Reddit Subreddit analysis

By Derik Vo

Background

1. Our team is studying social interaction by gender
2. We will incorporate this by using Reddit
3. We want to build a classification model

AskWomen vs. AskMen

Agenda

1. Data Collection
2. Exploratory Data analysis
3. Modeling
4. Conclusions
5. Recommendations
6. Q & A

Can we categorize a reddit post?

Data Collection

Record Count

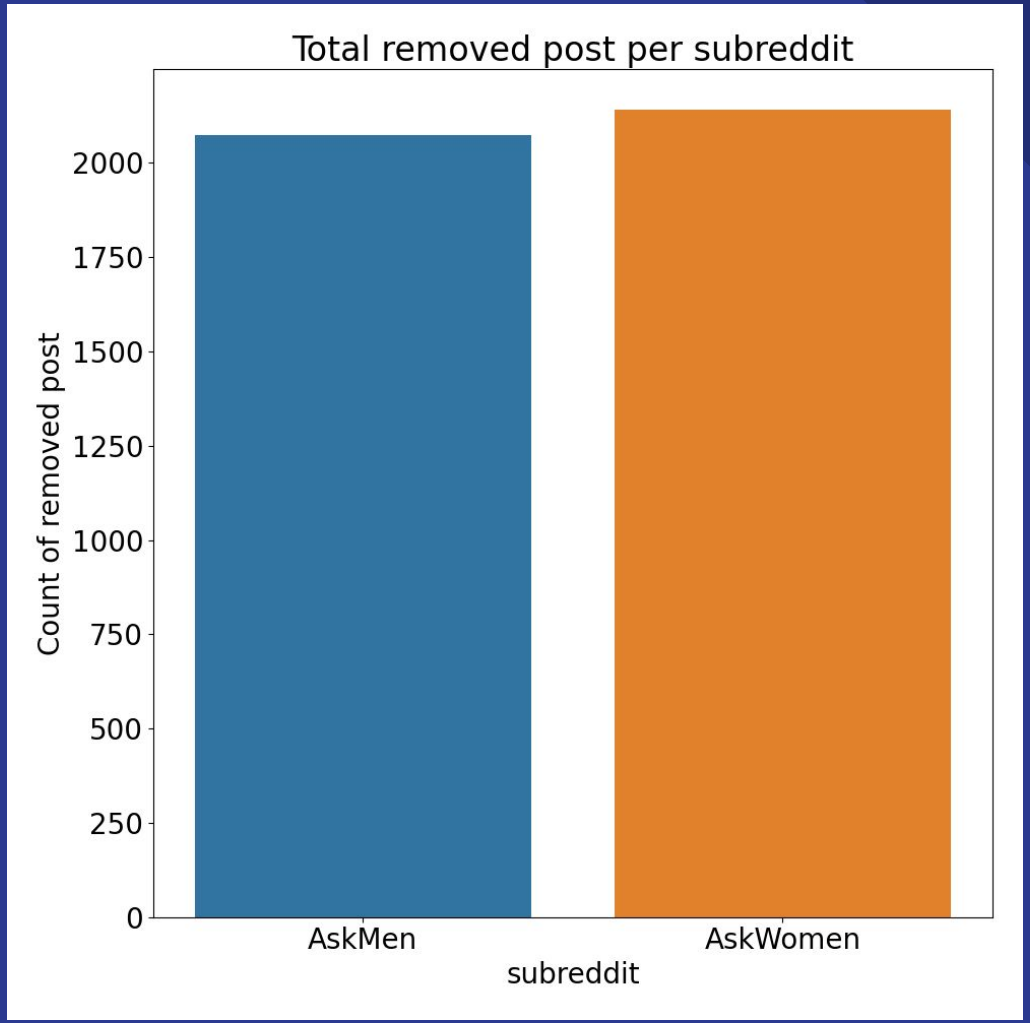
Ask Men: 3995 records

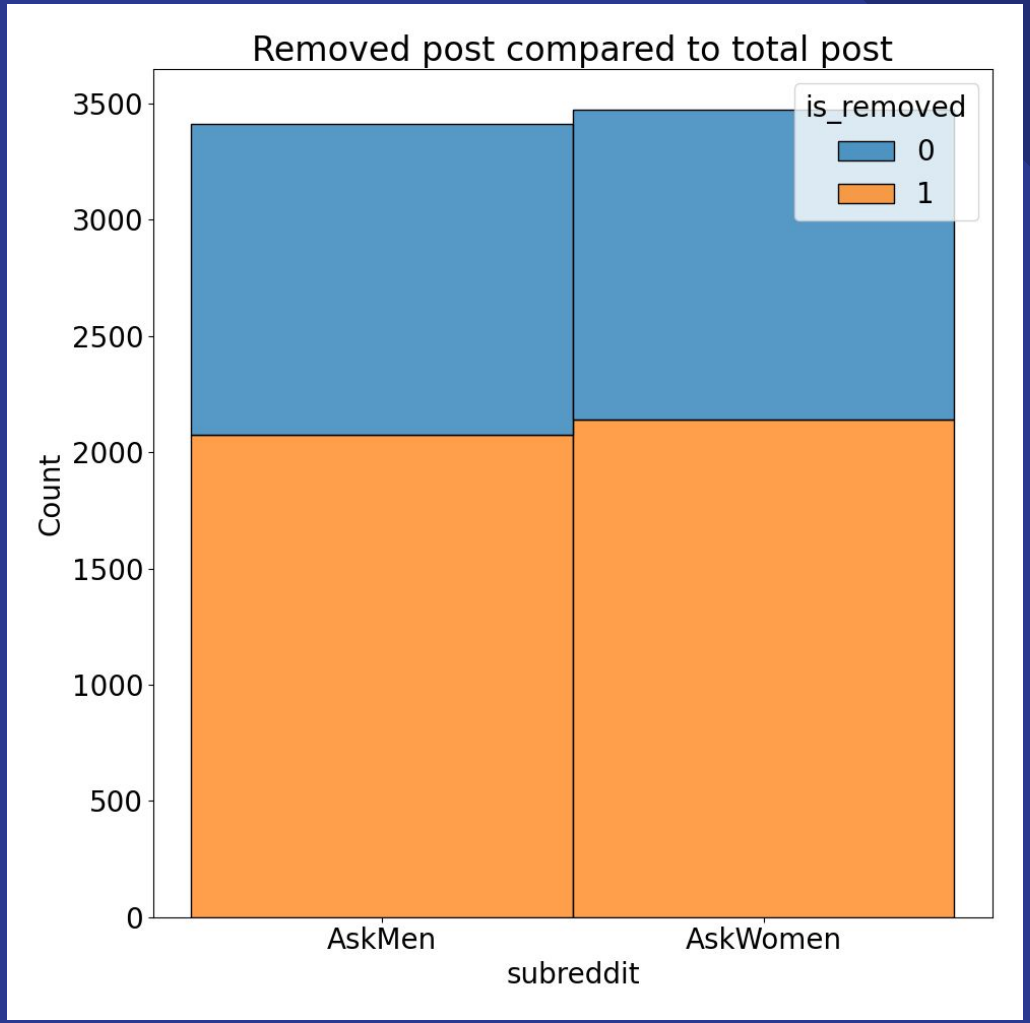
