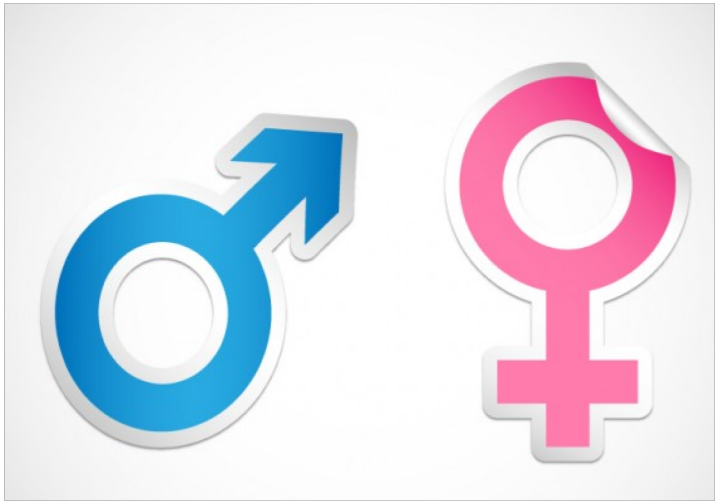


He said, She said: Discussions among Men and Women on Reddit

Amy DeSantis
General Assembly DSI
January 2, 2019

Research Question: How different are men's and women's subreddits?

How different are men's and women's subreddits?

- Who talks/posts more?
- What do they talk about?
- Can a machine tell the difference?

Subreddits by Gender

	Men	Women
Subscribers	2.9 k subscribers	38.2 k subscribers
Date conceived	8/1/2008	2/8/2008
# Posts in past 18 months	1569	8451
# Posts total	4,298	51,716
Alternate Threads	Nice guys	Nice girls
	Every Man Should Know	Girls Survival Guide
	Ask men	Ask women
	Men's rights /men's lib	Against men's rights
	Race-specific men	Race-specific women

Women's Subreddit Description

"A safe, respectful space to discuss the lives and stories of women of all backgrounds, and the current events which affect us. Trans people and especially trans feminine people are expressly welcome here. People of all genders are welcome; feminist cred appreciated but not required. Shaming women's choices and invalidating the perspectives of other women is not allowed here. Respect other life choices. We are baby and childless friendly. We are housewife and working woman friendly."

