, or thinking. 5. Do not post pictures looking for affirmation of your appearance or body features. 6. YES/NO/"Does anyone else" questions will be removed at our discretion. 7. Do not post negative/forever alone rants. 8. No gift threads. 9. No agenda posting, and no brigading/complaining about other subs. 10. Medical advice is not allowed here. 11. Don't comment/post self-deprecating content looking to get karma or attention. 12. We do not allow surveys or promotional content. 13. Overly political questions will be removed.

25 Most Commonly Used Words by Subreddit

r/AskMen



r/AskWomen



//////
r/AskMen Only

woman
guys
men
friend

sex
don't
know
git
go

//////
r/AskWomen Only

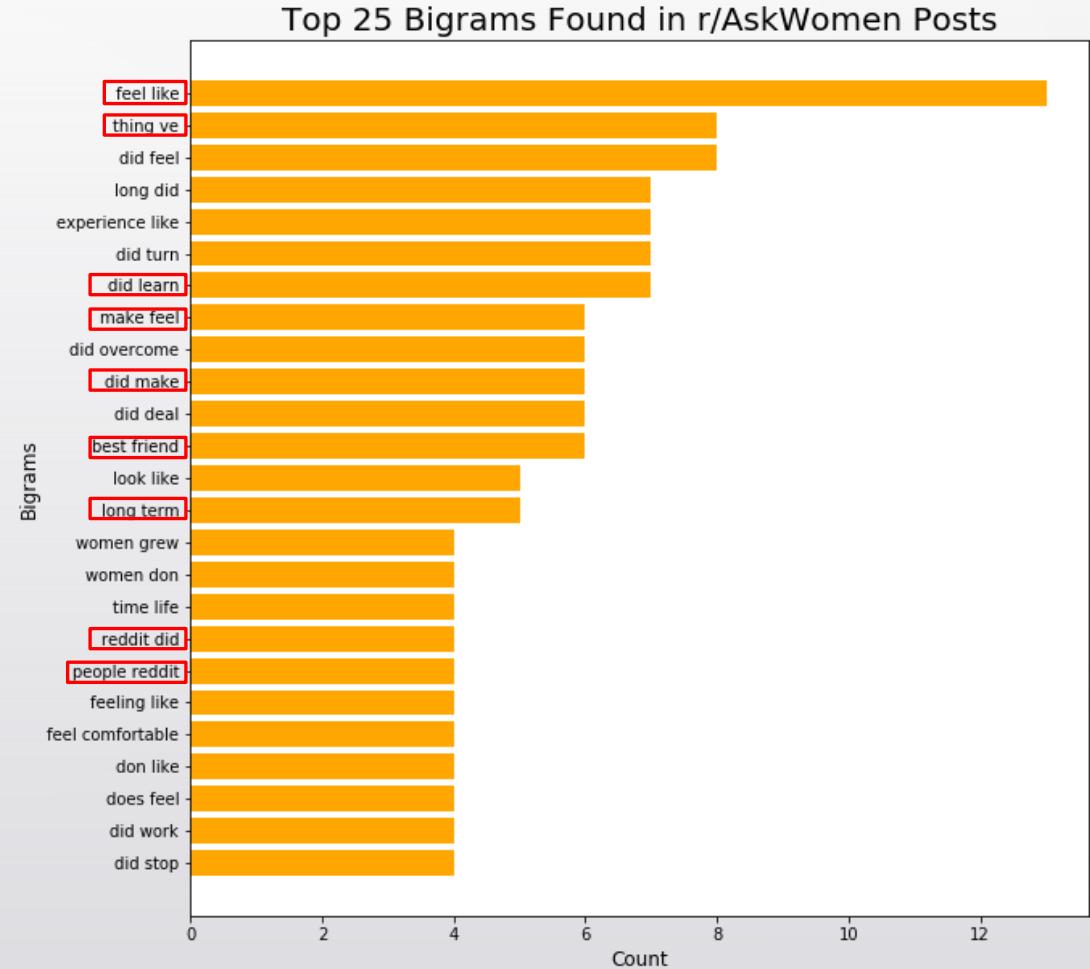
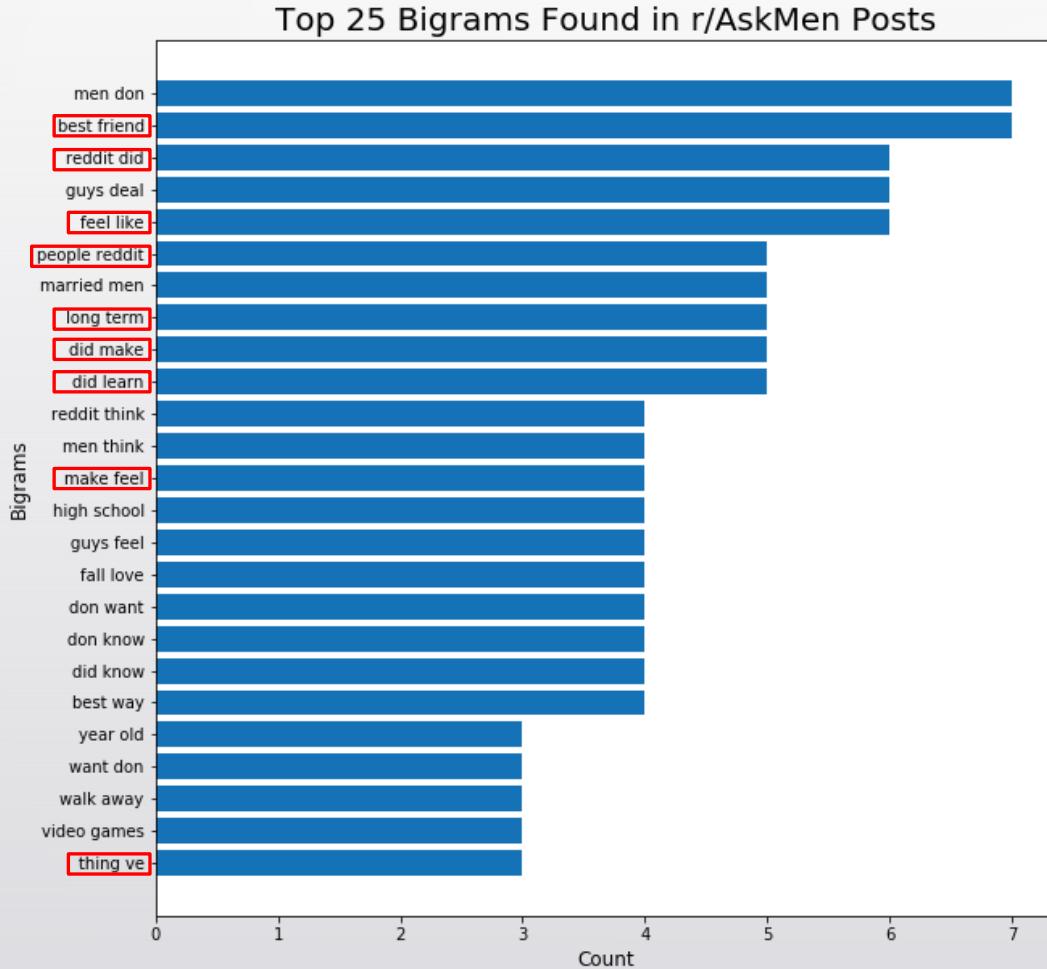
women
love
thing
work
experience
favorite
person

