Reddit NLP

CHRISTOPHER DENQ

(ASKREDDIT VS. ASKSCIENCE)

1. Problem Statement
2. Methodology
3. Analysis
4. Conclusion

1. Problem Statement
2. Methodology
3. Analysis
4. Conclusion

Goal

Give targeted ads to users

Goal

Give targeted ads to users

Classify both users

Goal

Give targeted ads to users

Classify both users

Solution

Pre-trained classification models

1. Problem Statement
2. Methodology
3. Analysis
4. Conclusion

Solution: Pre-trained Models

Solution: Pre-trained Models

1. Data Gathering w/ Pushshift

Solution: Pre-trained Models

1. Data Gathering w/ Pushshift
2. Cleaning & EDA w/ Python

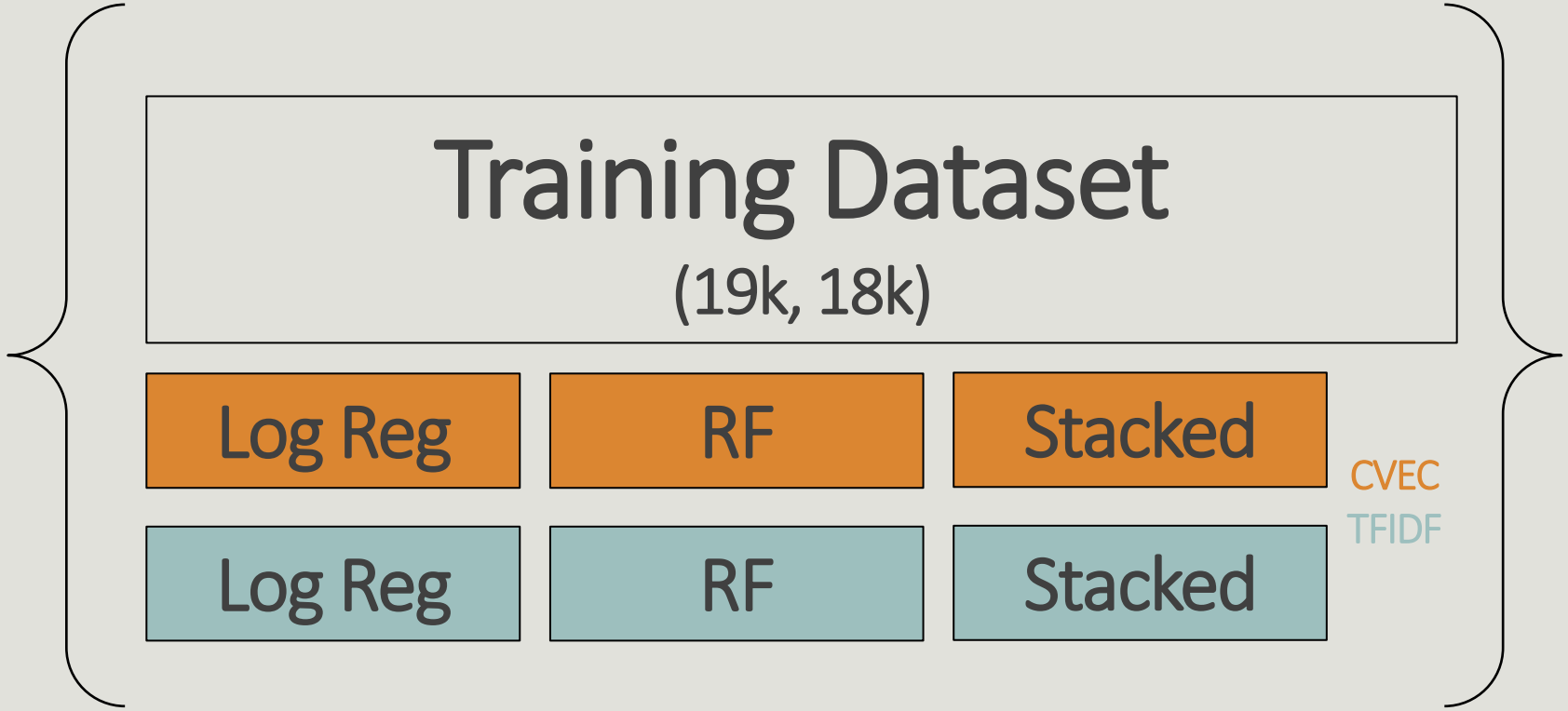
Solution: Pre-trained Models

1. Data Gathering w/ Pushshift
2. Cleaning & EDA w/ Python
3. **NLP Modeling w/ CVEC & TFIDF**

Solution: Pre-trained Models

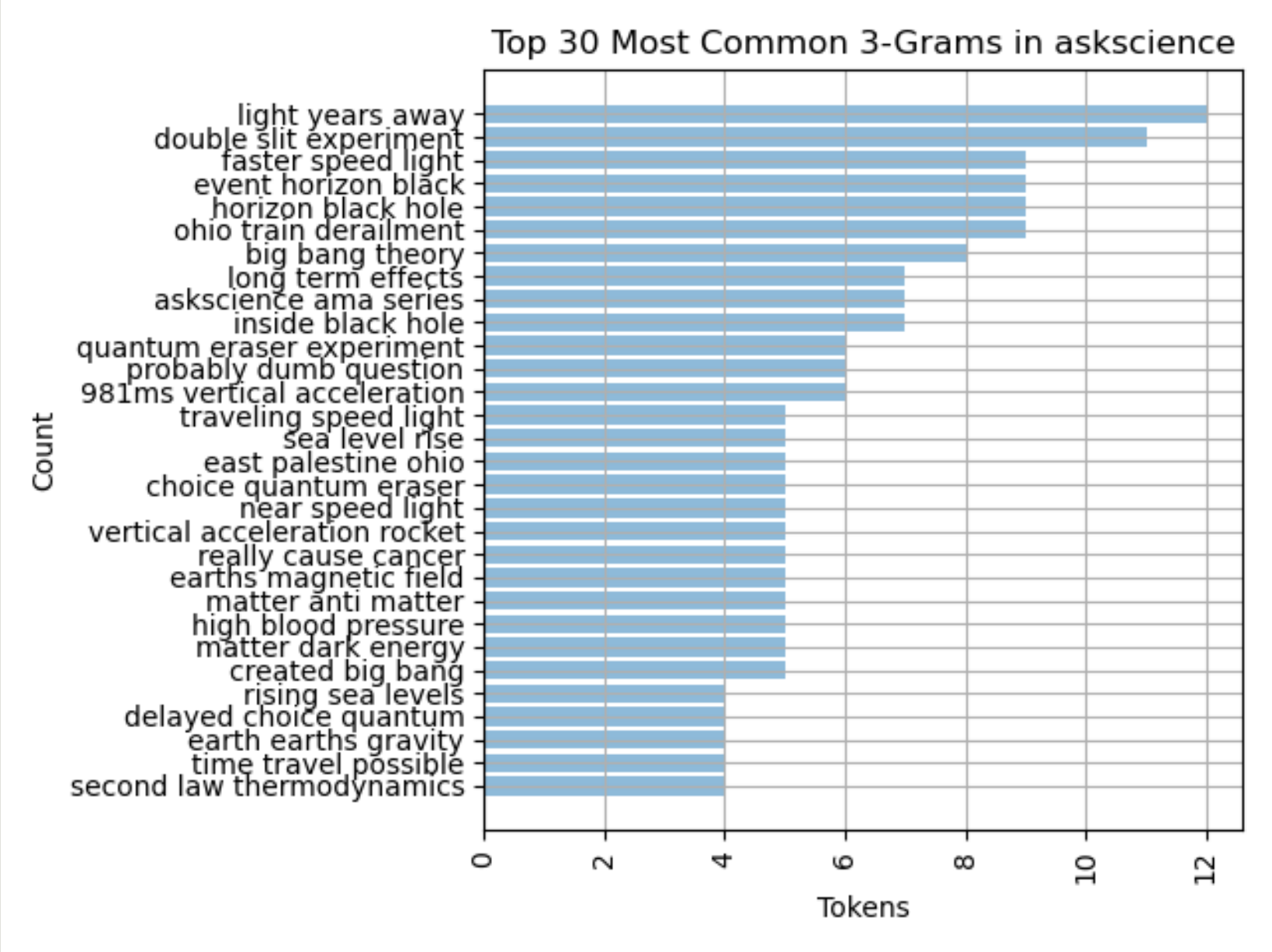
1. Data Gathering w/ Pushshift
2. Cleaning & EDA w/ Python
3. NLP Modeling w/ CVEC & TFIDF
4. Out-of-box Solution