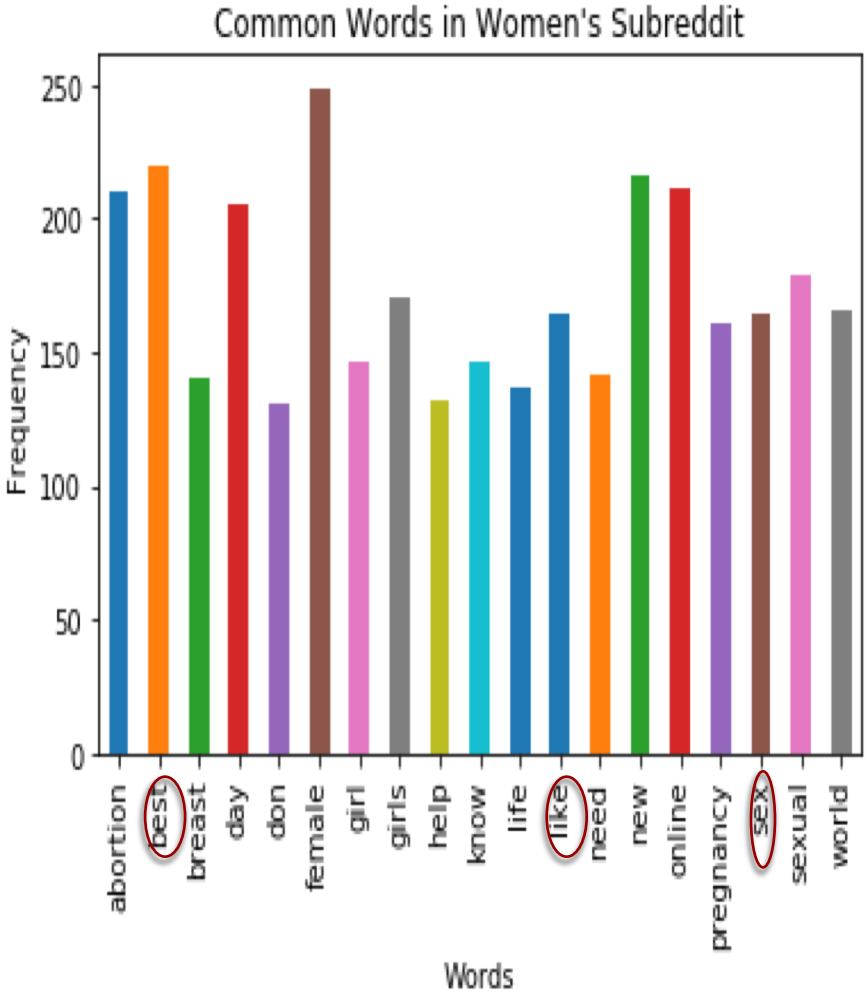
Women's Subreddit 12/31/2018

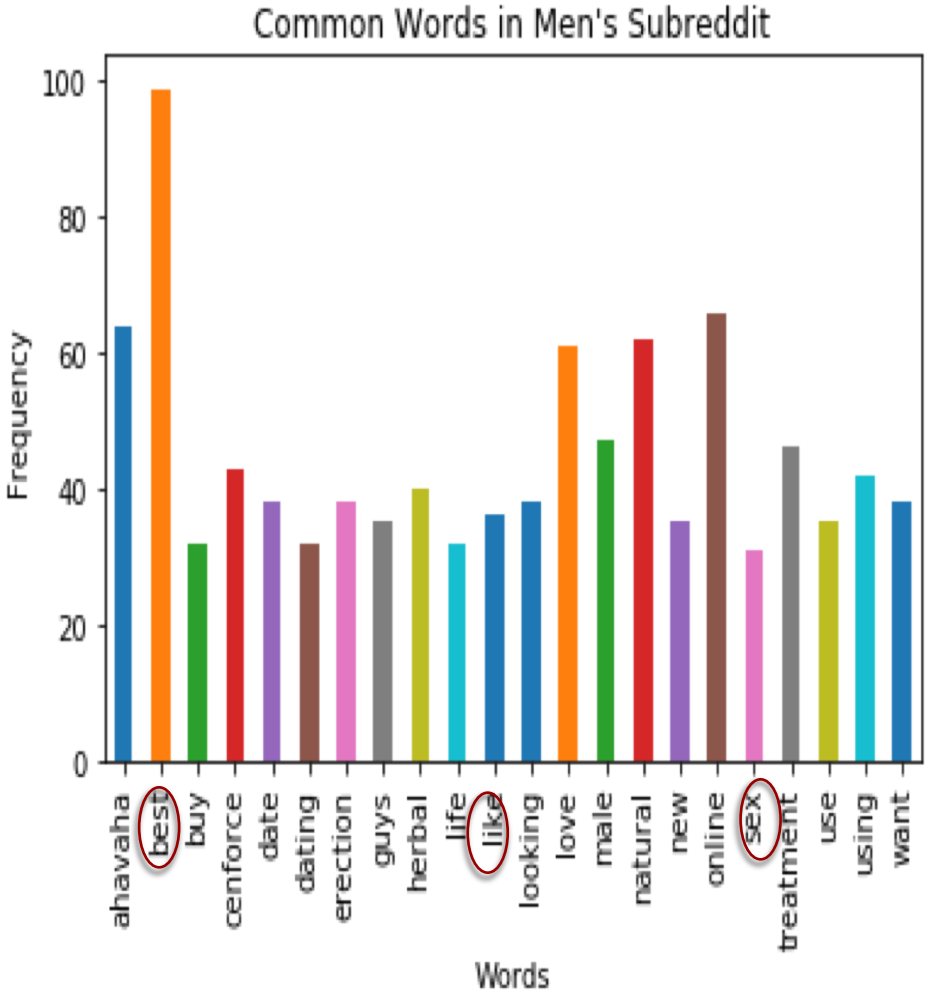
- Elizabeth Warren Announces She's Running For President in 2020. Who's Next?
- Do You Really Know What's In Your Tampons? This Woman Thinks You Should. There is no law requiring tampon makers to disclose their ingredients. Yet.
- Any other women get the feeling that the misogyny on Reddit is affecting their mental health?
- Third time I've been called "innocent". For some reason it feels like an insult. Thoughts?
- The Unstoppable Rise of 'Trash' Fashion - Many are upset that an Italian shoe designer is charging $535 for shoes held together by tape. But it's just the latest controversy in the long history of 'distressed' fashion.

Men's Subreddit 12/31/2018

- I Was a Bully in School. Now, My Son Was Becoming One Too. Here's How I Stopped It
- The Men 'Me Too' Left Behind: The movement toppled some alleged abusers, but others did just fine.
- How do you prepare yourself for a date?
- Man Period, Man-striation, Testosterone cycle… do you suffer from it, how do you deal with it?
- When did you realize you were gay?

