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

•••

Yonghe

Only one post came from "AskMen"

Our final model can predict them all correctly

Data Collection

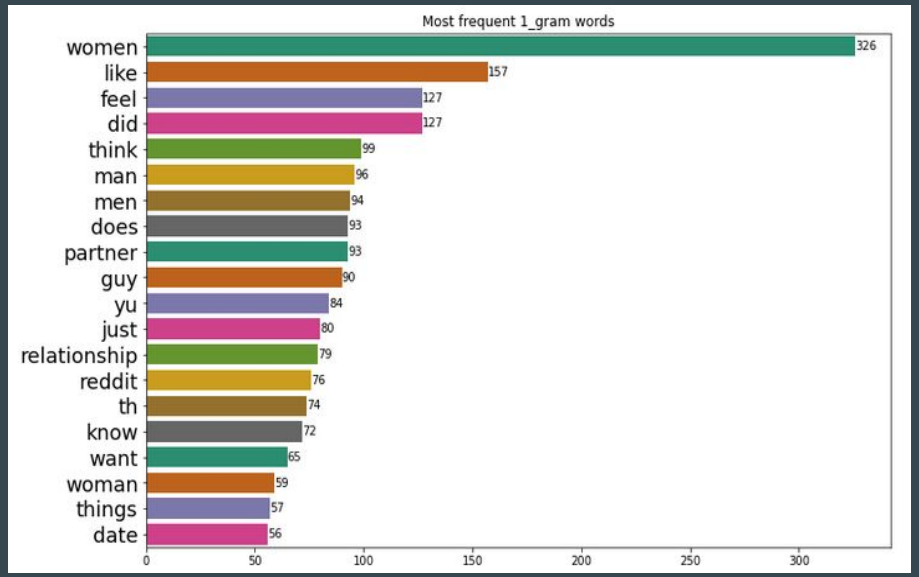
- Fetched 2,000 posts from each subreddit
- Only consider columns ‘title’ and ‘selftext’
- Balance Dataset

NLP

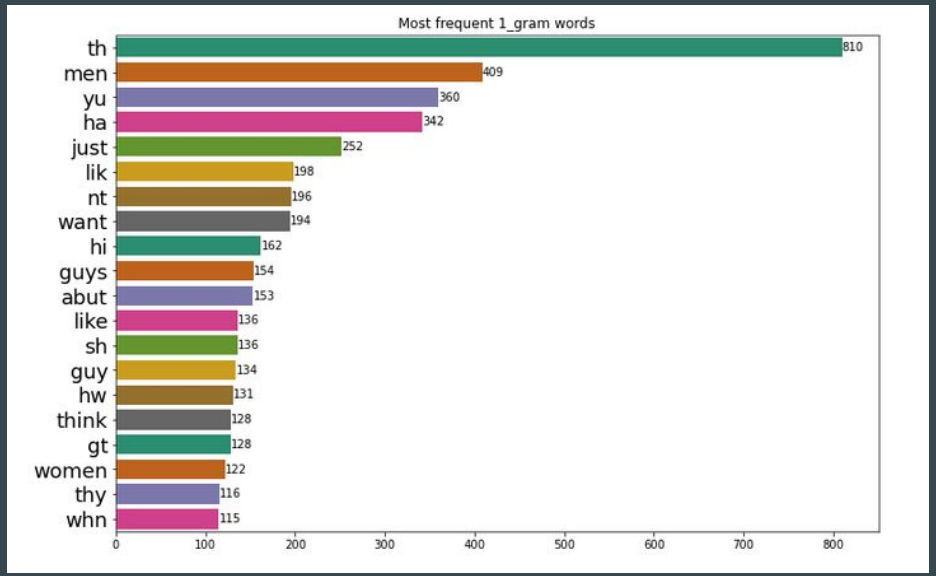
- Tried Count Vectorizer and TF-IDF Vectorizer
- Tried different NLP techniques during hyperparameter tuning
  - Tokenizer:
    - Regex_tokenizer r'(?u)\b\w\w+\b'
    - Regex_tokenizer + Lemmatizer
    - Regex_tokenizer + Stemmer
  - Stop_words: [None, ‘english’]
  - Ngram range: (1,1),(1,2),(1,3)
 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

Most Common 1_gram Words (excluding stopwords)