Ask Women: 3991 records

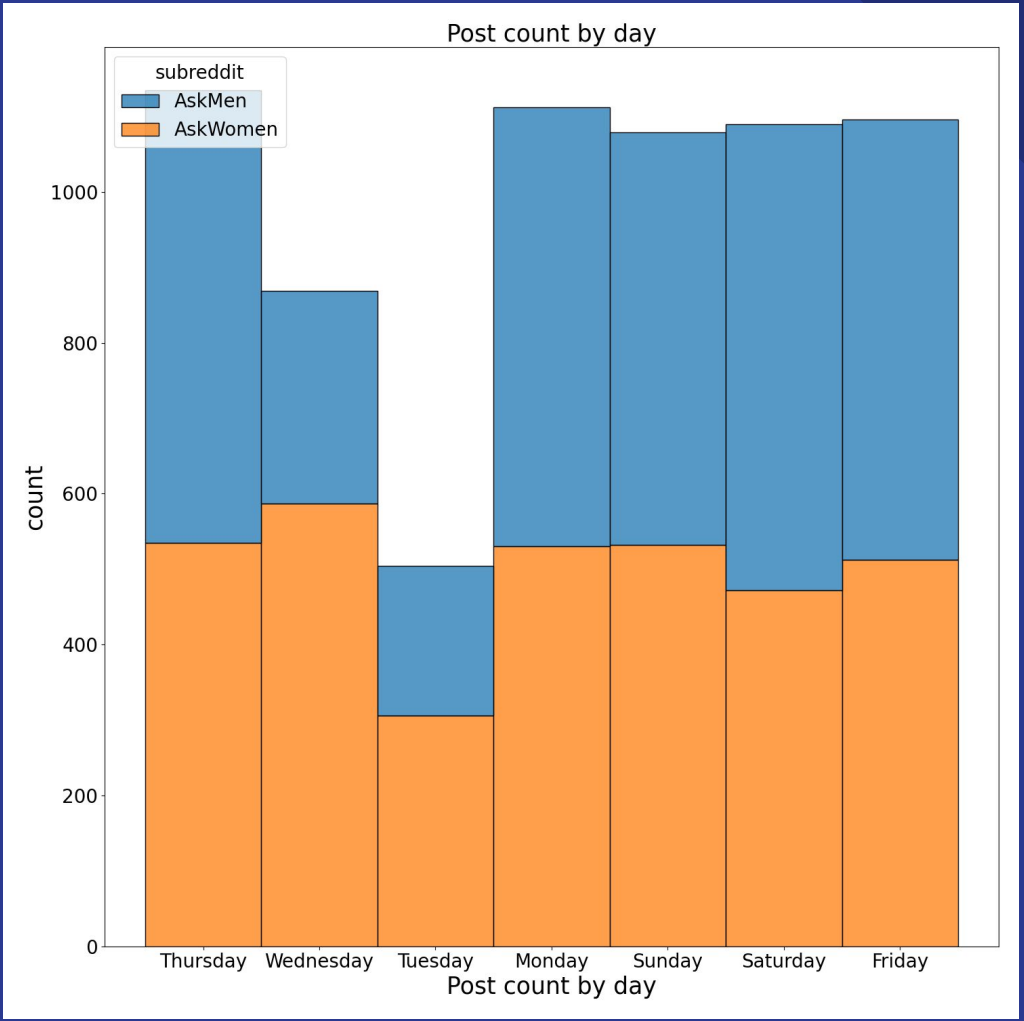
Filters used

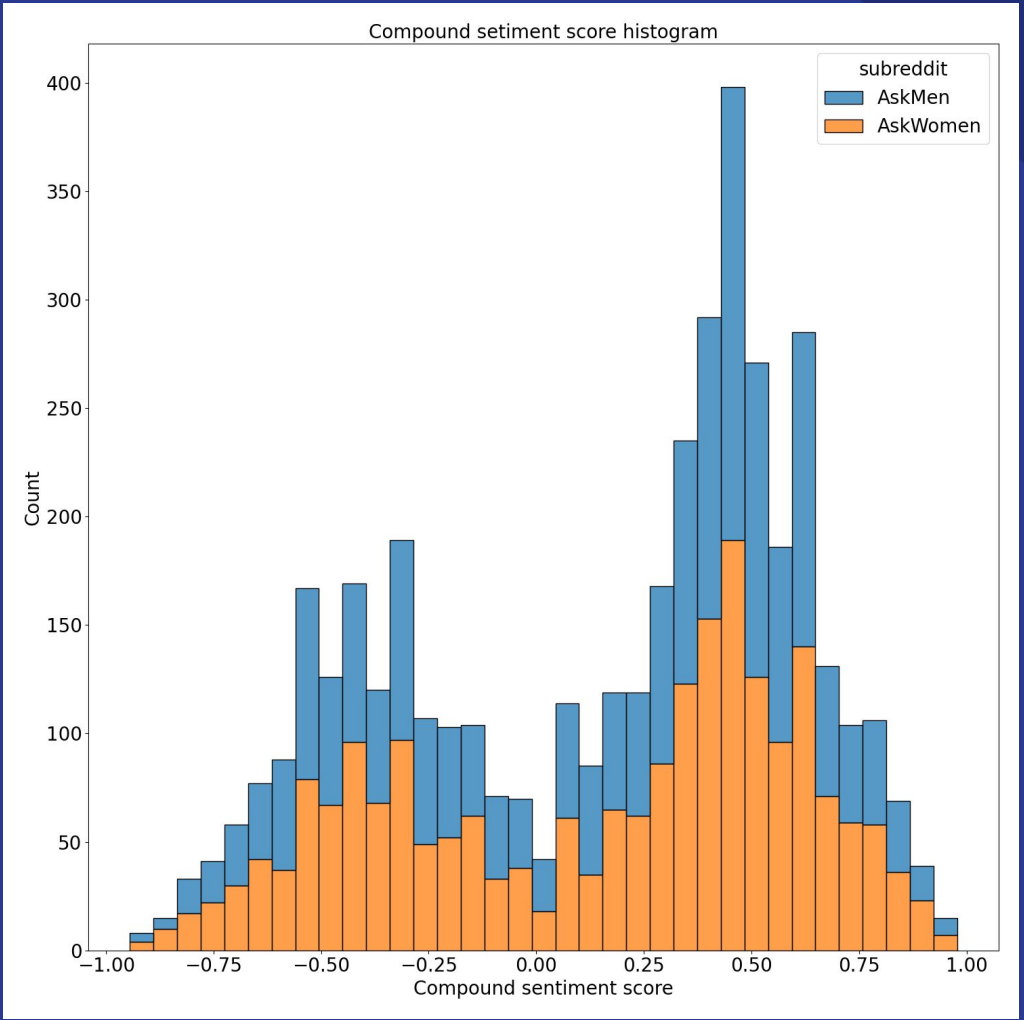
1. Since
2. min_comments

Exploratory Data analysis

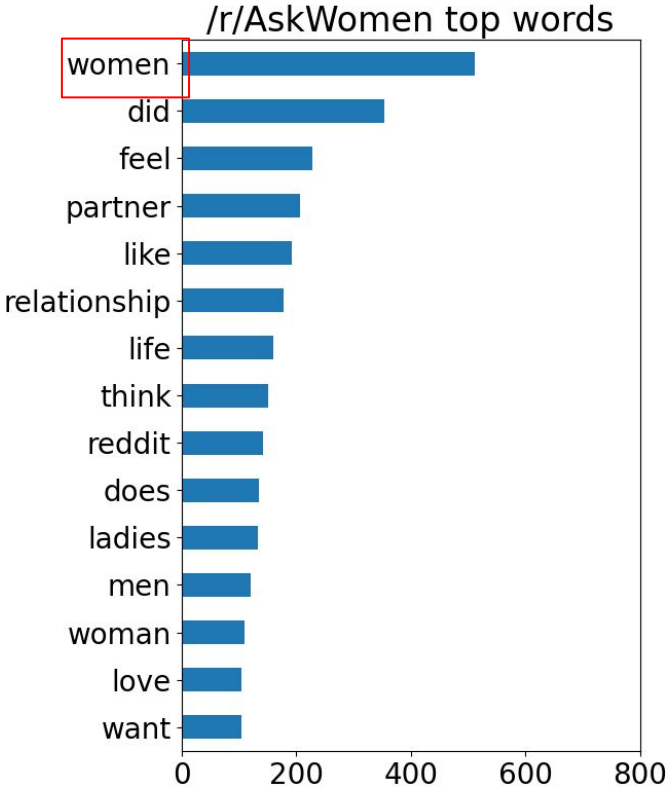

Total removed post per subreddit