Count Vector Analysis (ngrams 1 – 5)



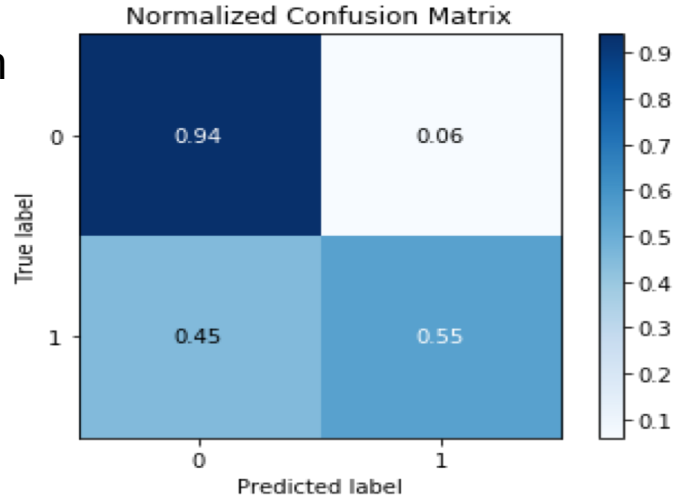


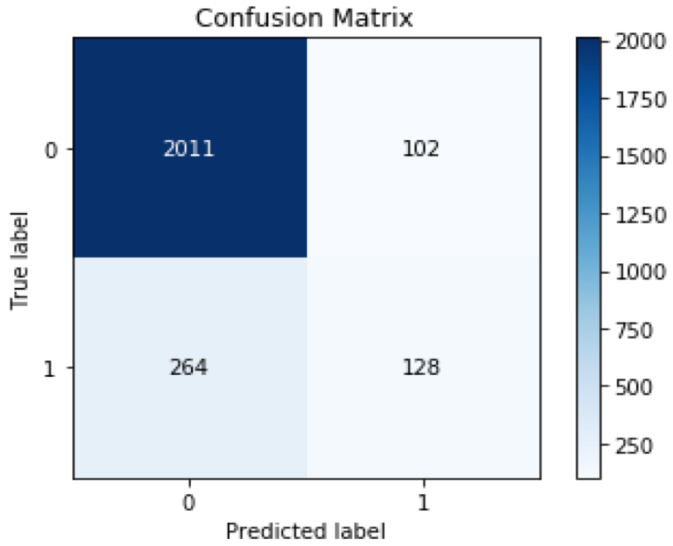
Logistic Regression (Weighted micro/macro Ave)

Ngram	Precision	Recall	Fscore
Count vector			
1,2	**0.97**	**0.97**	**0.97**
1,3	**0.98**	**0.98**	**0.98**
1,4 or 2,3 or 2,4	**0.97**	**0.97**	**0.97**
TF-IDF			
1,2 or 1,3 or 1,4	**0.89**	**0.87**	**0.83**
2,3	**0.89**	**0.88**	**0.85**
1,4	**0.89**	**0.88**	**0.84**

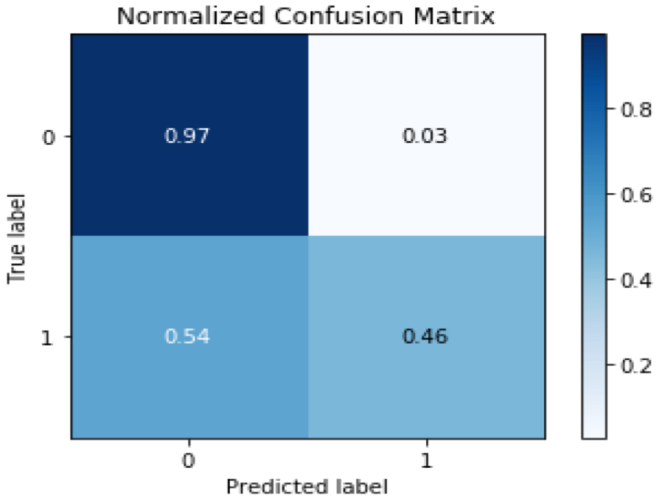
Random Forest: Confusion Matrices

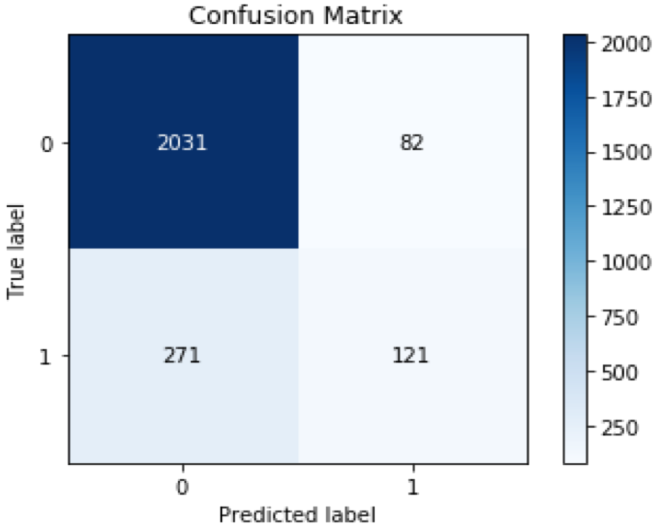
Ngram 1





Ngrams 1,2





Summary

- **Ngrams are flexible 1,2 and 1,3 very similar,** high accuracy for all models
- **Count Vectorizer more accurate than TF-IDF**
- **Random Forest – better accuracy when predicting women's subreddit.** More frequently mislabels the men's subreddit (as women's)
    - Possibly due to the larger sample of women's posts?
- Stratified on Y variable / Bootstrapping for Random Forest
    - Might improve with AUC-ROC analysis
- **Accuracy Scores** were generally .83 - .98
- **All methods overfit** –train scores 0.05 – 0.10 higher than test scores

Thank you!

