Classification models on subreddits:’AskWomen’ and ‘AskMen’

•••

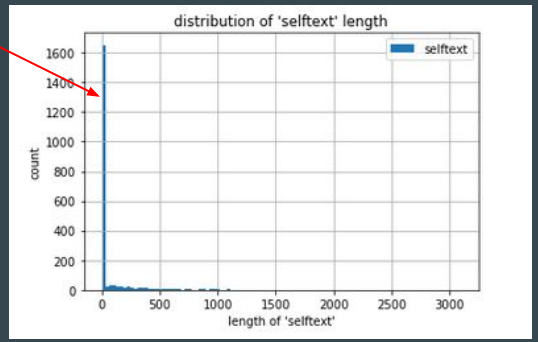
Yonghe

AskWomen vs AskMen

- Fetched 2,000 posts from each subreddit
- Only consider columns ‘title’ and ‘selftext’
- Balance Dataset

Data Cleaning

- No null values found
- Outliers - handle during hyperparameter tuning (max_features, max_depth)
- Column ‘selftext’: replace system-generated ‘[removed]’ with ‘’

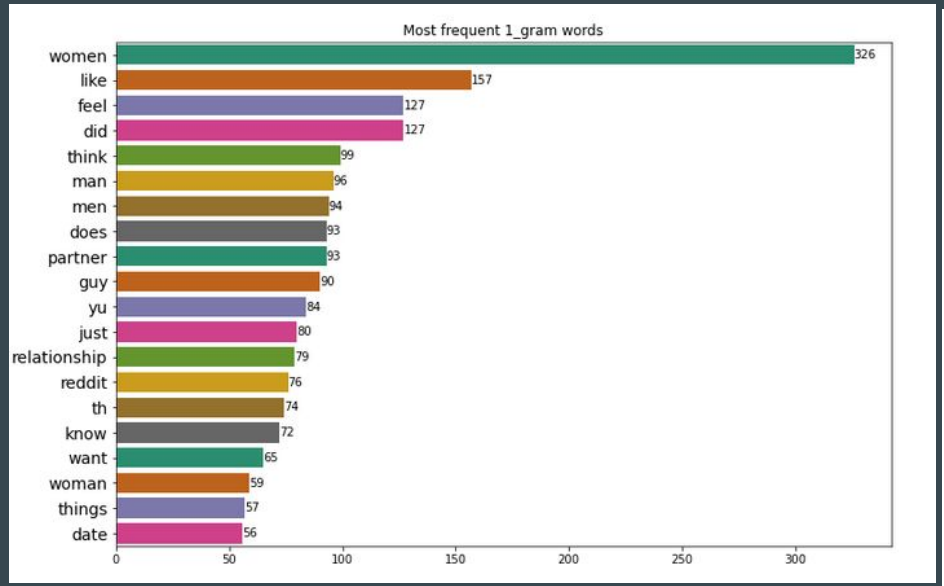


NLP

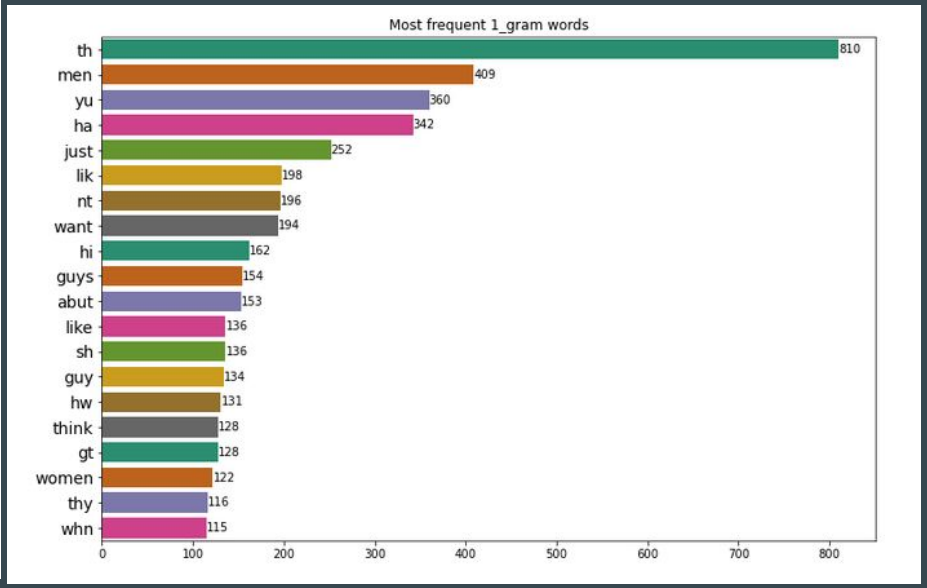
- Used TF-IDF Vectorizer
- Tried different NLP techniques during hyperparameter tuning
  - Tokenizer:
    - Regex_tokenizer r'(?u)\b\w\w+\b'
    - Regex_tokenizer + Lemmatizer
    - Regex_tokenizer + Stemmer
  - Stop_words: [None, ‘english’]
  - Ngram range: (1,1),(1,2),(1,3)
  - max_features: [400,500,600,700,900]

Top 15 Most Frequent 1_gram Words (excluding stopwords)

