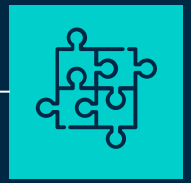
NLP Reddit Analysis

-Blago Ugrinov,
GA Data Scientist

Problem Statement

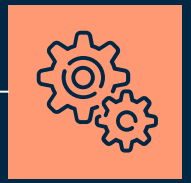
Based on Reddit discourse, to what extent, and in which ways, are the fields of general engineering and machine learning similar?

TABLE OF CONTENTS



01

Data Processing



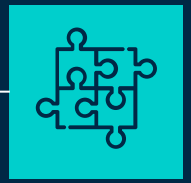
02

Model &
Visualizations

03

Recommendations
& Looking Forward

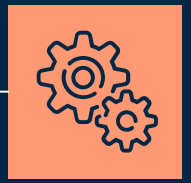
Data Gathering/Cleaning/Engineering



Gathering

Reddit PushShift API

-2600 pulled total
- >15 comments
- >15 Reddit score



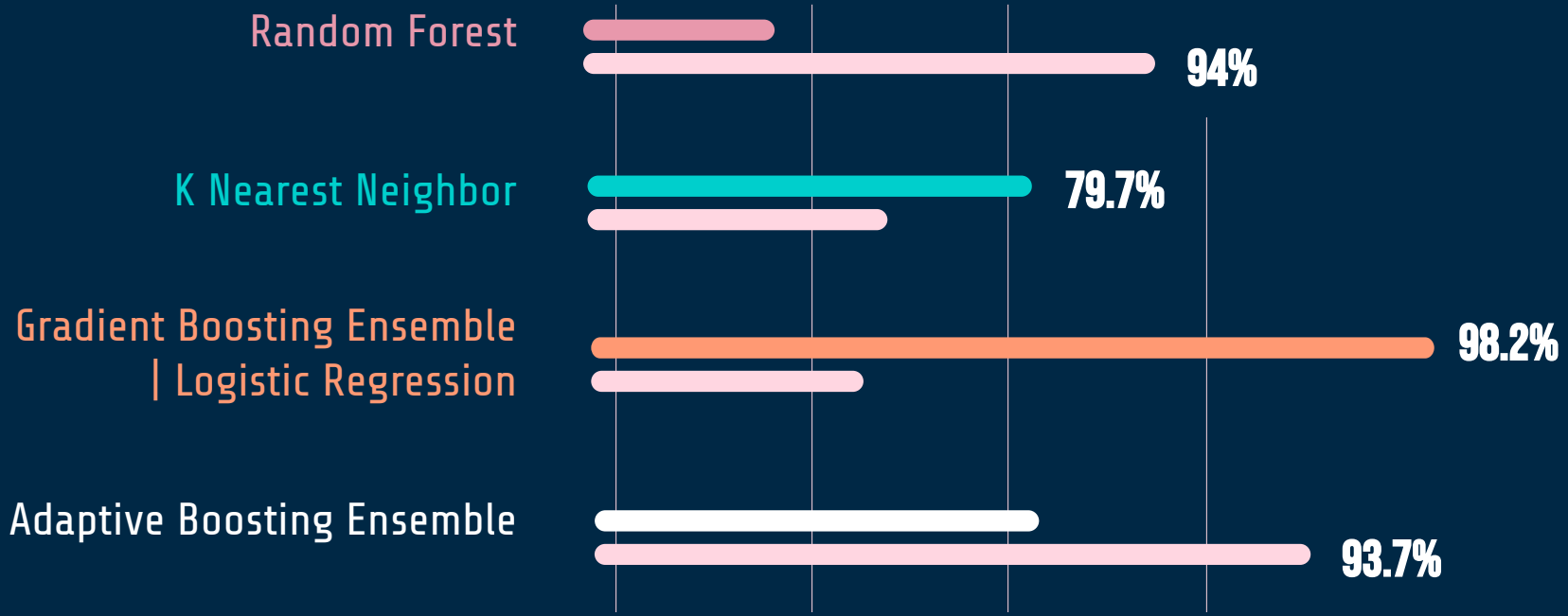
Cleaning

- Binarize subreddit target
- Binarize selftext

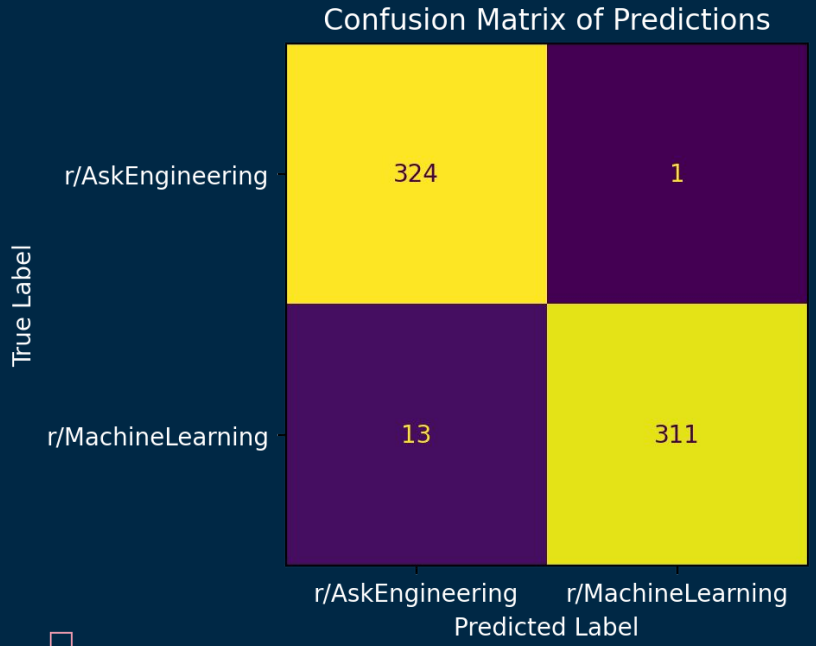
Feature Engineering

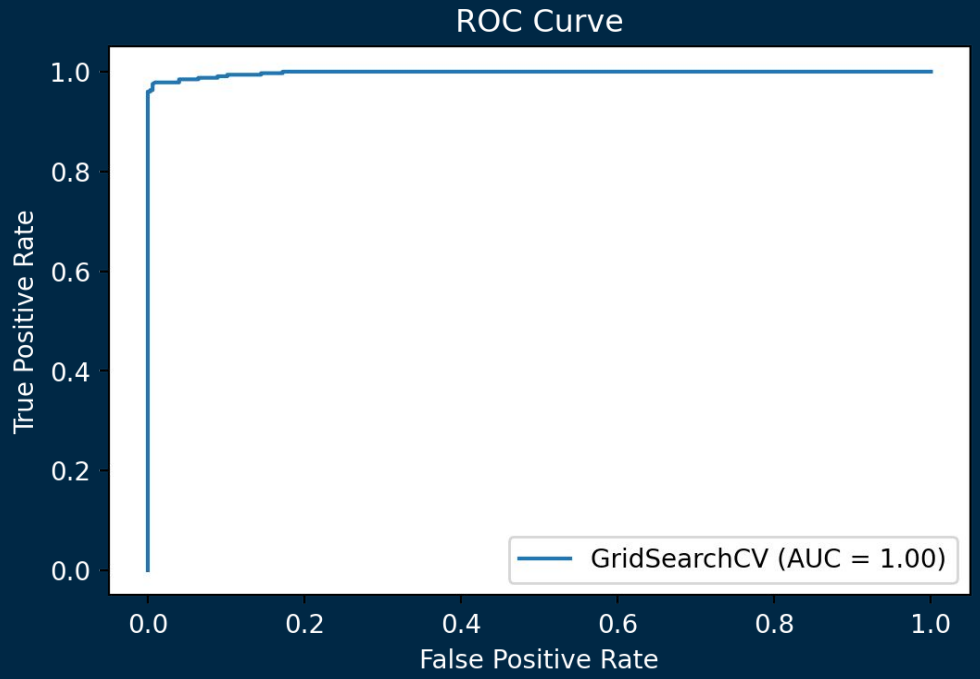
- Total text feature
- Title and selftext length feature
- Title and selftext word count feature