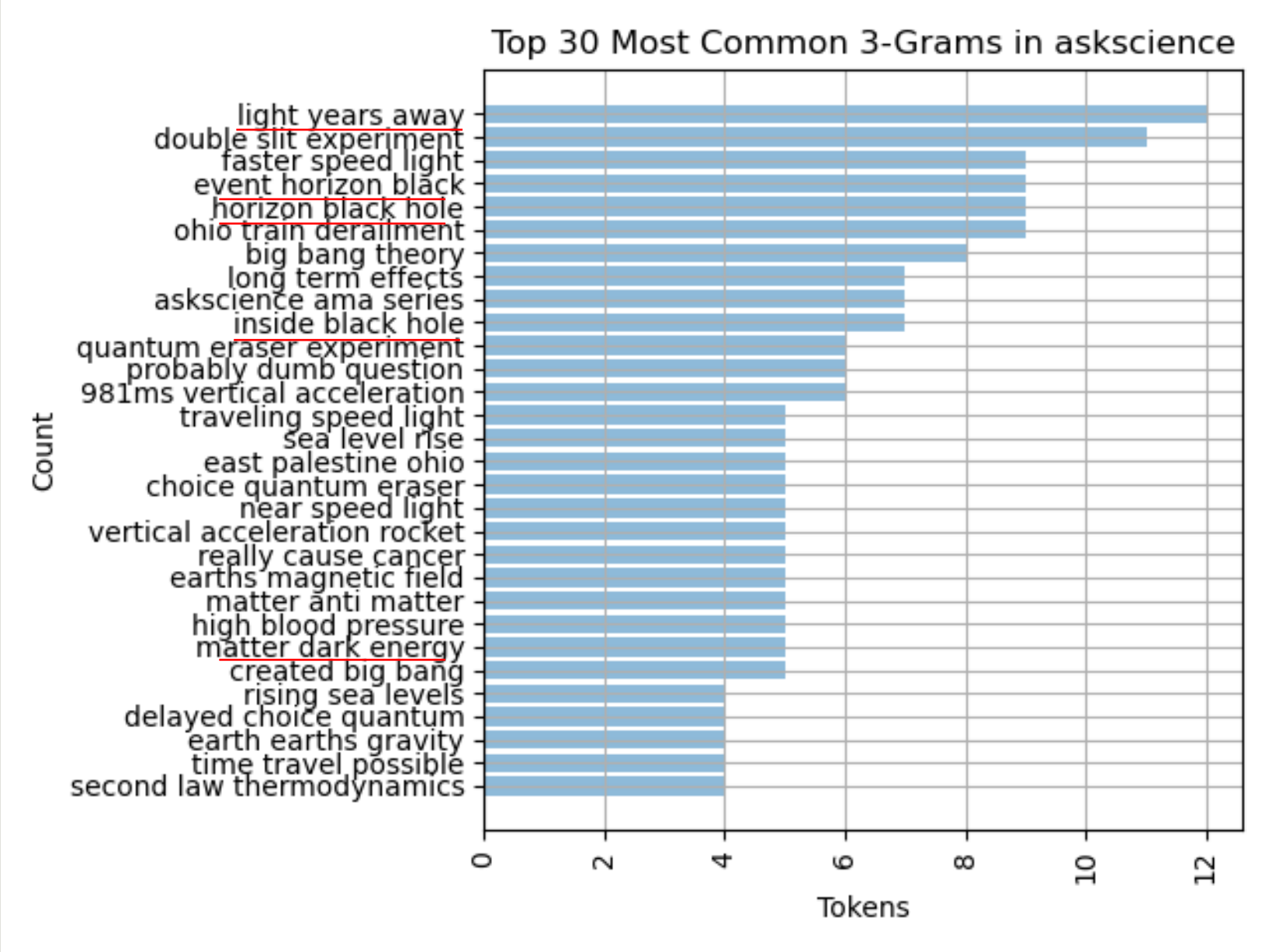
Question: By how much?