AskWomen

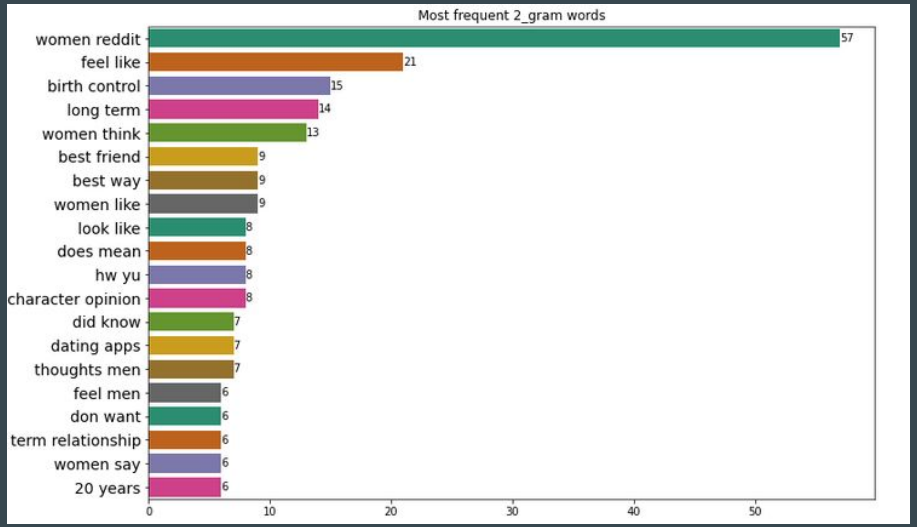


AskMen

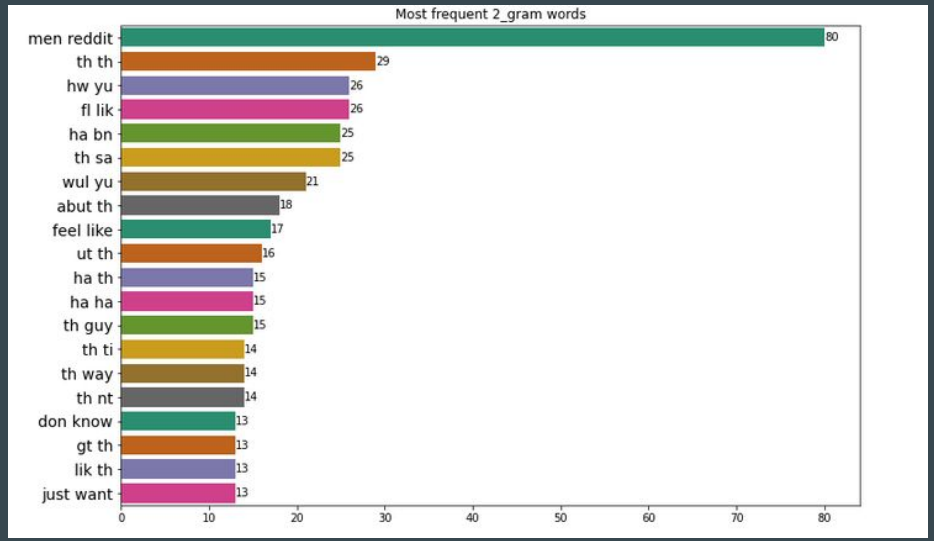


Top 15 Most Frequent 2_gram Words (excluding stopwords)

AskWomen



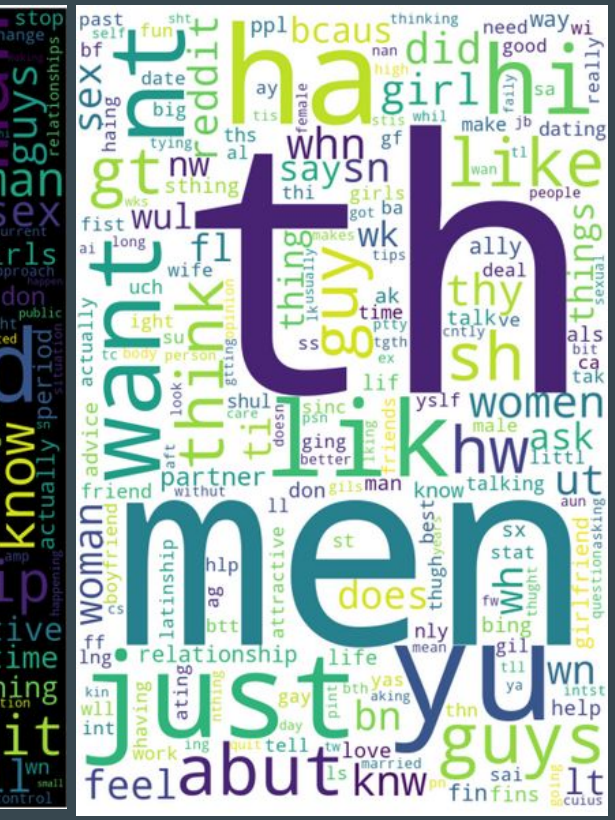
AskMen



Word Clouds of Top 200 Most Frequent Words

AskWomen

AskMen



- Abbreviations occur more frequently in 'AskMen' than 'AskWomen'
- Many abbreviations in ‘AskMen’ are stopwords
- Comparing abbreviations and stopwords will be critical to classify these two subreddits. We can verify it thought feature importance analysis

Metrics

- Roc_auc_score
- Accuracy score:
  - baseline score: 0.5 (balance dataset)
- perc_diff < 5%

Baseline Accuracy_Score: 0.5

- Dataset contains 2000 posts from ‘AskWomen’ and 2000 posts from “AskMen”