Modeling Pipelines

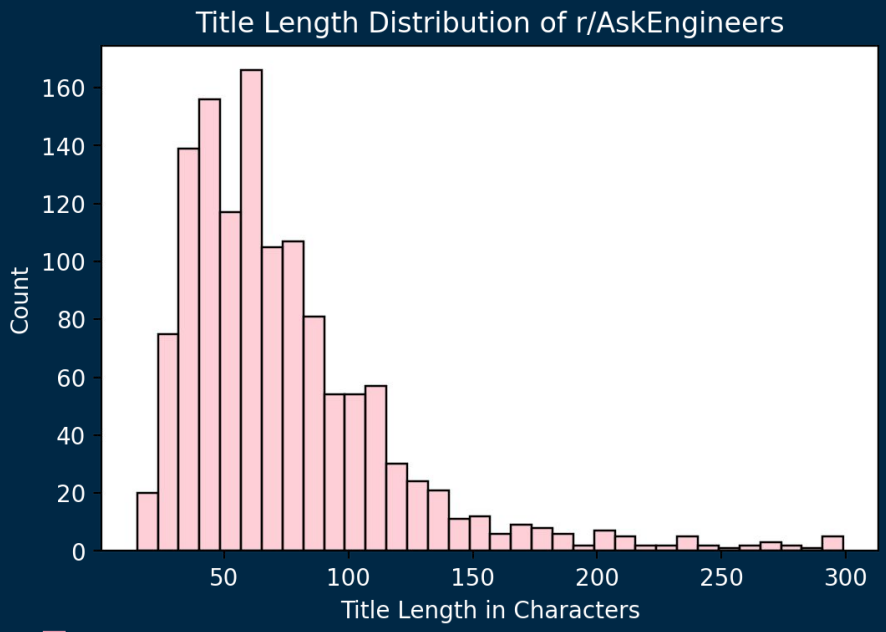


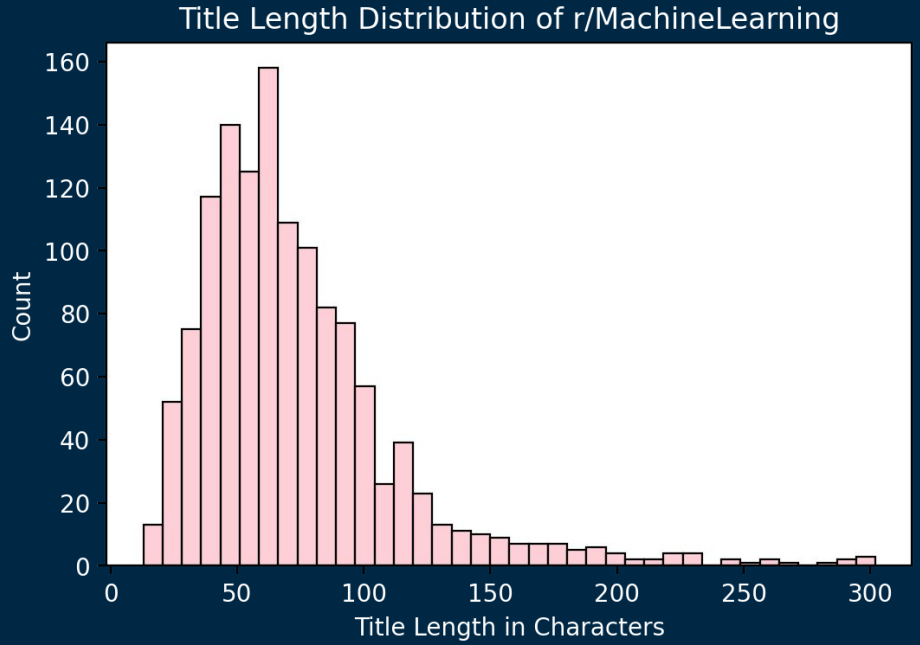
Visualizations



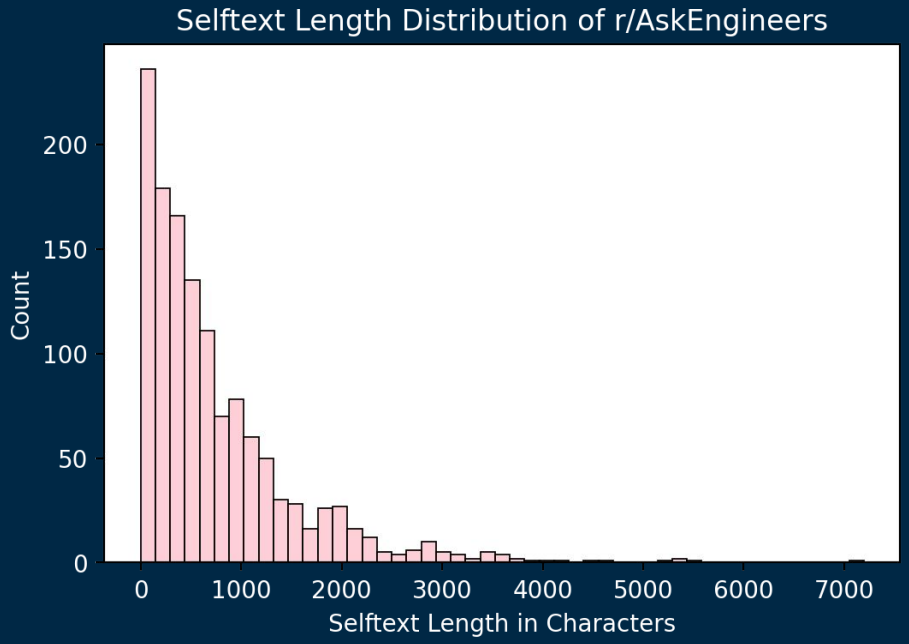


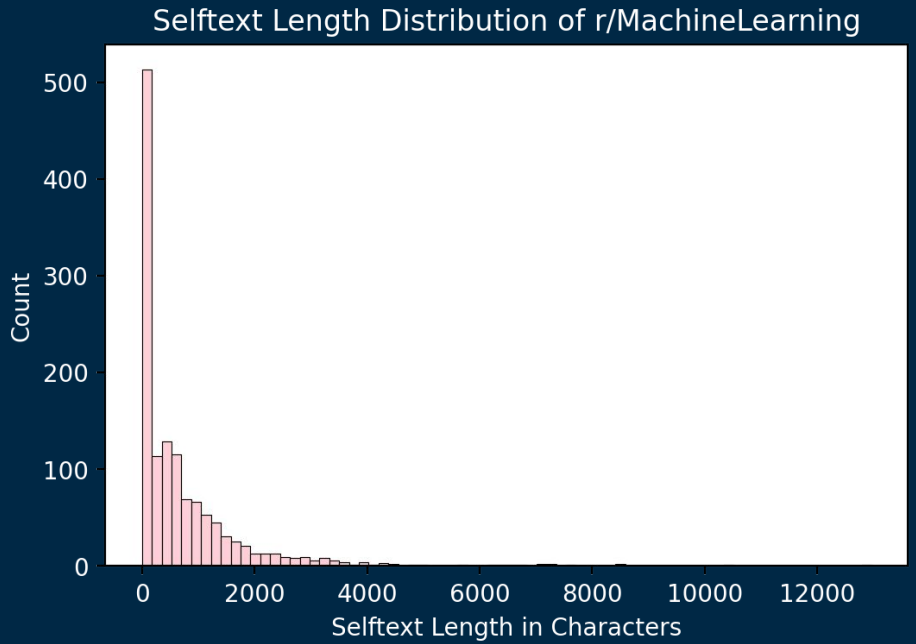
Visualizations



