Project 3
WebAPI & NLP

by Radchenko Gleb

DSI-822

Project statement:

Our marketing company is looking for a way to improve effectiveness of advertising campaignes running on Reddit. For that purpose stakeholders would like to know how accurately we can predict, if a user a woman or a man based on the subreddits they are publishing submissions to.

Project success will be evaluated using resulting f1 score and accuracy score.

Subreddits

/r AskWoman

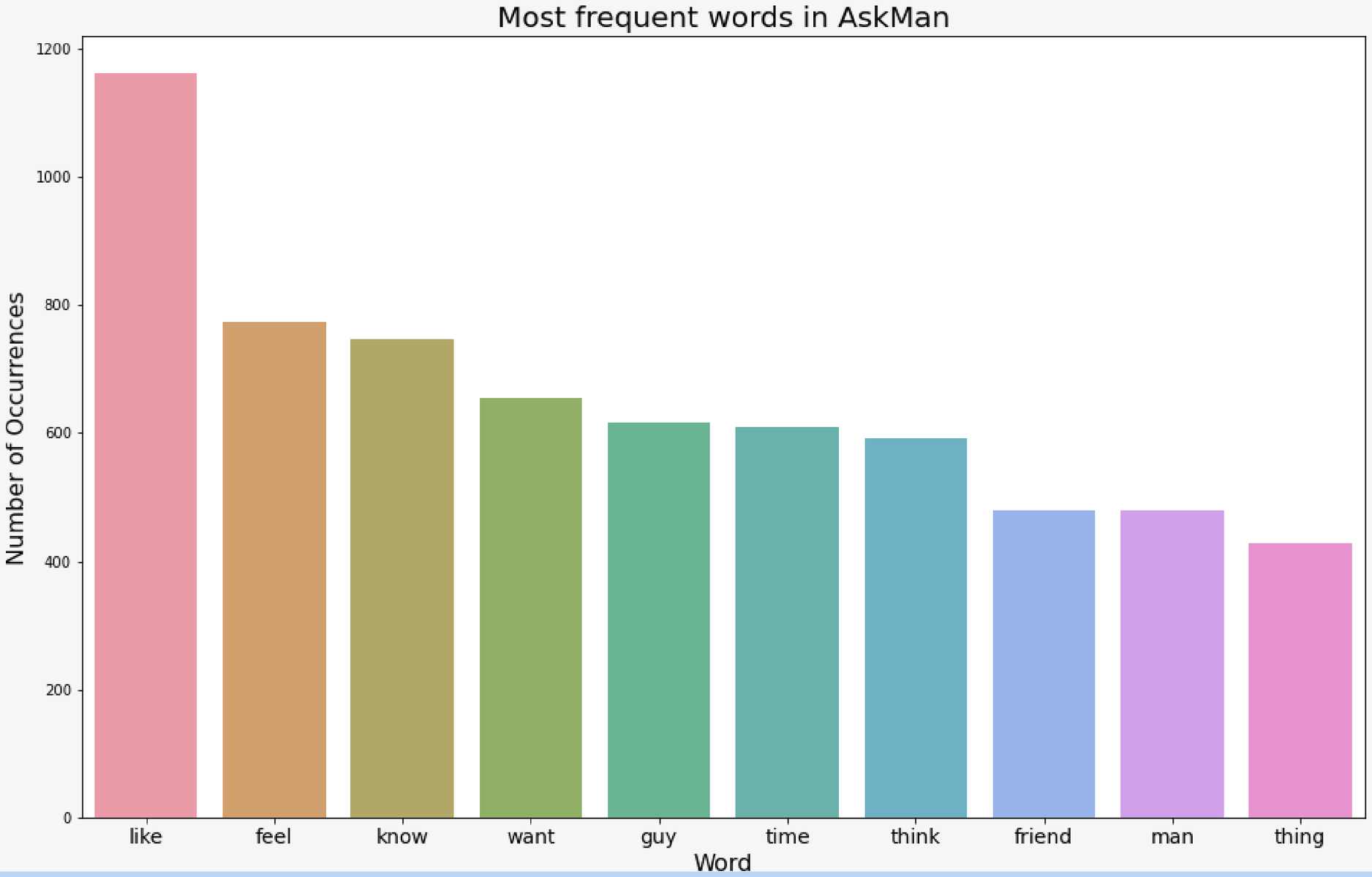
/r AskMan

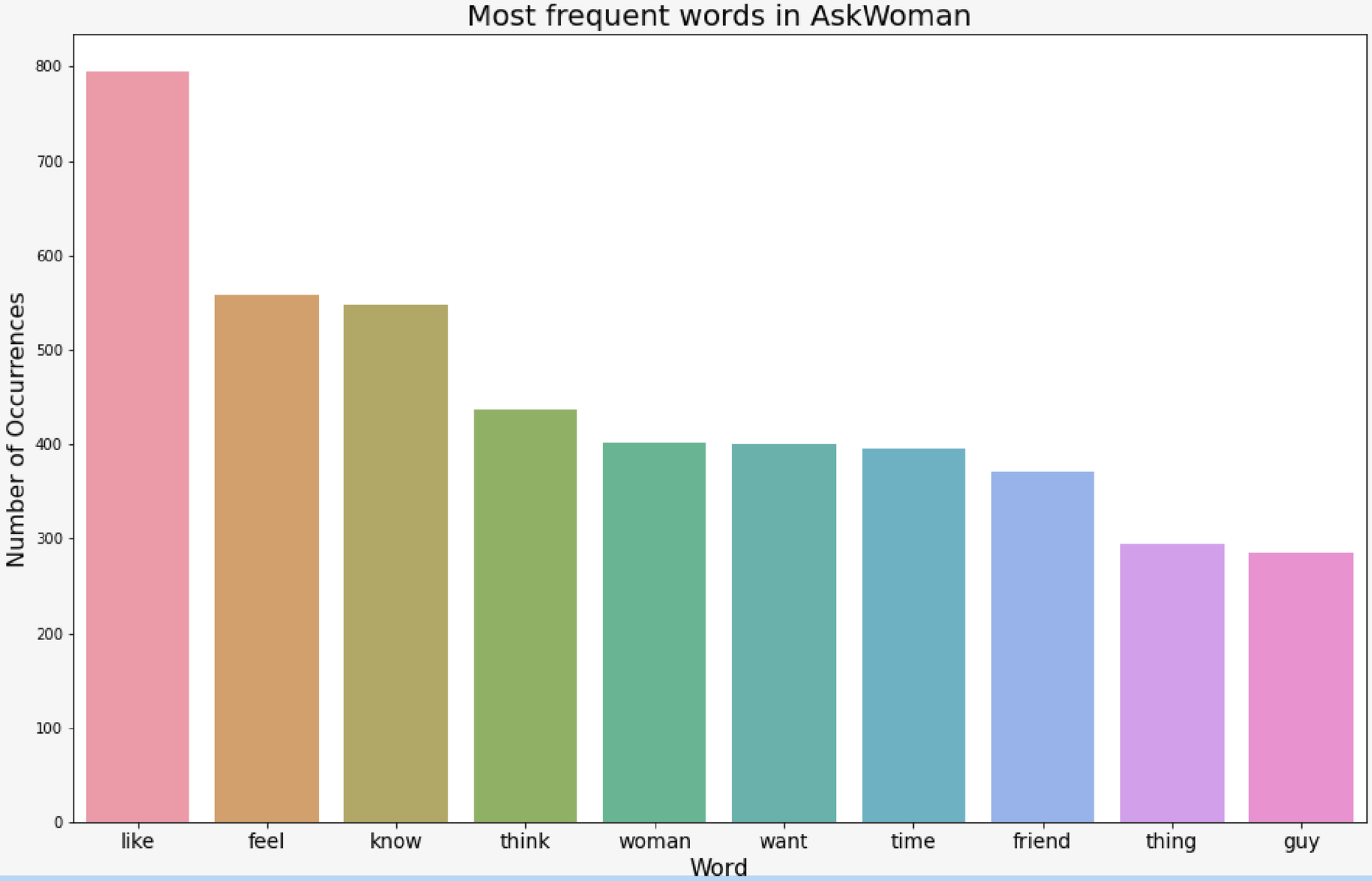
Reasons:

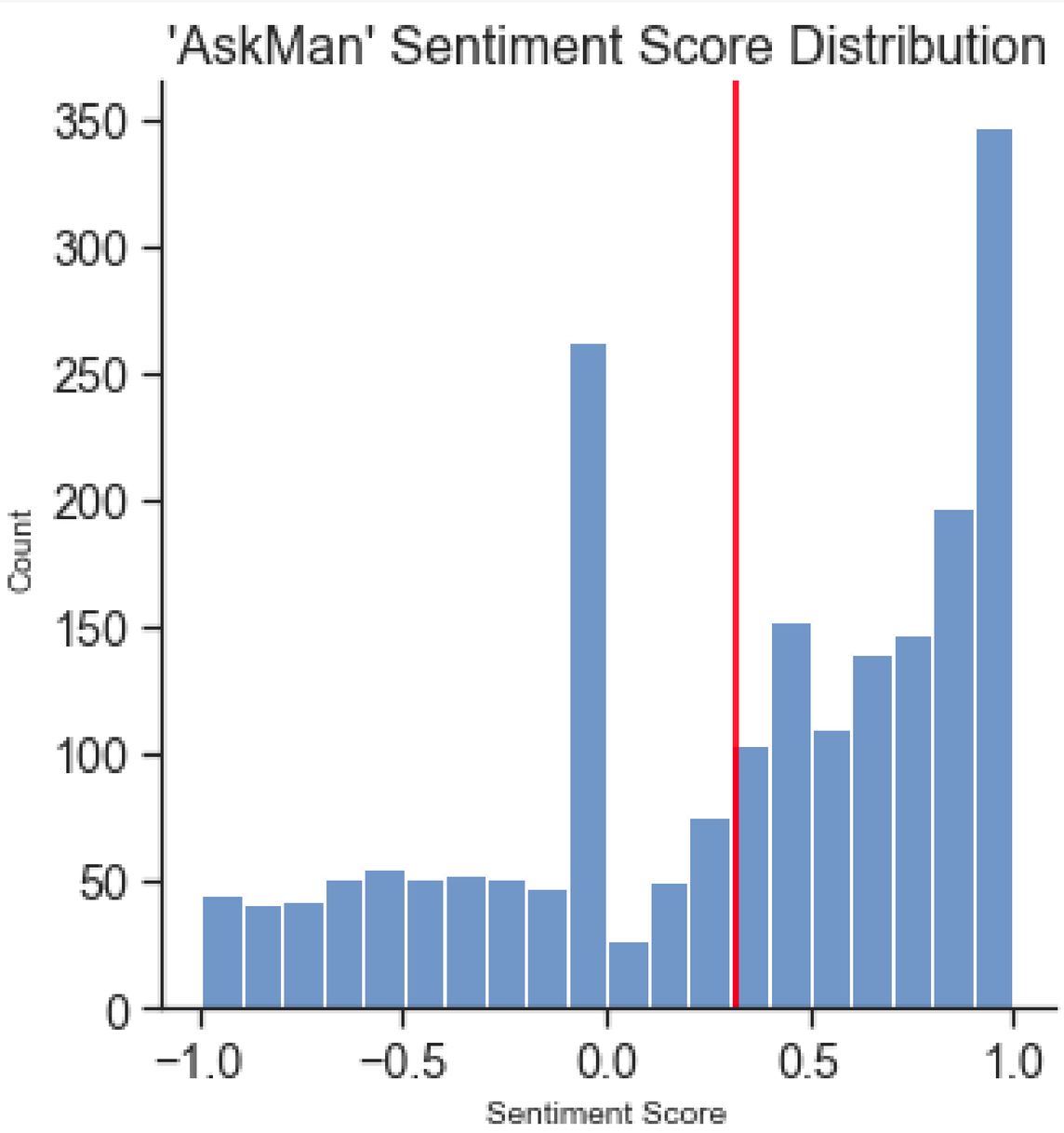
A lot of text data