Extra Slides

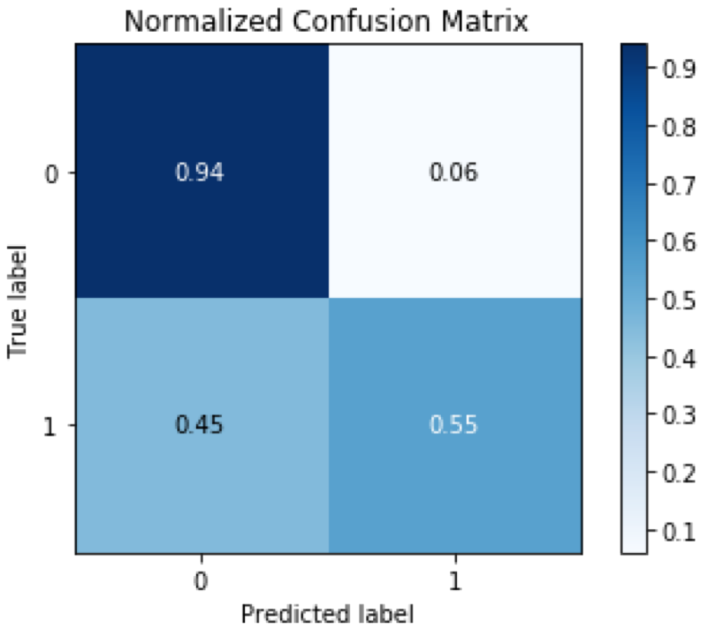
Grid Search Params

```
gs.best_score:  0.8435129740518962
gs.best_params:  {'max_depth': 5,
'min_samples_split': 3, 'n_estimators': 10}
best_gs = RandomForestClassifier(bootstrap=True,
class_weight=None, criterion='gini',
            max_depth=5, max_features='auto',
max_leaf_nodes=None,
            min_impurity_decrease=0.0,
min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=3,
            min_weight_fraction_leaf=0.0,
n_estimators=10, n_jobs=None,
            oob_score=False, random_state=None,
verbose=0,
            warm_start=False)
best_gs.score:  0.8435129740518962
```

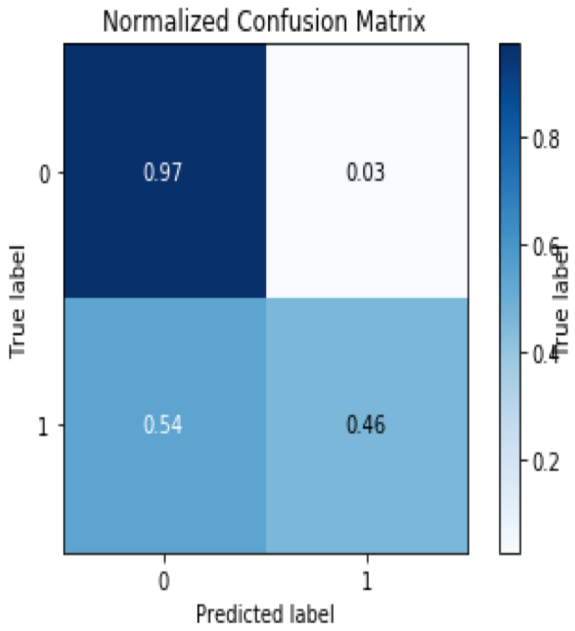
```
gs.best_score:  0.8439121756487026
gs.best_params:  {'max_depth': 5,
'min_samples_split': 4, 'n_estimators': 10}
best_gs = RandomForestClassifier(bootstrap=True,
class_weight=None, criterion='gini',
            max_depth=5, max_features='auto',
max_leaf_nodes=None,
            min_impurity_decrease=0.0,
min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=4,
            min_weight_fraction_leaf=0.0,
n_estimators=10, n_jobs=None,
            oob_score=False, random_state=None,
verbose=0,
            warm_start=False)
best_gs.score:  0.8435129740518962
```