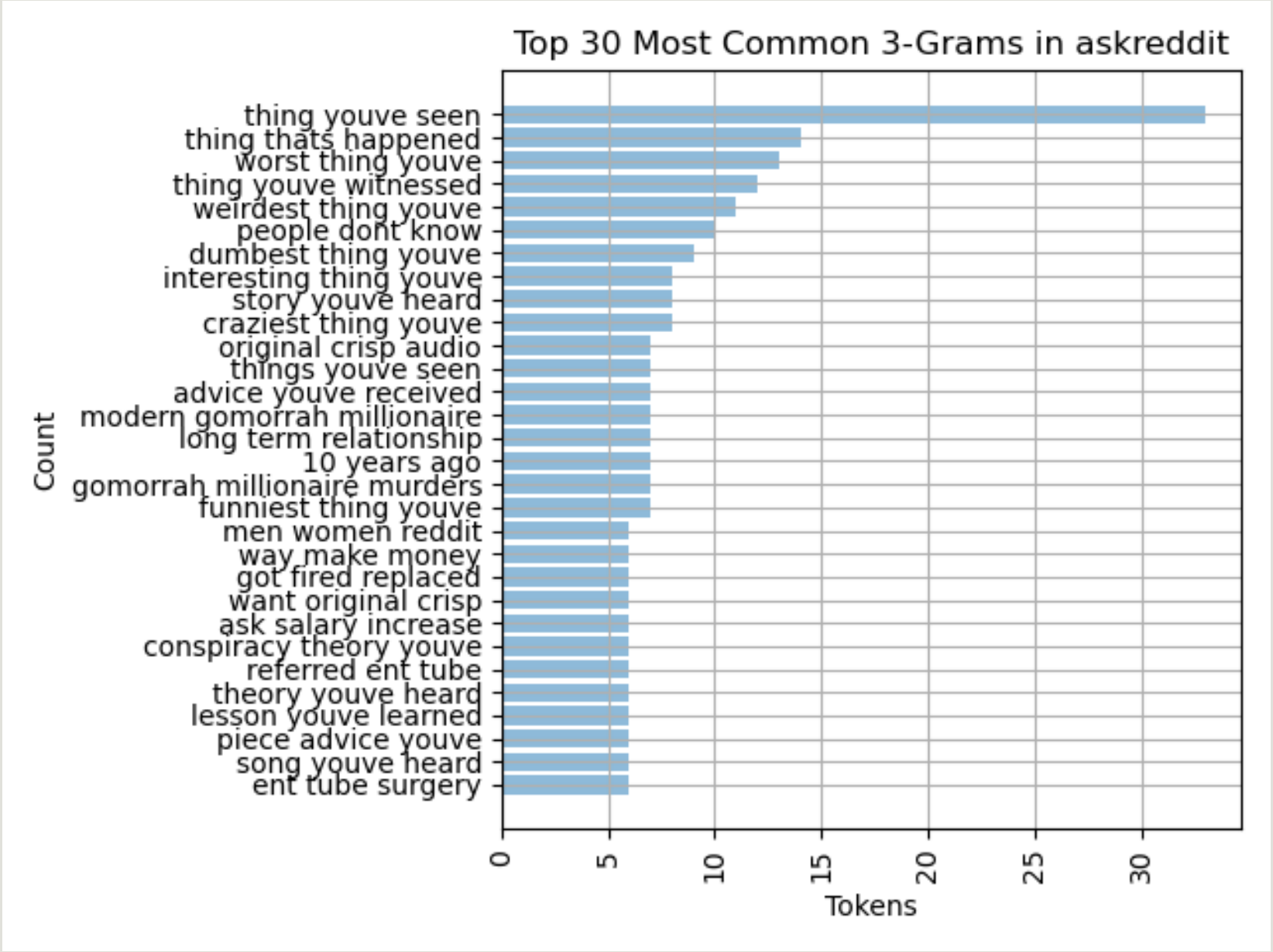


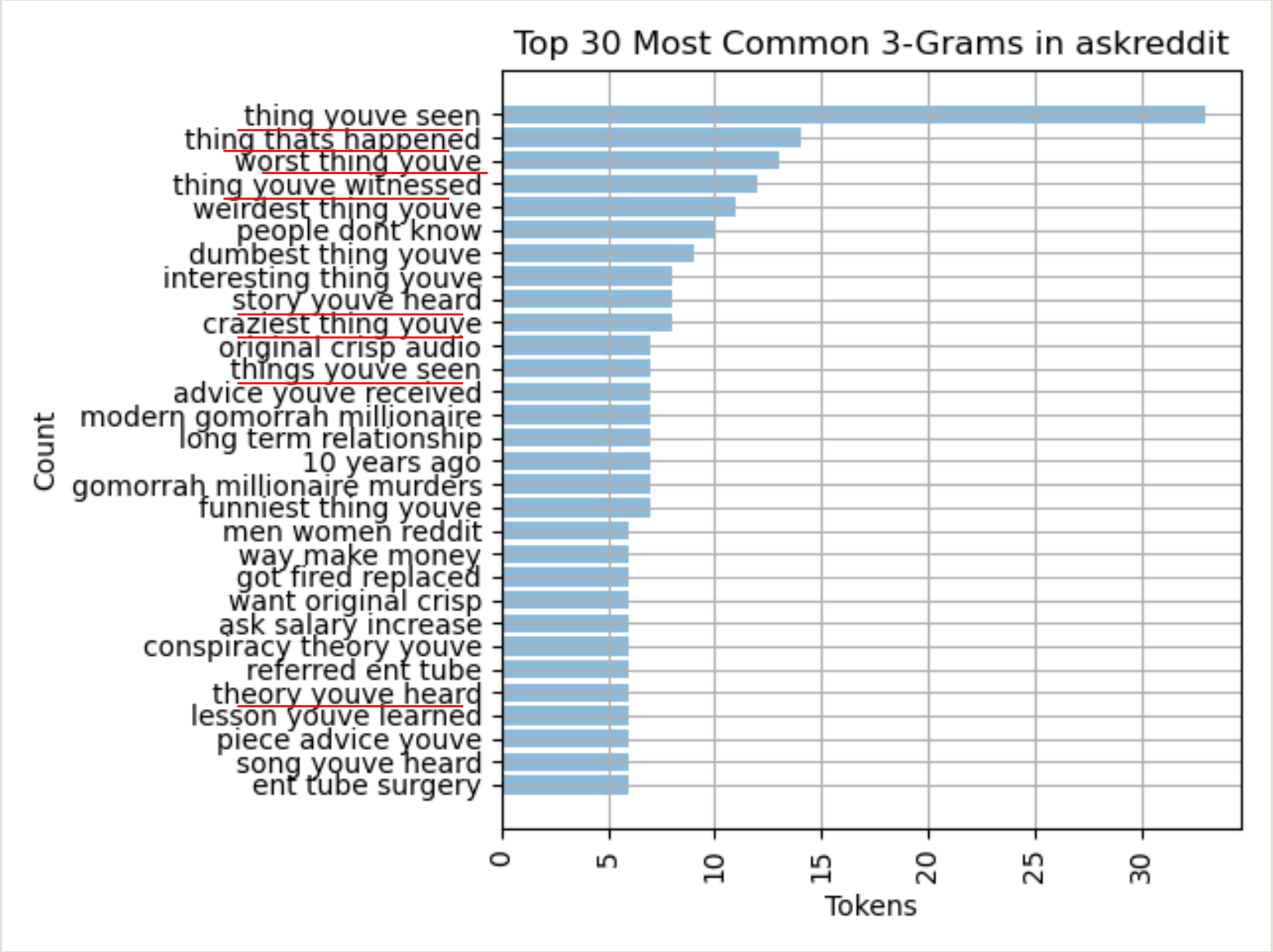
1. Problem Statement
2. Methodology
3. Analysis
4. Conclusion

Goal Classify both users equally


Top 30 Most Common 3-Grams in askscience
light years away
double slit experiment
faster speed light
event horizon black
horizon black hole
ohio train derailment
big bang theory
long term effects
askscience ama series
inside black hole
quantum eraser experiment
probably dumb question
981ms vertical acceleration
traveling speed light
sea level rise
east palestine ohio
choice quantum eraser
near speed light
vertical acceleration rocket
really cause cancer
earths magnetic field
matter anti matter
high blood pressure
matter dark energy
created big bang
rising sea levels
delayed choice quantum
earth earths gravity
time travel possible
second law thermodynamics
Count
0
2
4
6
8
10
12
Tokens