ladies

Both

best reddit feel life
long timemake think
relationship friends ve partner
people deal

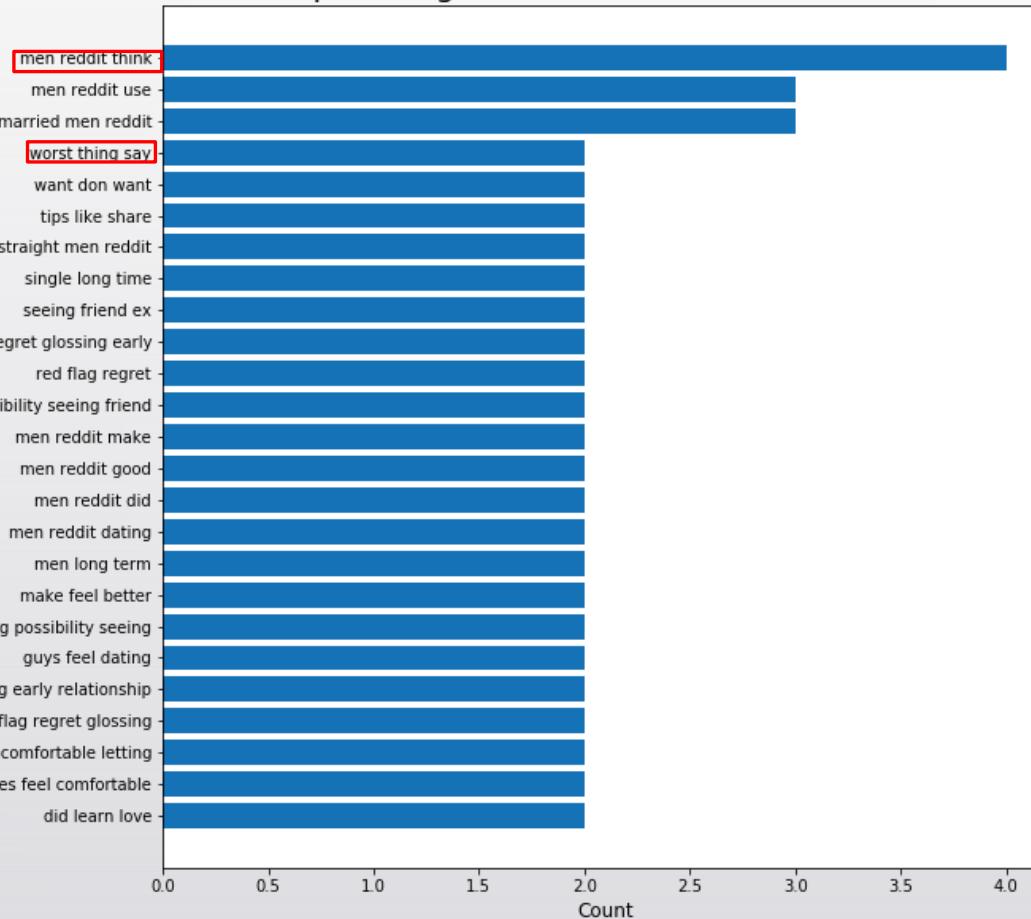
25 Most Common Bigrams



25 Most Common Trigrams

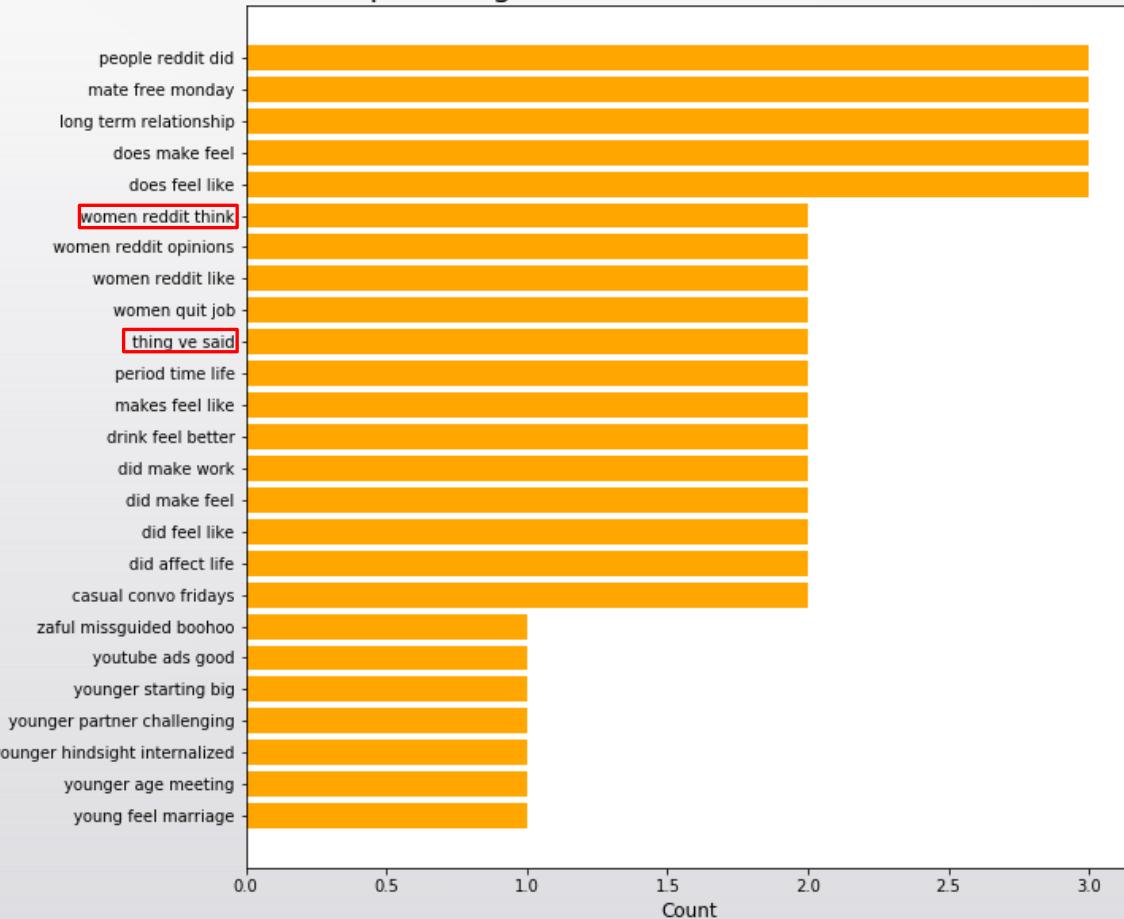
Trigrams