Random Forest: 18 MONTHS

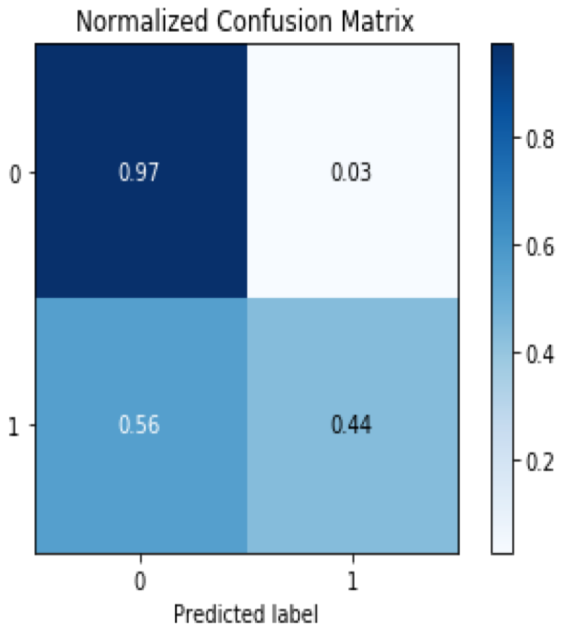
Test Ngrams 1,1



Test Ngrams 1,2

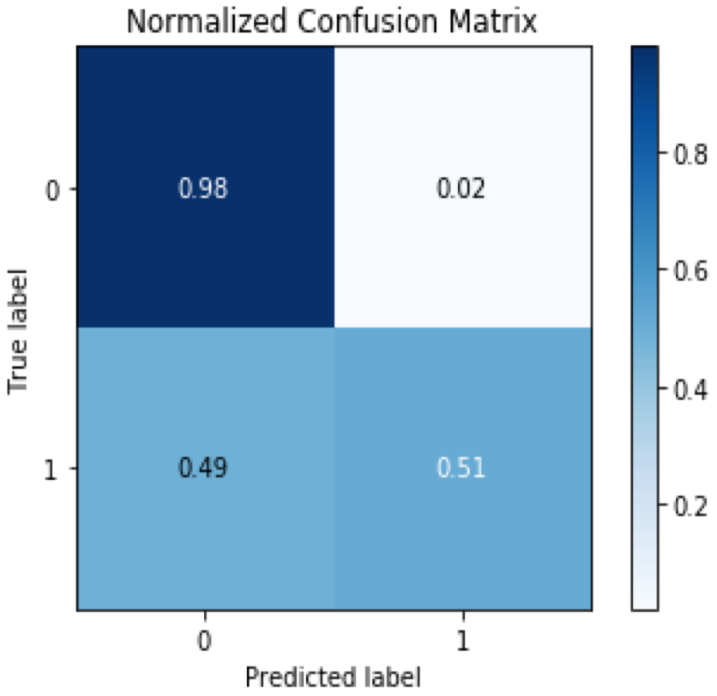


Test Ngrams 1,3

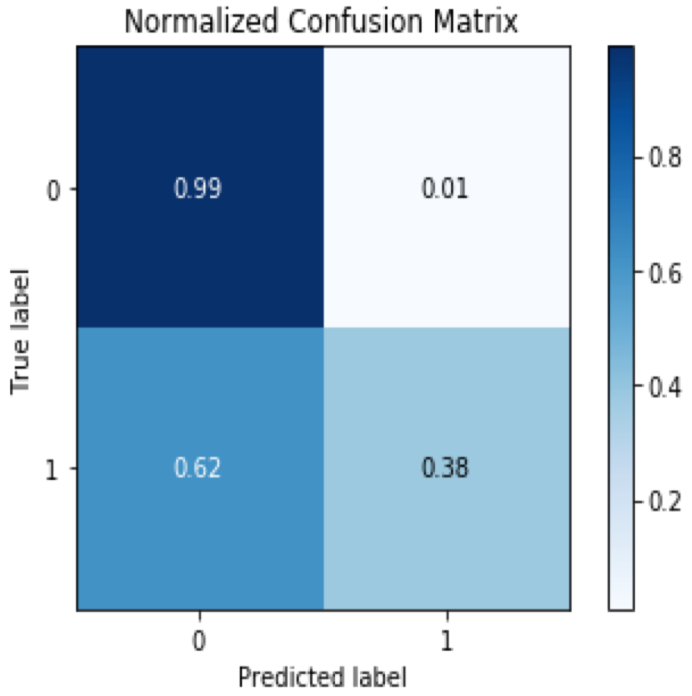


Random Forest: ALL POSTS

Test Ngrams 1,2

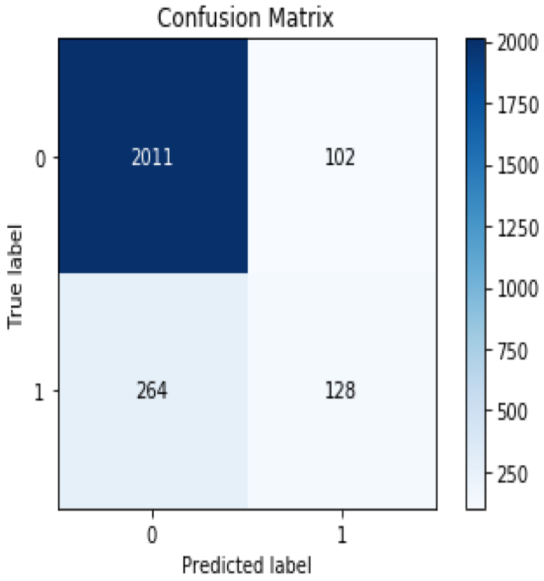