Count of removed post
2000
1750
1500
1250
1000
750
500
250
0
AskMen
AskWomen
subreddit

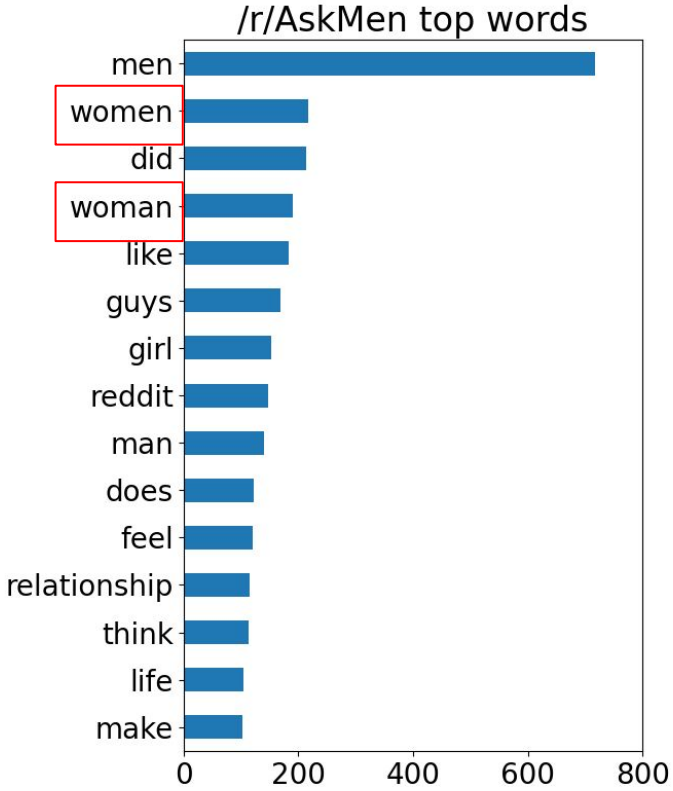

Removed post compared to total post
is_removed
0
1
Count
3500
3000
2500
2000
1500
1000
500
0
AskMen
AskWomen
subreddit