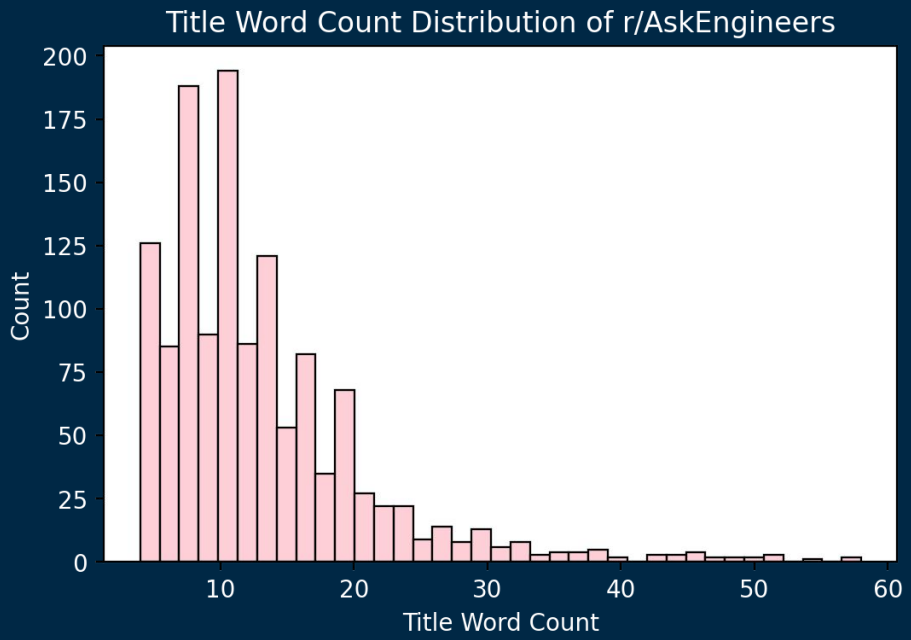


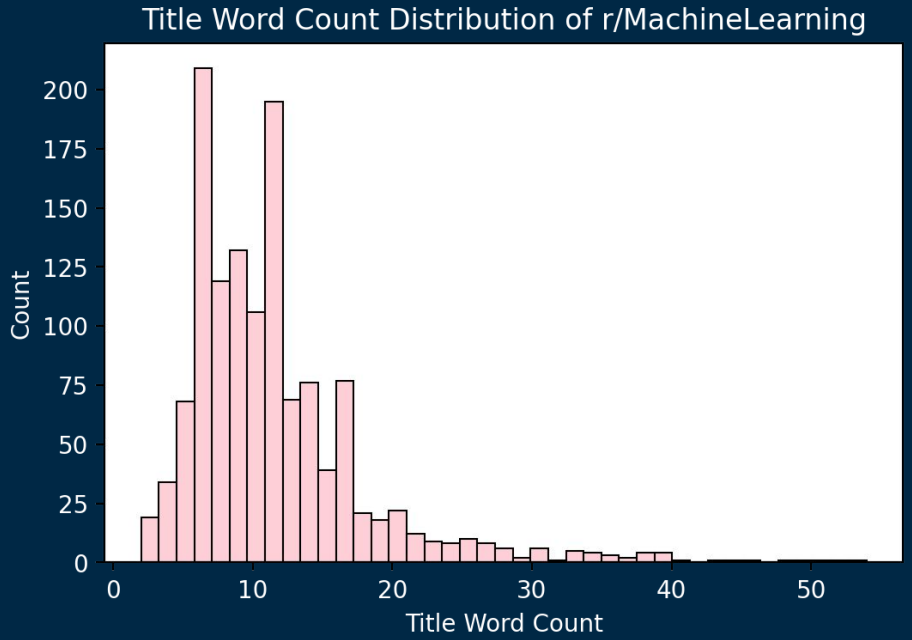
Visualizations



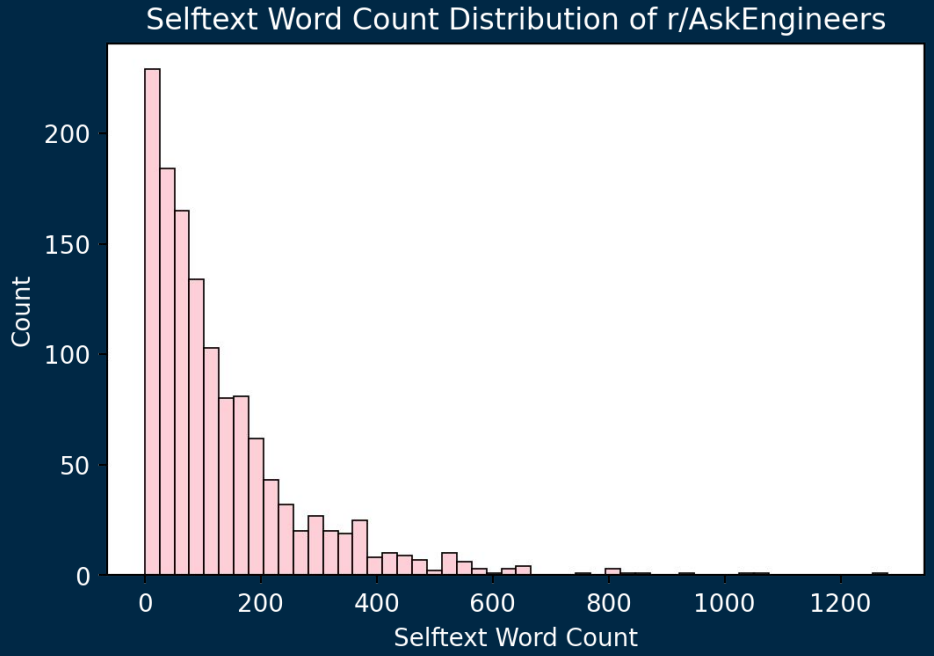


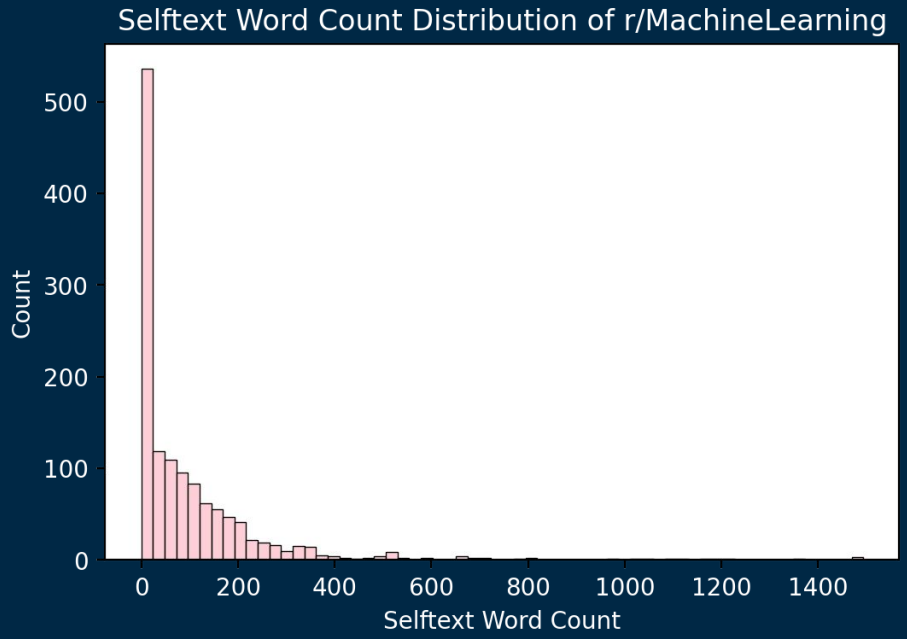
Visualizations



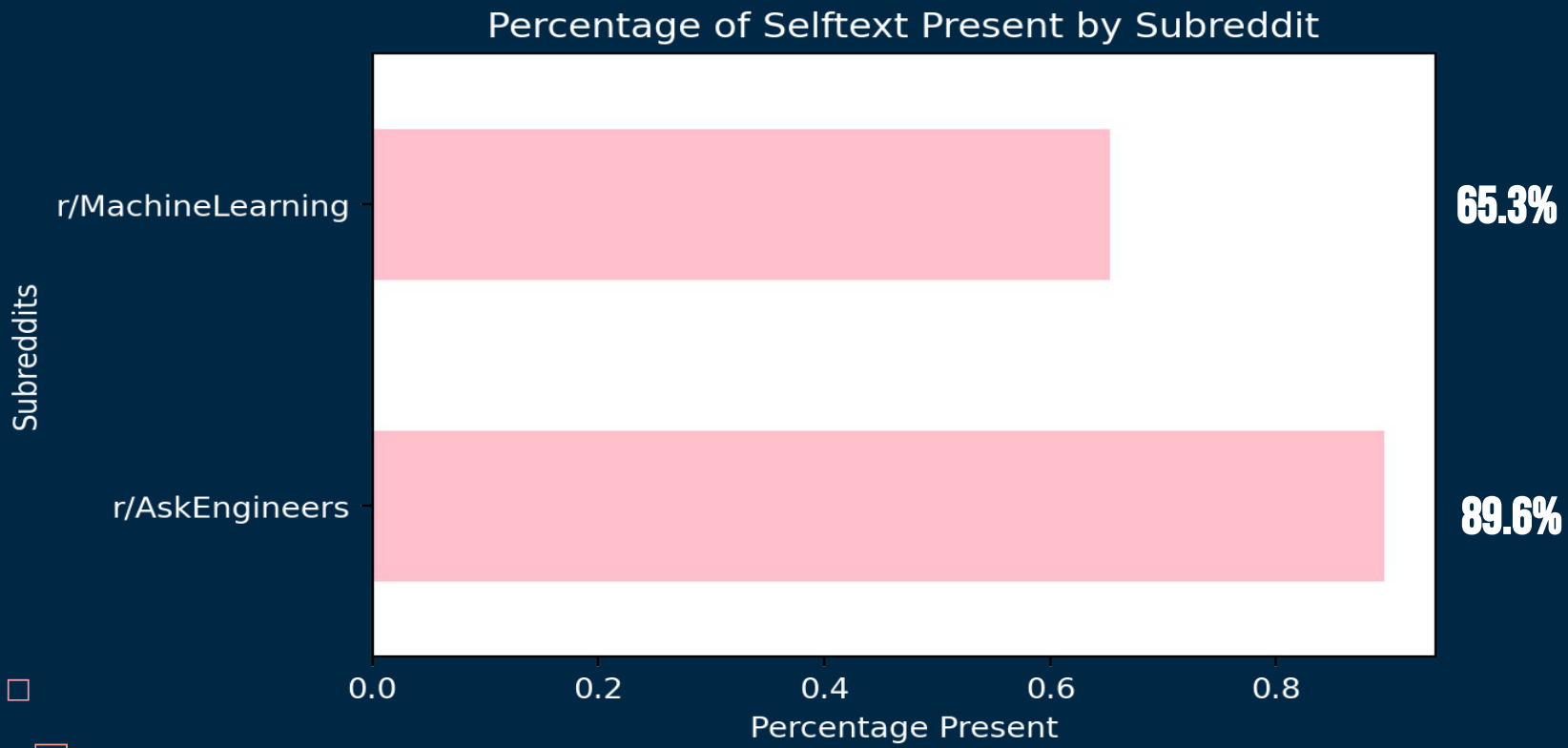