Overall Ngrams 1,2

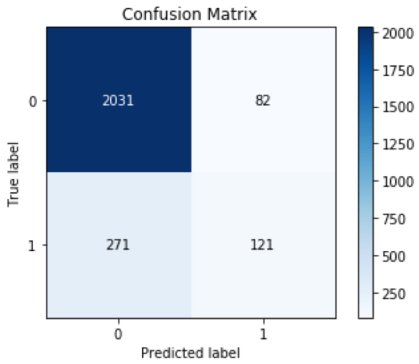


Random Forest: 18 MONTHS

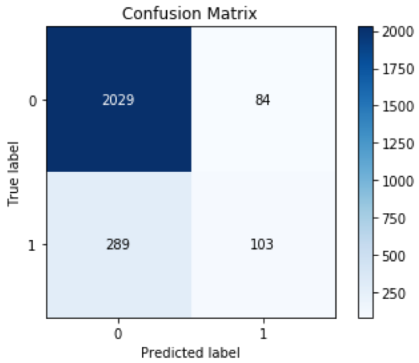
Test Ngrams 1,1



Test Ngrams 1,2

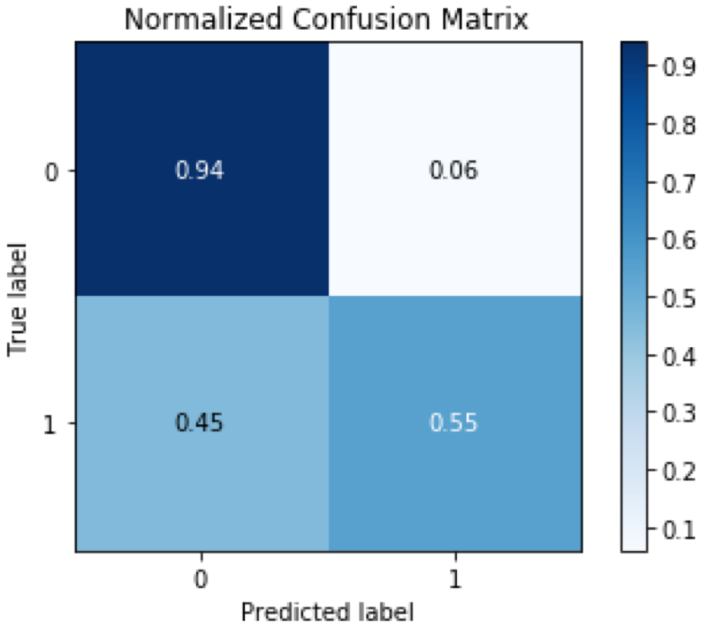


Test Ngrams 1,3

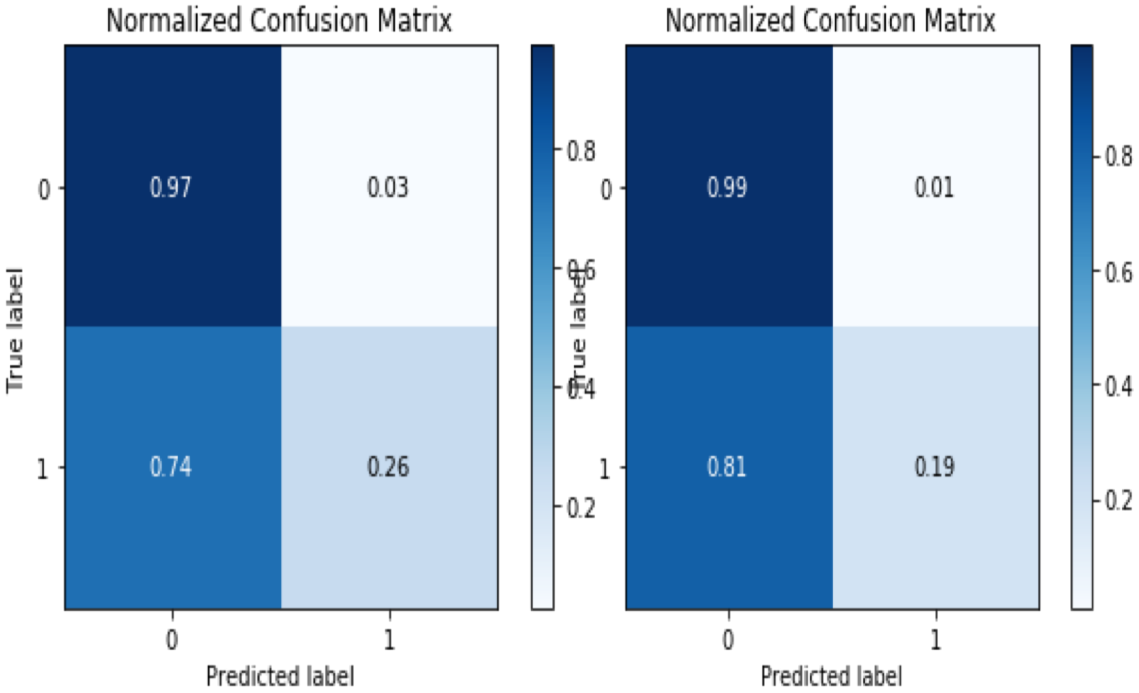