Top 25 Trigrams Found in r/AskMen Posts



Trigrams

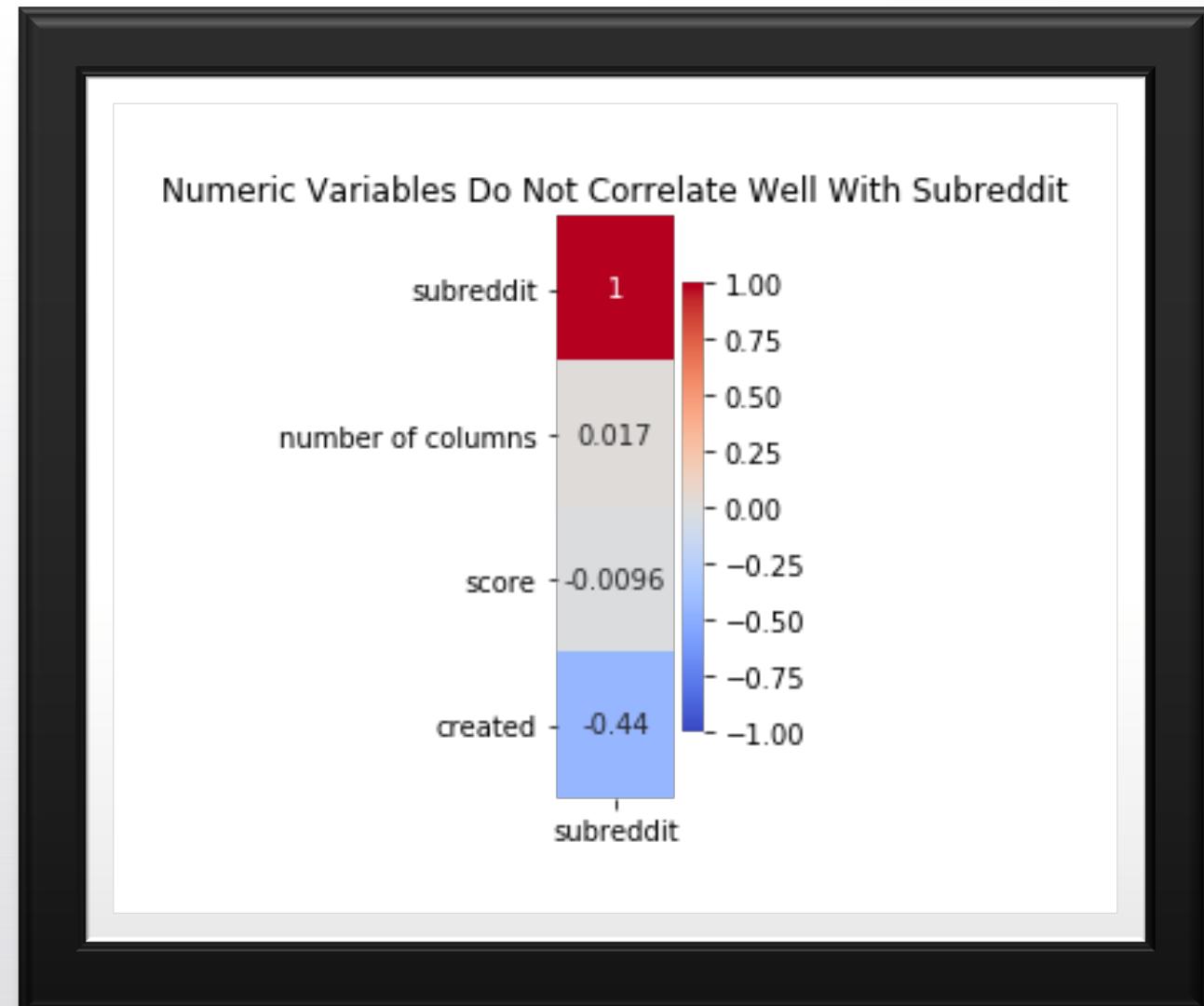
Top 25 Trigrams Found in r/AskWomen Posts





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

<u>Model</u>	<u>Vectorizer</u>	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?