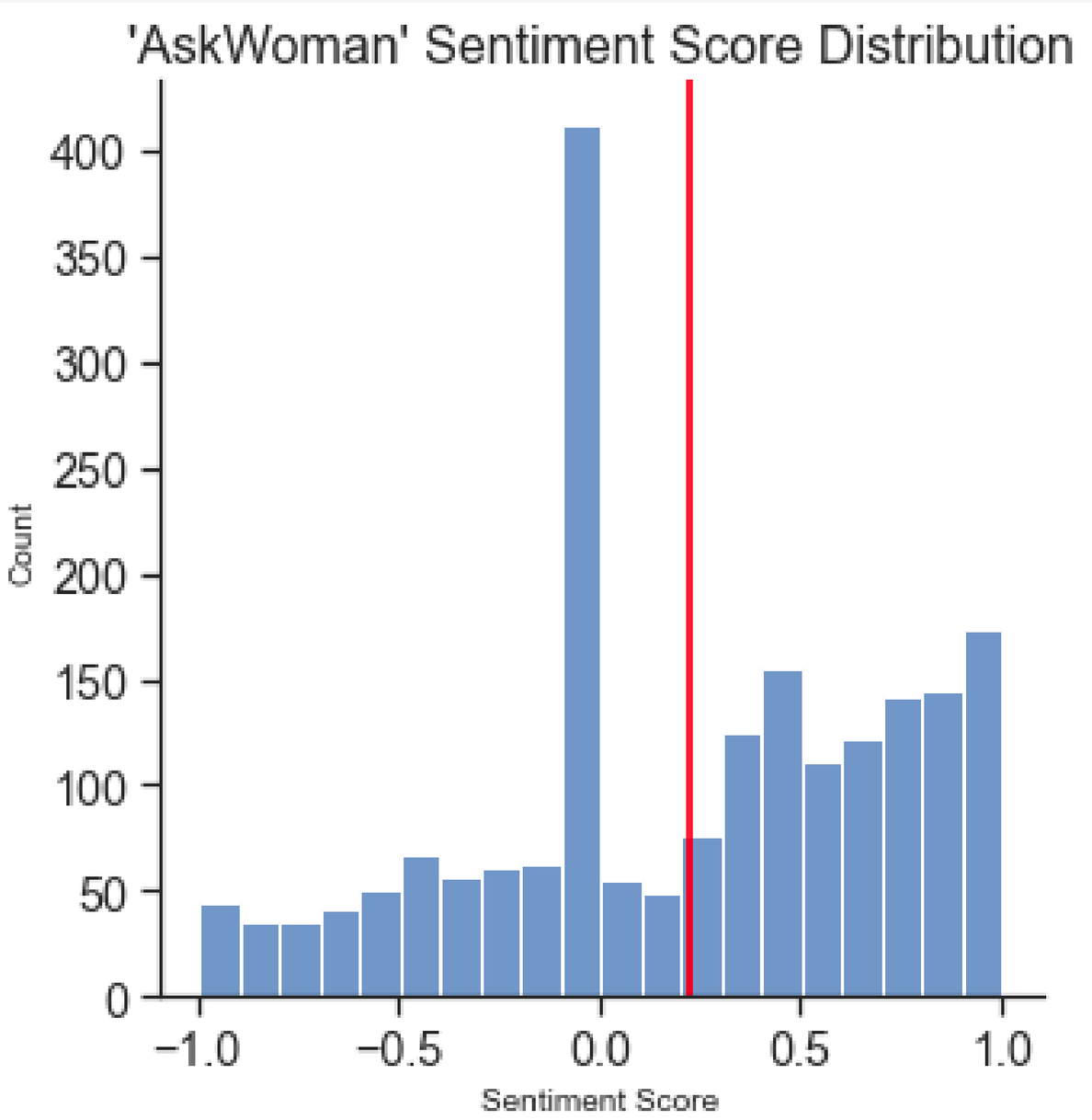
Significant amount of submission and commentaries