Random Forest: 18 MONTHS

Overall Ngrams 1,2

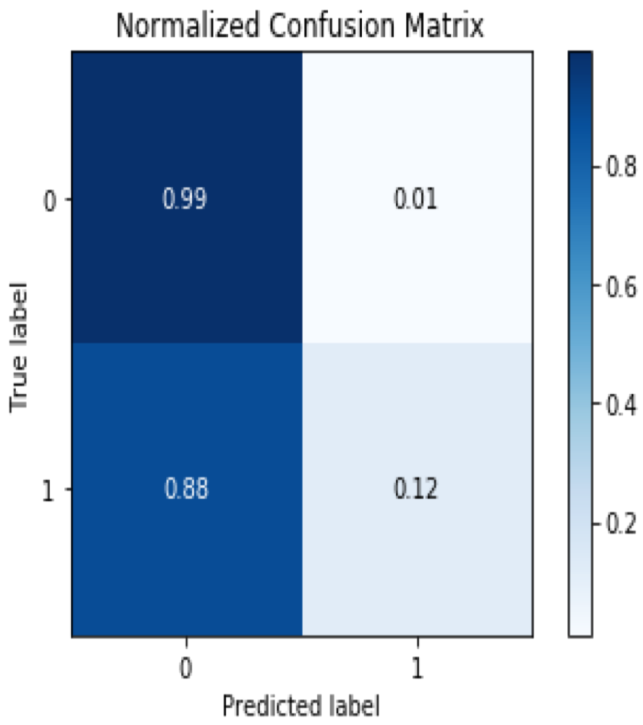


10k posts Ngrams 2,5

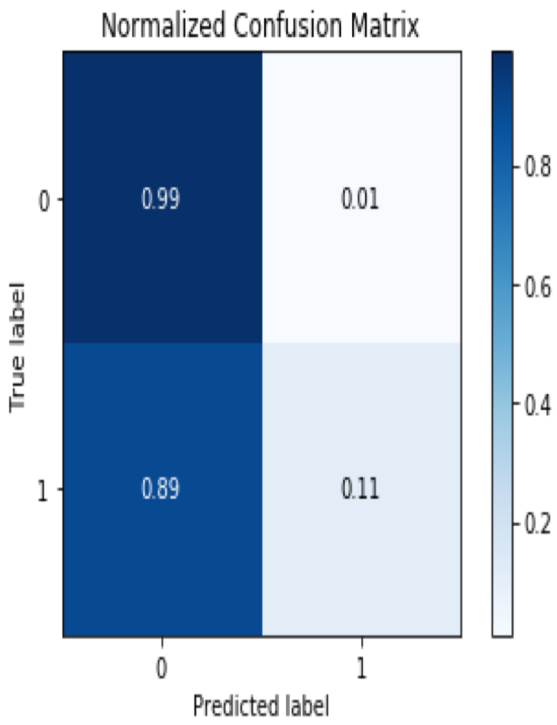


Random Forest: 18 MONTHS

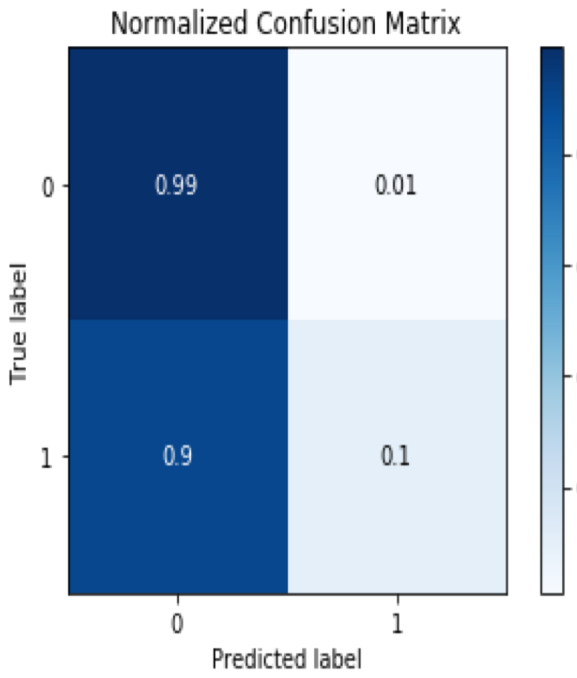
Test Ngrams 2,3