Post count by day
subreddit
AskMen
AskWomen
count
0
200
400
600
800
1000
Thursday
Wednesday
Tuesday
Monday
Sunday
Saturday
Friday
Post count by day

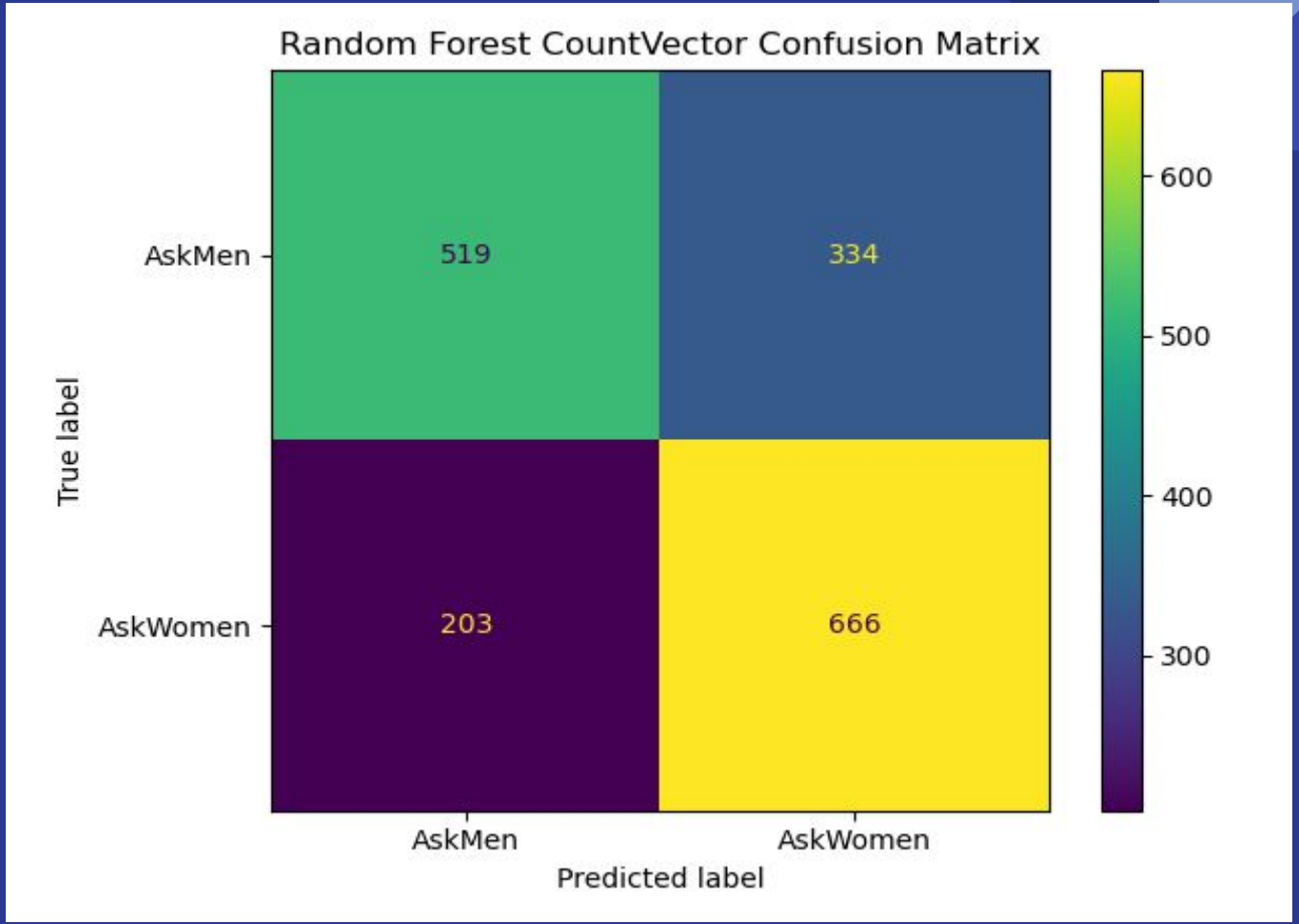

Compound setiment score histogram
subreddit
AskMen
AskWomen
Count
0
50
100
150
200
250
300
350
400
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Compound sentiment score