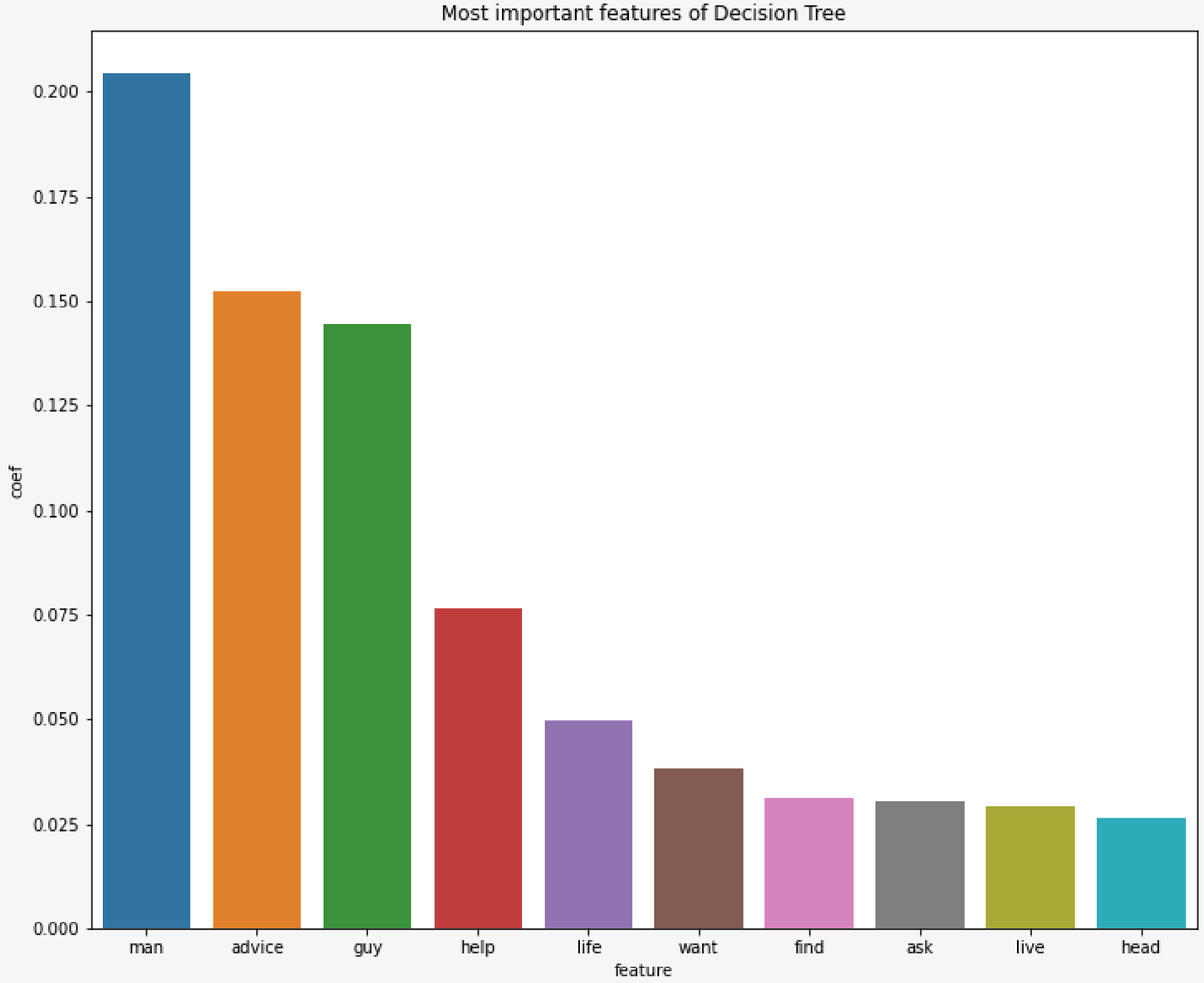
Interesting problem

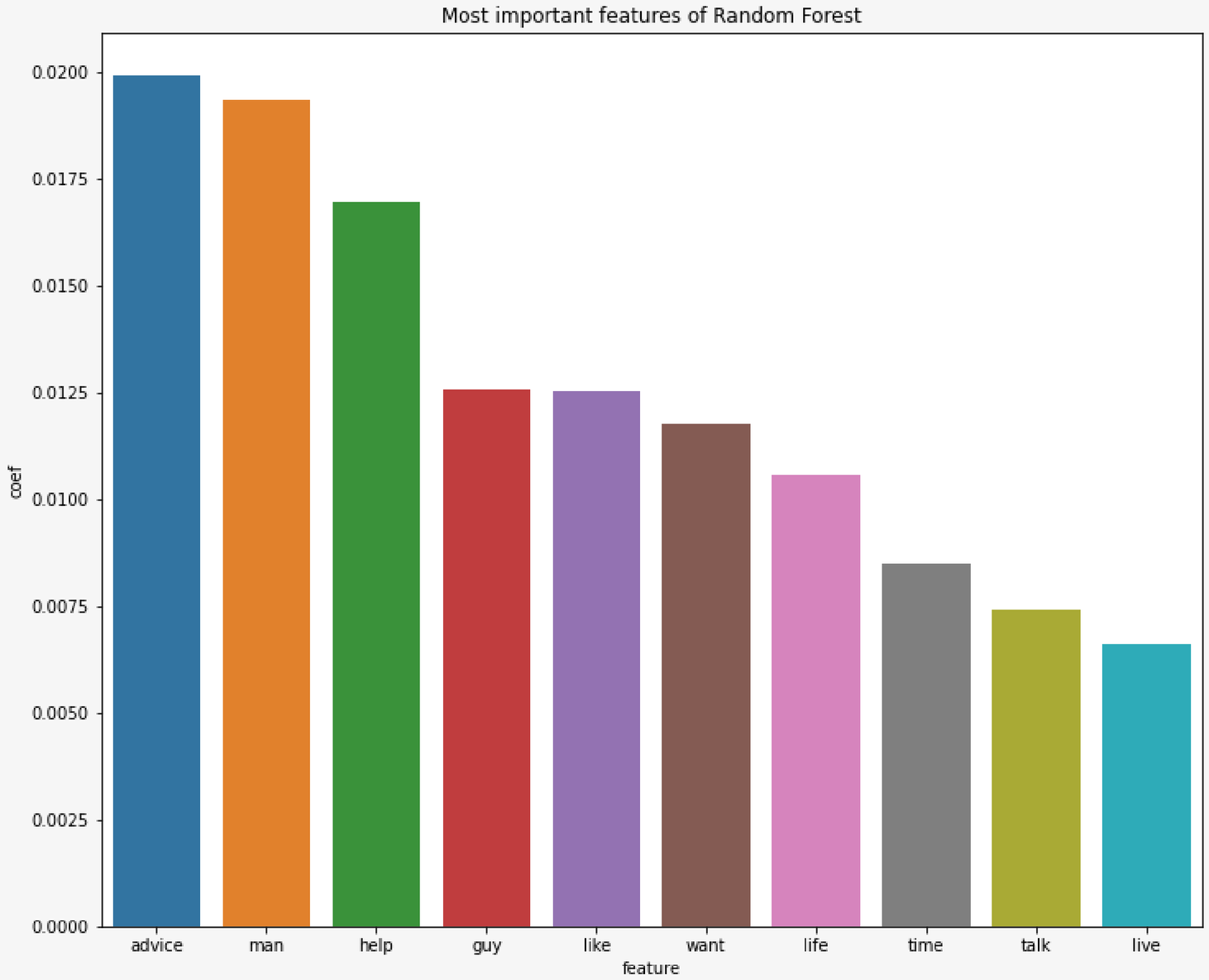

Most frequent words in AskMan
Number of Occurrences
1200
1000
800
600
400
200
0
like
feel
know
want
guy
time
think
friend
man
thing
Word