AskWomen

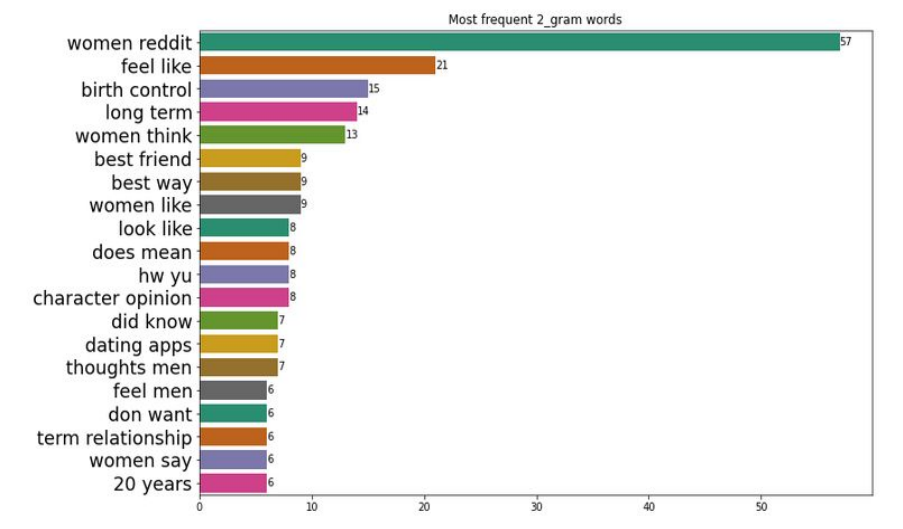


AskMen

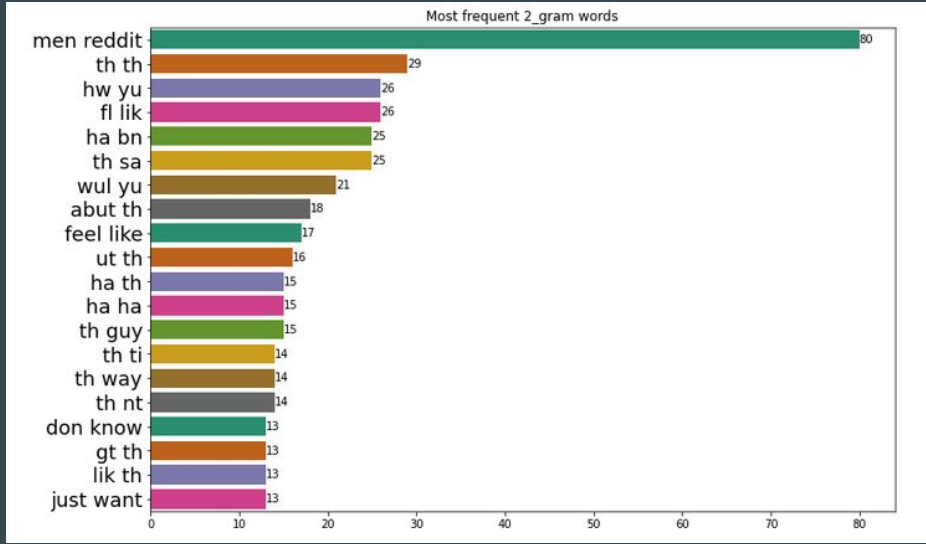


Most Frequent 2_gram Words (excluding stopwords)

AskWomen



AskMen



Top 200 Most Common Words

AskWomen

AskMen

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

Metrics

- Roc_auc_score
  - baseline score: 0.5

- Accuracy score:
  - baseline score: 0.5 (balance dataset)

- perc_diff < 5%

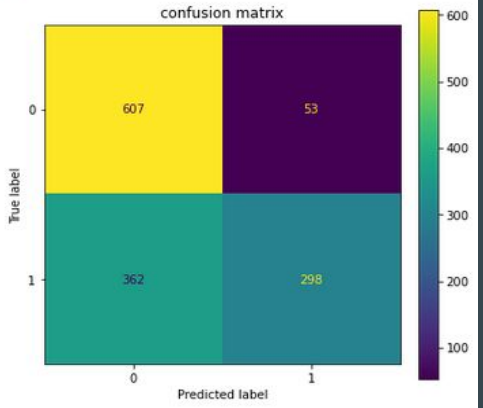
Model 1: CountVectorizer + RandomForest

```
roc_auc_score on training set: 0.837
roc_auc_score on testing set: 0.819
perc_diff: 2.2 %


accuracy_score on training set: 0.705
accuracy_score on testing set: 0.686
perc_diff: 2.8 %
```

```
confusion matrix:
                pred AskWomen  pred AskMen
actual AskWomen            607           53
actual AskMen              362          298

plot confusion matrix and ROC curve
```




```
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
```