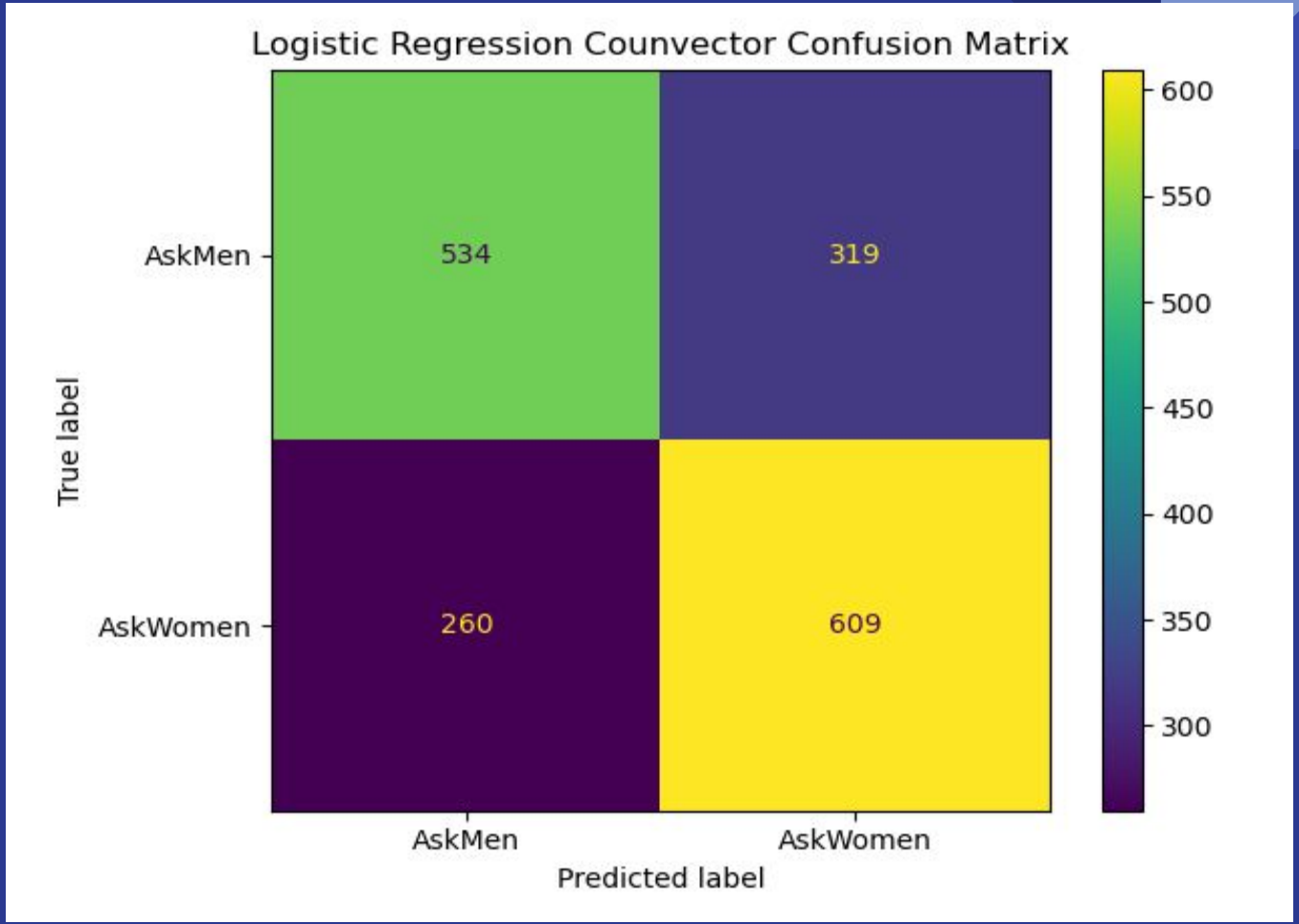
Top 15 words in each subreddit

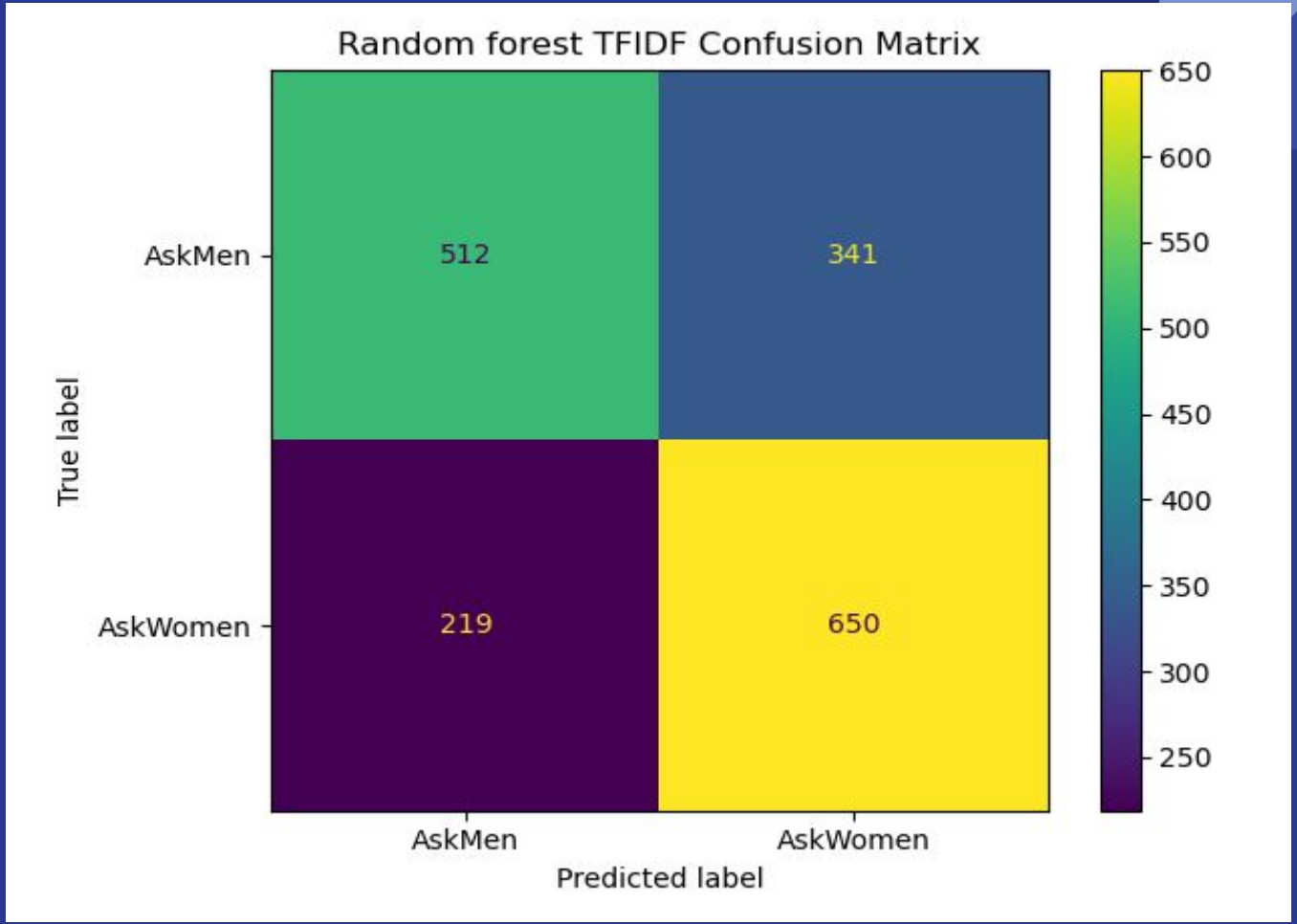


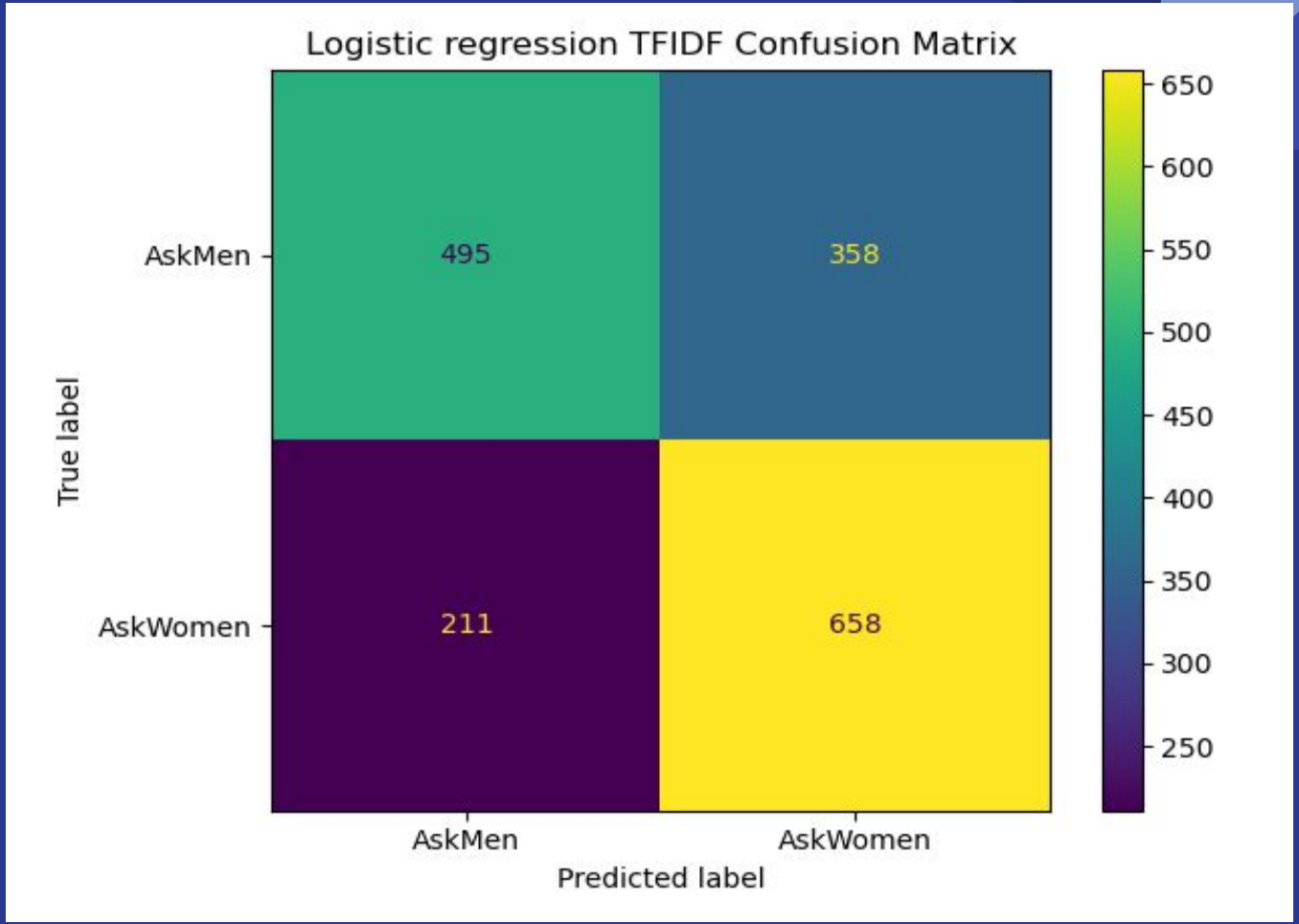


Modeling


Random Forest CountVector Confusion Matrix
AskMen
AskWomen
True label
519
334
203
666
AskMen
AskWomen
Predicted label
600
500
400
300


Logistic Regression Counvector Confusion Matrix
AskMen
AskWomen
534
319
260
609
True label
AskMen
AskWomen
Predicted label
600
550
500
450
400
350
300


Random forest TFIDF Confusion Matrix
AskMen
AskWomen
True label
512
341
219
650
AskMen
AskWomen
Predicted label
650
600
550
500
450
400
350
300
250


Logistic regression TFIDF Confusion Matrix
AskMen
AskWomen
True label
495
358
211
658
AskMen
AskWomen
Predicted label
650
600
550
500
450
400
350
300
250

Conclusion

Models could not categorize

1. Models were relatively consistent in the predictions
2. Models and transformer usage had marginal effect
3. The behavior of each subreddit had too much similarities

Recommendations

1. Filter out common words shared between subreddits
2. Incorporate the comments within the subreddits
3. Share our dataset to build a bot to identify post that should be removed