- The expected chance to guess correctly a post from "AskMen" or "AskWomen" is 50%

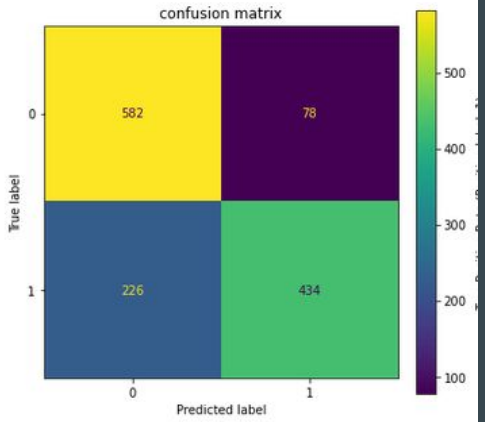
Model 1: Random Forest

```
roc_auc_score on training set: 0.860
roc_auc_score on testing set: 0.824
perc_diff: 4.1 %


accuracy_score on training set: 0.767
accuracy_score on testing set: 0.770
perc_diff: 0.3 %
```

```
best_params:
    classifier__max_depth  :  6
    classifier__n_estimators  :  500
    tvec__max_features  :  350
    tvec__ngram_range  :  (1, 3)
    tvec__stop_words  :  None
    tvec__tokenizer  :  None
```

```
confusion matrix:
                pred AskWomen  pred AskMen
actual AskWomen            582           78
actual AskMen              226          434
```