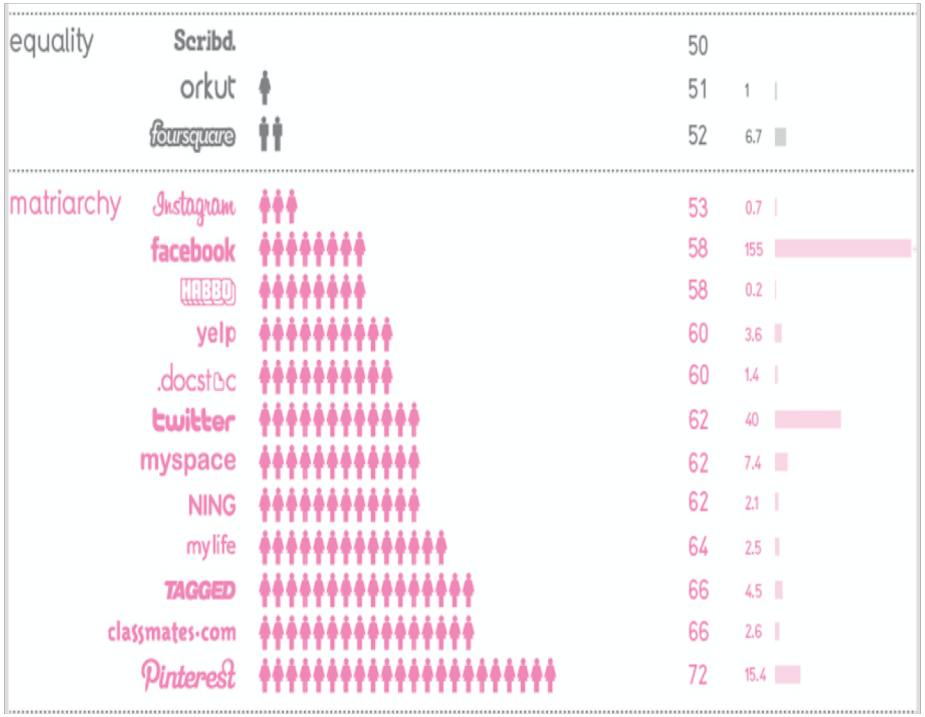
Overall Ngrams 2,4

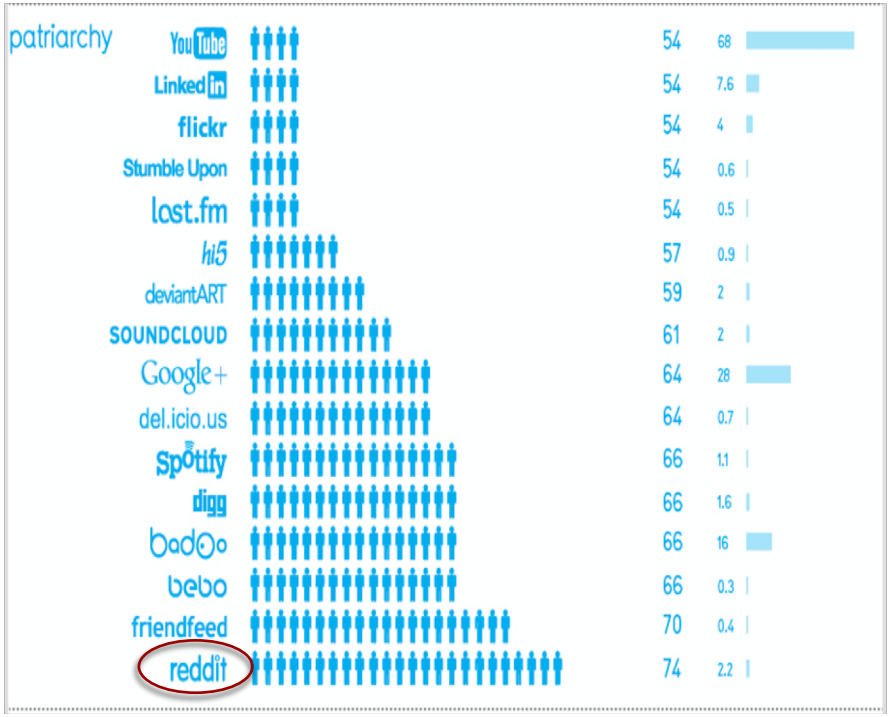


10k posts Ngrams 2,5



Social Networking: Monthly Use by Gender





https://www.huffingtonpost.com/2012/06/20/social-media-by-gender-women-pinterest-men-reddit-infographic_n_1613812.html