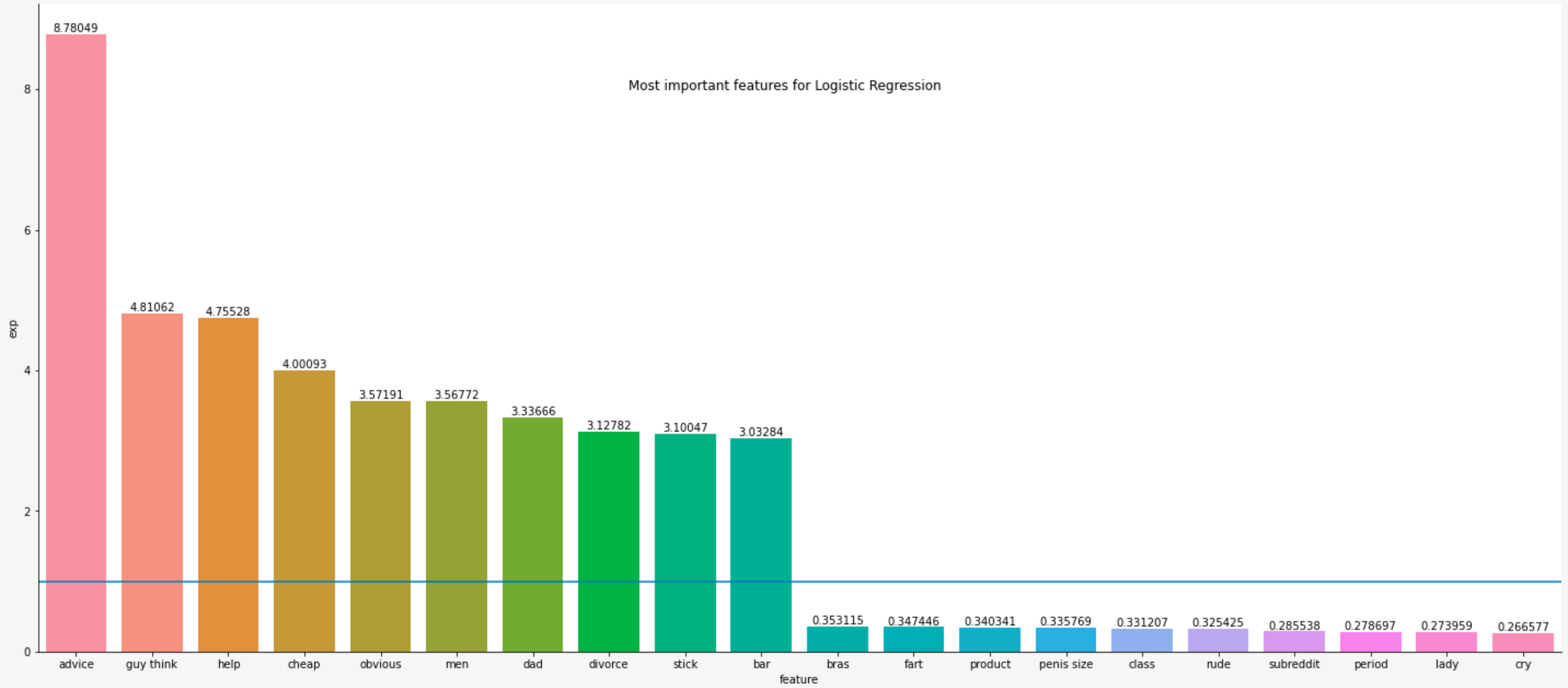

Most frequent words in AskWoman
Number of Occurrences
0
100
200
300
400
500
600
700
800
like
feel
know
think
woman
want
time
friend
thing
guy
Word


'AskMan' Sentiment Score Distribution
Count
Sentiment Score
0
50
100
150
200
250
300
350
−1.0
−0.5
0.0
0.5
1.0


'AskWoman' Sentiment Score Distribution
Count
Sentiment Score
0
50
100
150
200
250
300
350
400
−1.0
−0.5
0.0
0.5
1.0


Most important features of Decision Tree
coef
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
man
advice
guy
help
life
want
find
ask
live
head
feature


Most important features of Random Forest
coef
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
advice
man
help
guy
like
want
life
time
talk
live
feature


Most important features for Logistic Regression
exp
feature
0
2
4
6
8
8.78049
4.81062
4.75528
4.00093
3.57191
3.56772
3.33666
3.12782
3.10047
3.03284
0.353115
0.347446
0.340341
0.335769
0.331207
0.325425
0.285538
0.278697
0.273959
0.266577
advice
guy think
help
cheap
obvious
men
dad
divorce
stick
bar
bras
fart
product
penis size
class
rude
subreddit
period
lady
cry