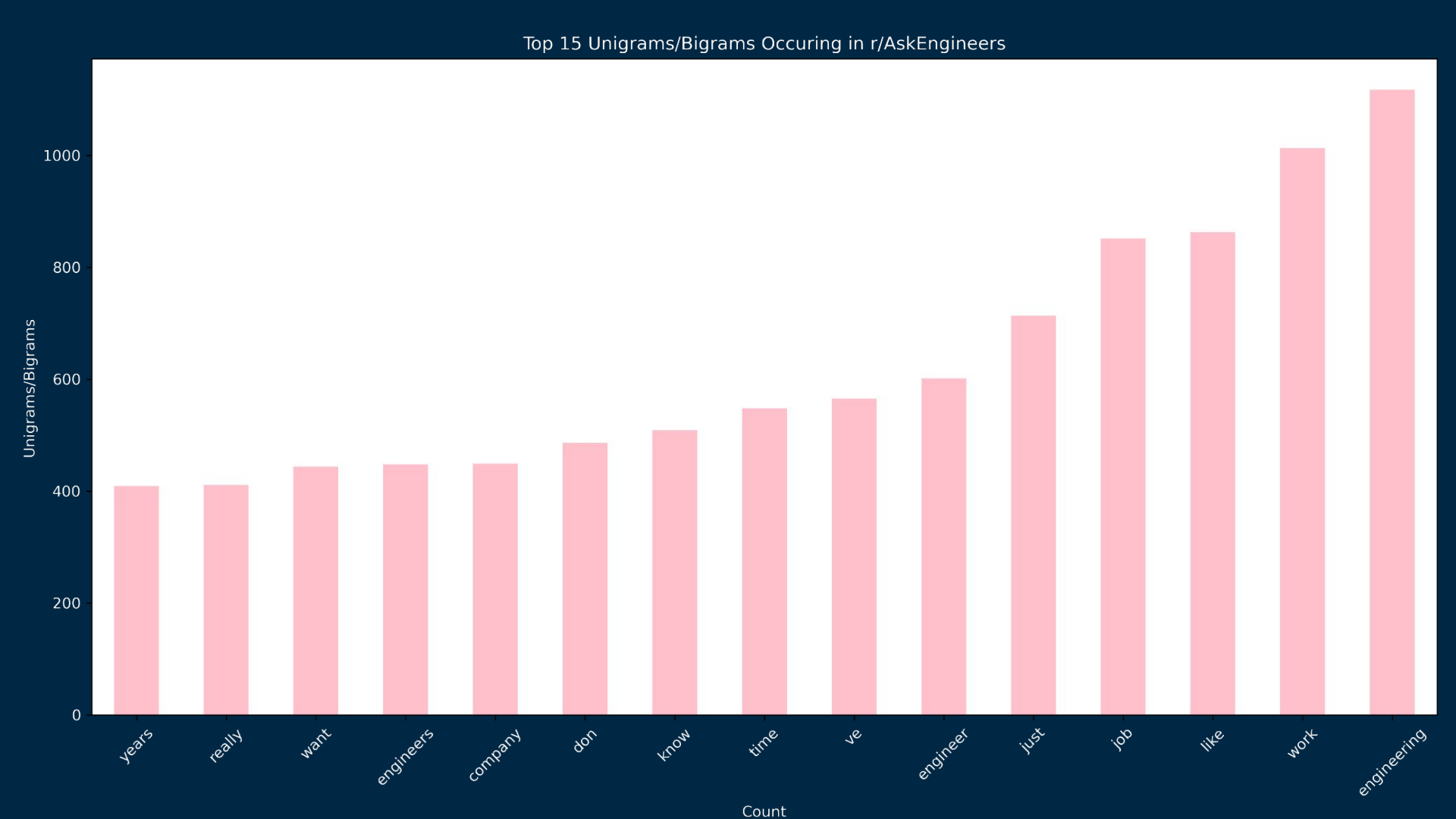
Visualizations

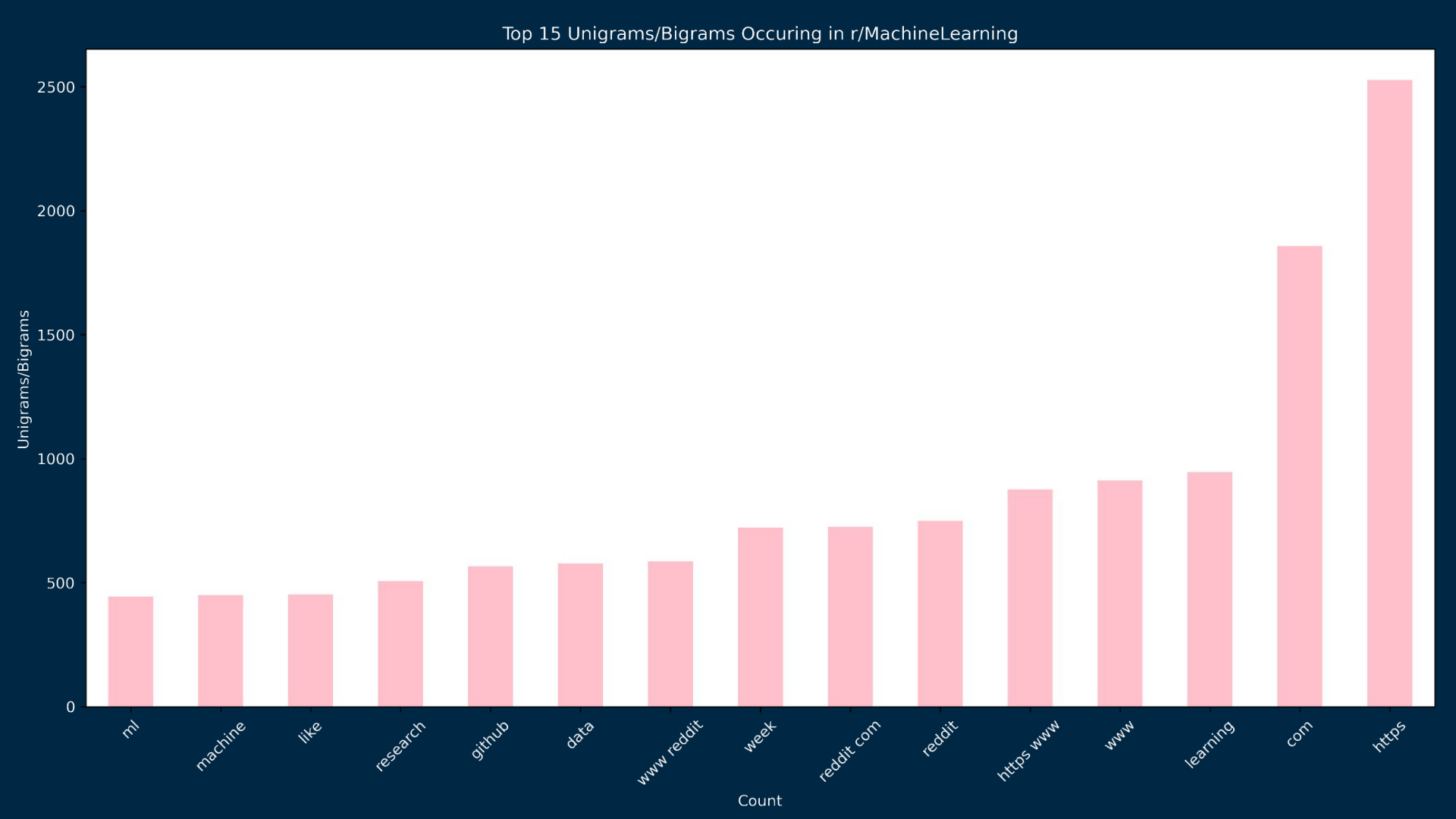




Visualizations




Top 15 Unigrams/Bigrams Occuring in r/AskEngineers
Unigrams/Bigrams
0
200
400
600
800
1000
years
really
want
engineers
company
don
know
time
ve
engineer
just
job
like
work
engineering
Count


Top 15 Unigrams/Bigrams Occuring in r/MachineLearning
Unigrams/Bigrams
Count
0
500
1000
1500
2000
2500
ml
machine
like
research
github
data
www reddit
week
reddit com
reddit
https www
www
learning
com
https

Conclusion & Next Steps

- The fields of machine learning and general engineering **do not seem to have much overlap** at this point in time in regards to subreddit discourse.
- **Common words** in top 50: 'com', 'https', 'use', 'new', 'people', 'think', 'know', 'time', 've', 'just', **'like'**, 'work'.
- **Moving forward**, more data would allow us a more comprehensive and accurate model.

Thank You!