Metrics:

- roc_auc_score increase from 0.5 to 0.82,
  - perc_diff: 2.2%
- accuracy_score increases from 0.5 to 0.68
  - perc_diff : 2.8%

Observations of best_params:

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

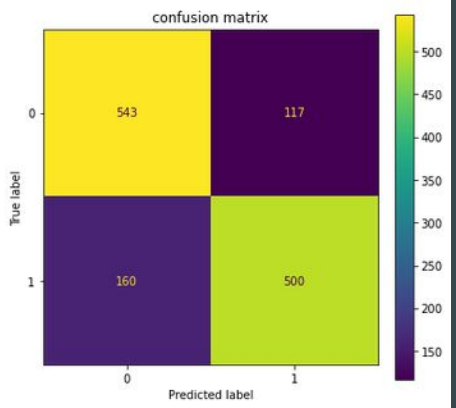
Model 2: TfidfVectorizer + Logistic Regression

```
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
```



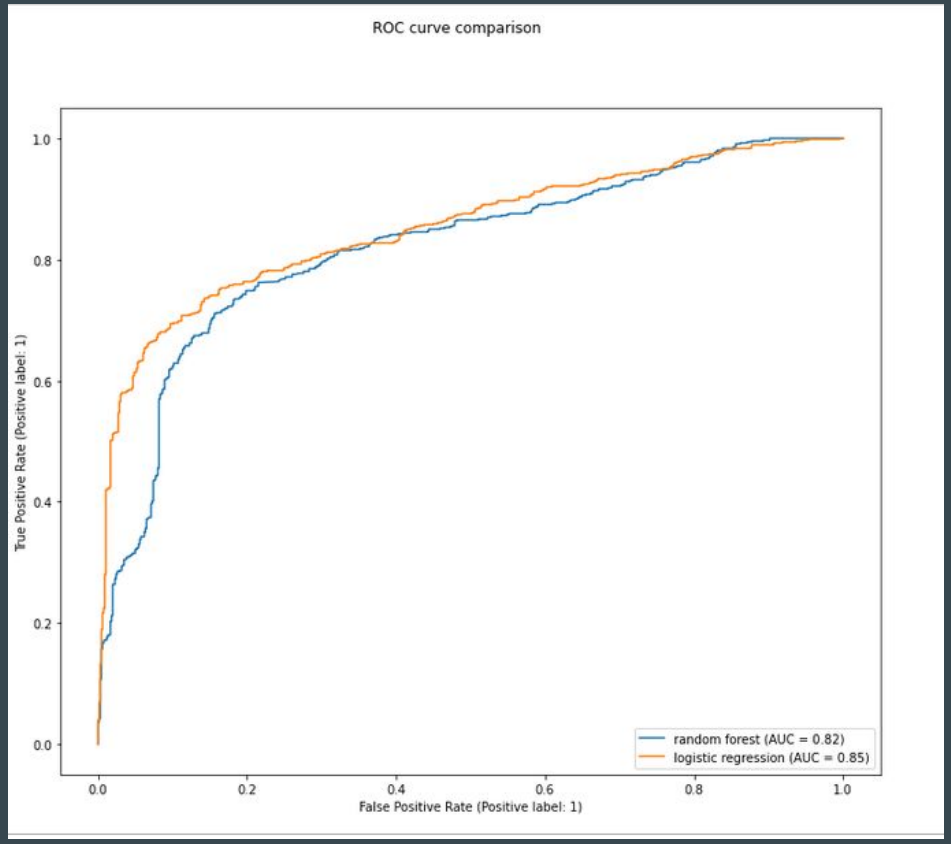
Metrics:

- Roc_auc_score increase from 0.82 to 0.85
  - perc_diff: 0.9%
- accuracy_score increases from 0.68 to 0.79
  - perc_diff: 0.7%

Observations of best_params:

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

Model performance comparison



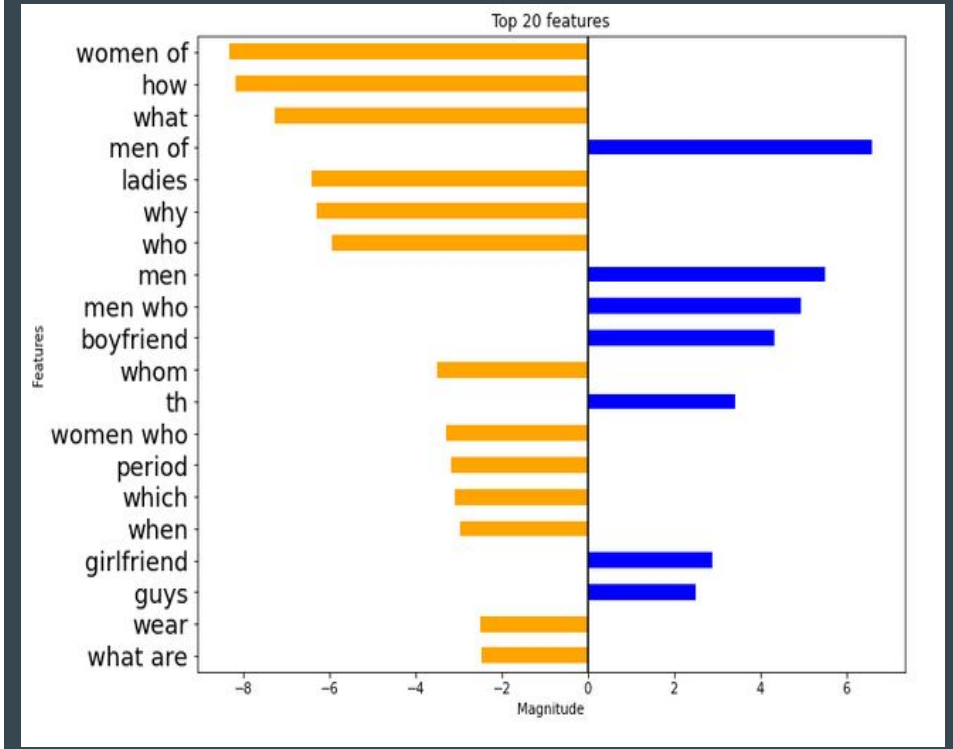
- Logistic Regression Model outperforms Random Forest Model in almost all thresholds
- Logistic Regression Model is chosen to be final model

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
- Positive coefs, shown in blue color in the chart, indicate the features are more common in “AskMen’
- Negative coefs, shown in orange color in the chart, indicate the features are more common in “AskWoMen’
- Top features contains a lot of stopwords, mostly because abbreviations of stopwords such as 'whn','hw' occurs much more frequently in "AskMen".

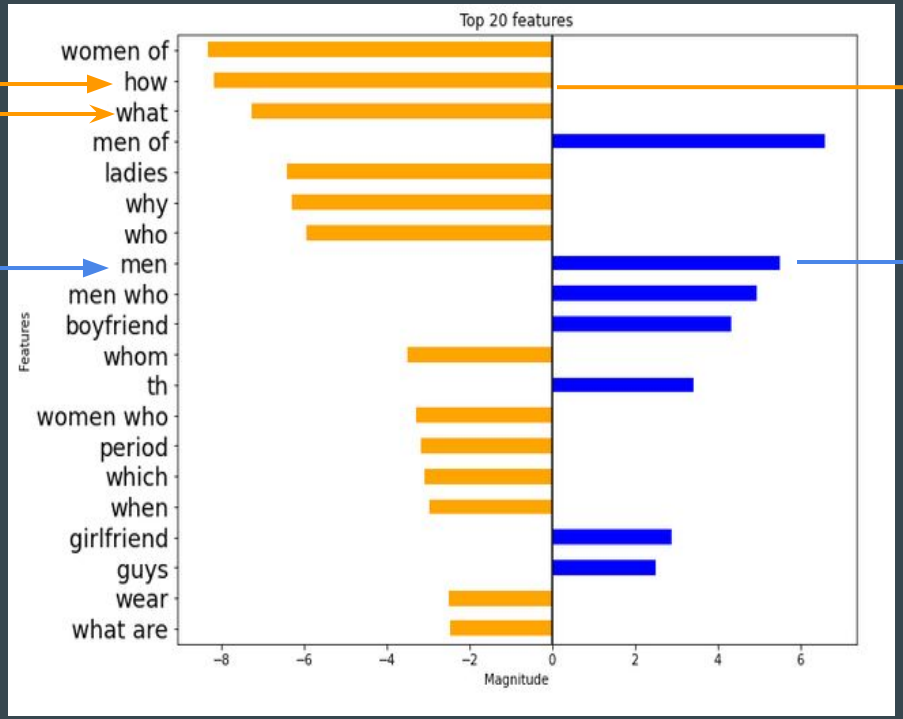
Important Feature Demonstration

how is this fair or even legal

What's your best decision so far in 2022?

Men! Is it bullshit or should I let him win me back?

What are your thoughts on Andrew Tate and his takes?



AskWomen: 0

AskMen: 1

Q & A