AskWoman

Mean amount of words **21**

The longest sub **529** words

Mean amount of letters **122**

Max amount of letters **2702**

AskMan

Mean amount of words **30**

The longest sub **369** words

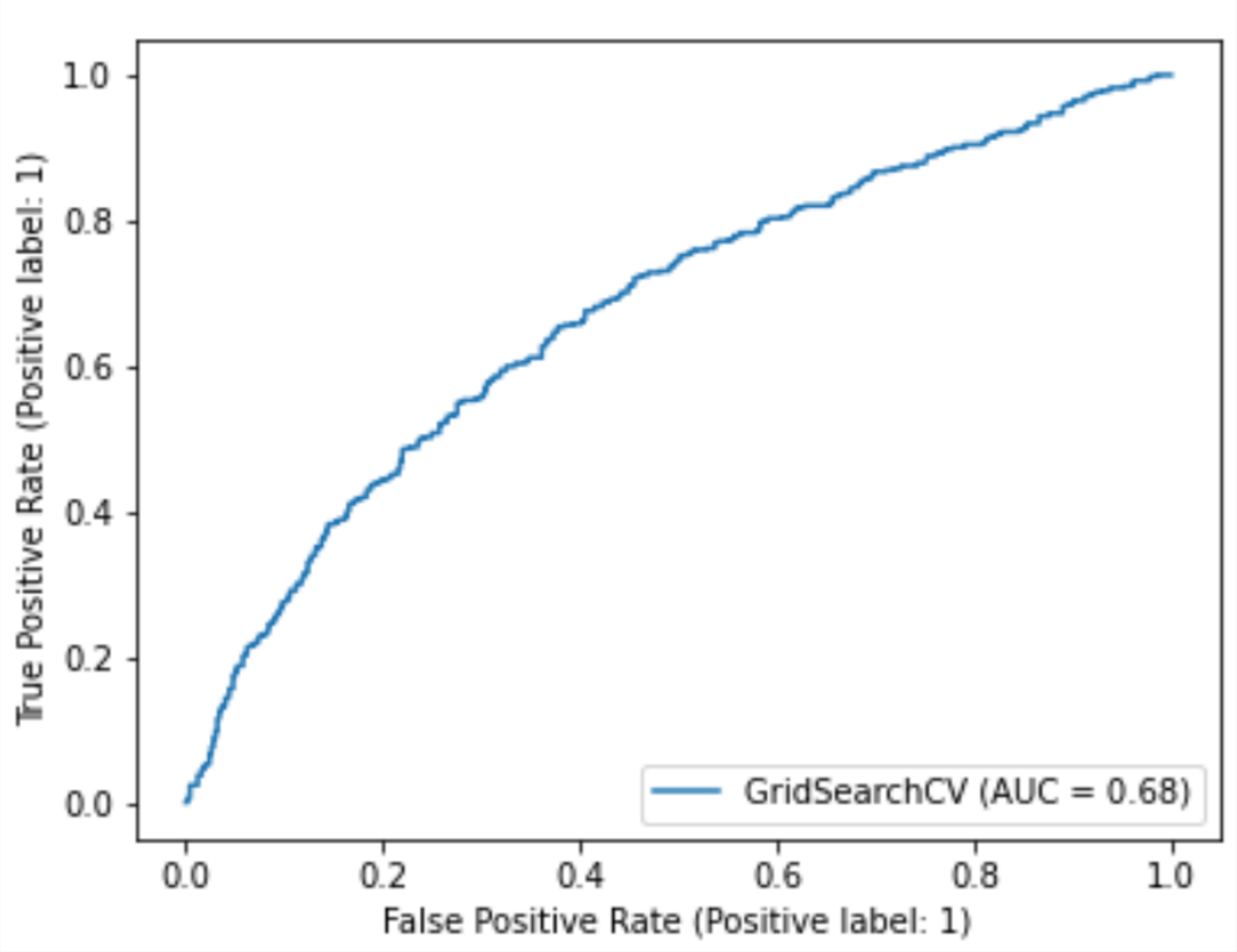
Mean amount of letters **167**

Max amount of letters **2062**

Top 3 models

Model	Accuracy	F1 score
BASELINE	0.5	0.504
AdaBoost	**0.63**	**0.59**
RandomForest	0.62	0.59
Naive Bayes	0.59	0.59

Logistic Regression ROC plot



Model	Submissions	Commentaries
AdaBoost	0.63	0.605
DecisionTree	0.62	0.579
Naive Bayes	0.59	0.635

Recommendations

Model shows noticeable better performance over baseline model

Let's give it a try!

2 ways to improve results

Dive deeper into the dataset

Find new data which would show more differences between man and woman


Thanks!
Thanks!
Thanks!
Thanks!
Thanks!
Thanks!
Thanks!
Thanks!
Thanks!
Thanks!
Thanks!