Metrics:

- roc_auc_score: 0.824,
  - perc_diff: 4.1%
- accuracy_score increases from 0.5 to 0.77
  - perc_diff : 0.3%

Observations from best_params:

- Only use 350 features
- Better result without removal of stopwords
- Better result without lemmatizing or stemming

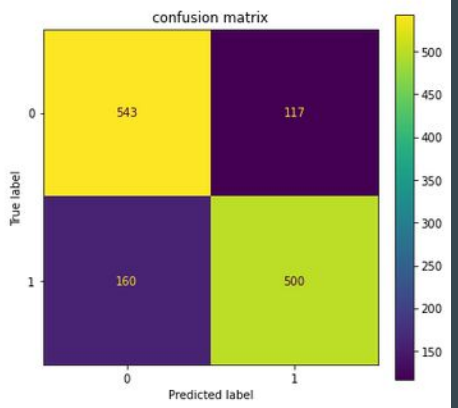
Model 2: Logistic Regression

```
roc_auc_score on training set: 0.859
roc_auc_score on testing set: 0.851
perc_diff: 0.9 %


accuracy_score on training set: 0.785
accuracy_score on testing set: 0.790
perc_diff: 0.7 %
```

```
best_params:
    classifier__C  :  1
    classifier__penalty  :  l1
    tvec__max_features  :  1000
    tvec__ngram_range  :  (1, 3)
    tvec__stop_words  :  None
    tvec__tokenizer  :  None
```

```
confusion matrix:
                pred AskWomen  pred AskMen
actual AskWomen            543          117
actual AskMen              160          500
```



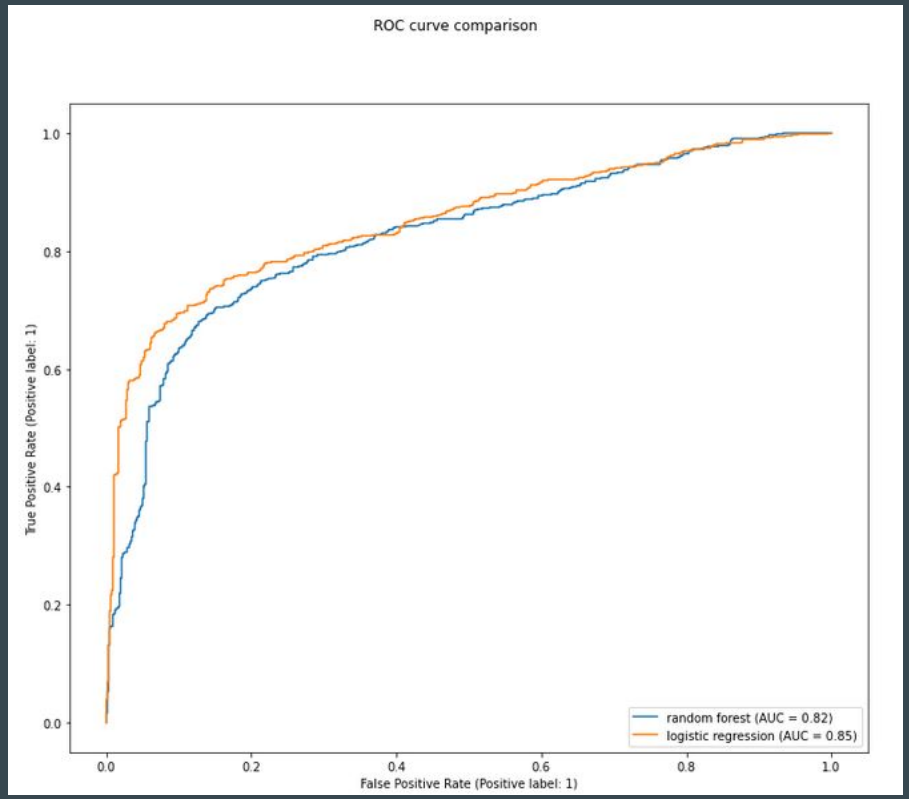
Metrics:

- roc_auc_socre: 0.851
- accuracy_score increases from 77% to 79%

Observations from best_params:

- Better result without removal of stopwords
- Better result without lemmatizing or stemming

Model performance comparison



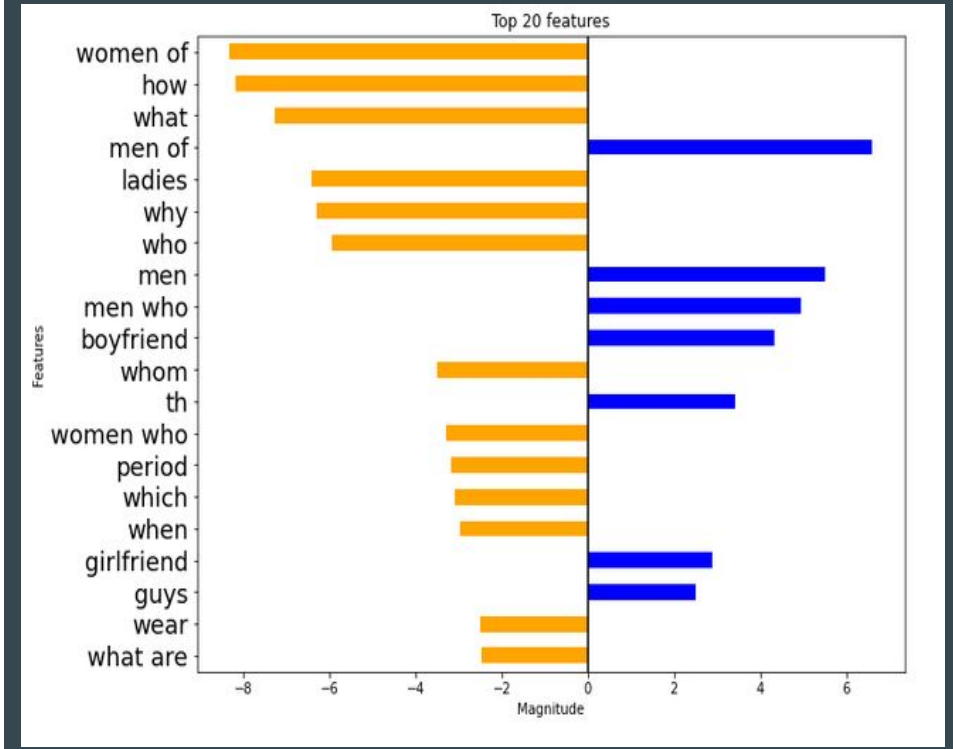
- Logistic Regression Model outperforms Random Forest Model in almost all thresholds
- Logistic Regression Model is chosen to be final model

Feature Importance Analysis for Logistic Regression Model

AskWomen: 0

AskMen: 1

feature	coef
women of	-8.32
how	-8.17
what	-7.25
men of	6.62
ladies	-6.41
why	-6.29
who	-5.93
men	5.52
men who	4.96
boyfriend	4.33
whom	-3.49
th	3.42
women who	-3.28
period	-3.16
which	-3.08
when	-2.98
girlfriend	2.89
guys	2.52
wear	-2.50
what are	-2.47



- We mapped “AskMen” : 1, “AskWomen” : 0
- Features with positive coef such as 'men of', 'men' and 'men who' are favorable to 'AskMen
- Features with negative coef such as 'women of', 'how' and 'what' are supporting 'AskWomen'
- Many stopwords like ‘how’,’what’ and ‘why’ are features with negative coef. In other words, they are favorable to “AskWomen”

Q & A