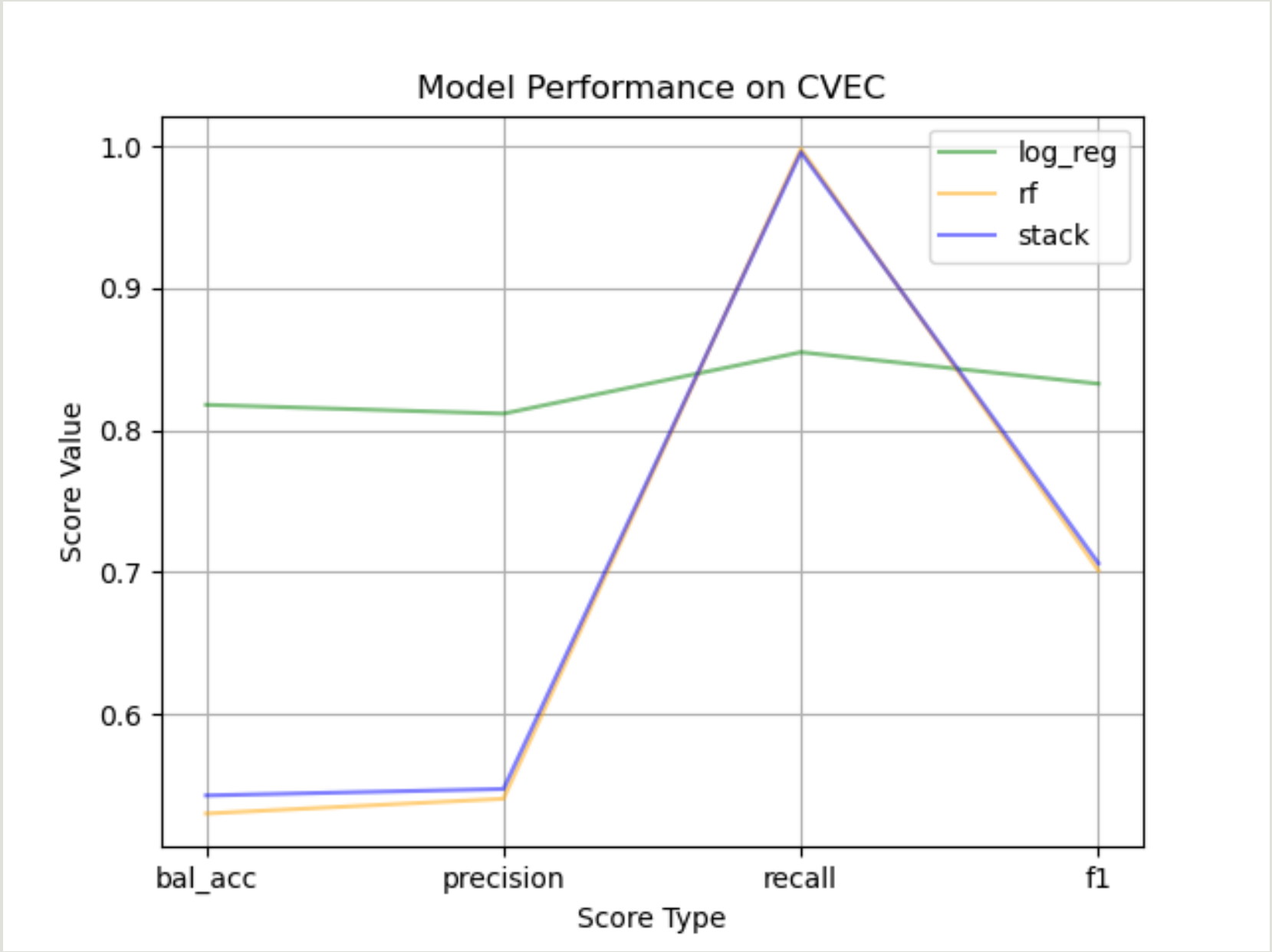

Top 30 Most Common 3-Grams in askscience
light years away
double slit experiment
faster speed light
event horizon black
horizon black hole
ohio train derailment
big bang theory
long term effects
askscience ama series
inside black hole
quantum eraser experiment
probably dumb question
981ms vertical acceleration
traveling speed light
sea level rise
east palestine ohio
choice quantum eraser
near speed light
vertical acceleration rocket
really cause cancer
earths magnetic field
matter anti matter
high blood pressure
matter dark energy
created big bang
rising sea levels
delayed choice quantum
earth earths gravity
time travel possible
second law thermodynamics
Count
0
2
4
6
8
10
12
Tokens

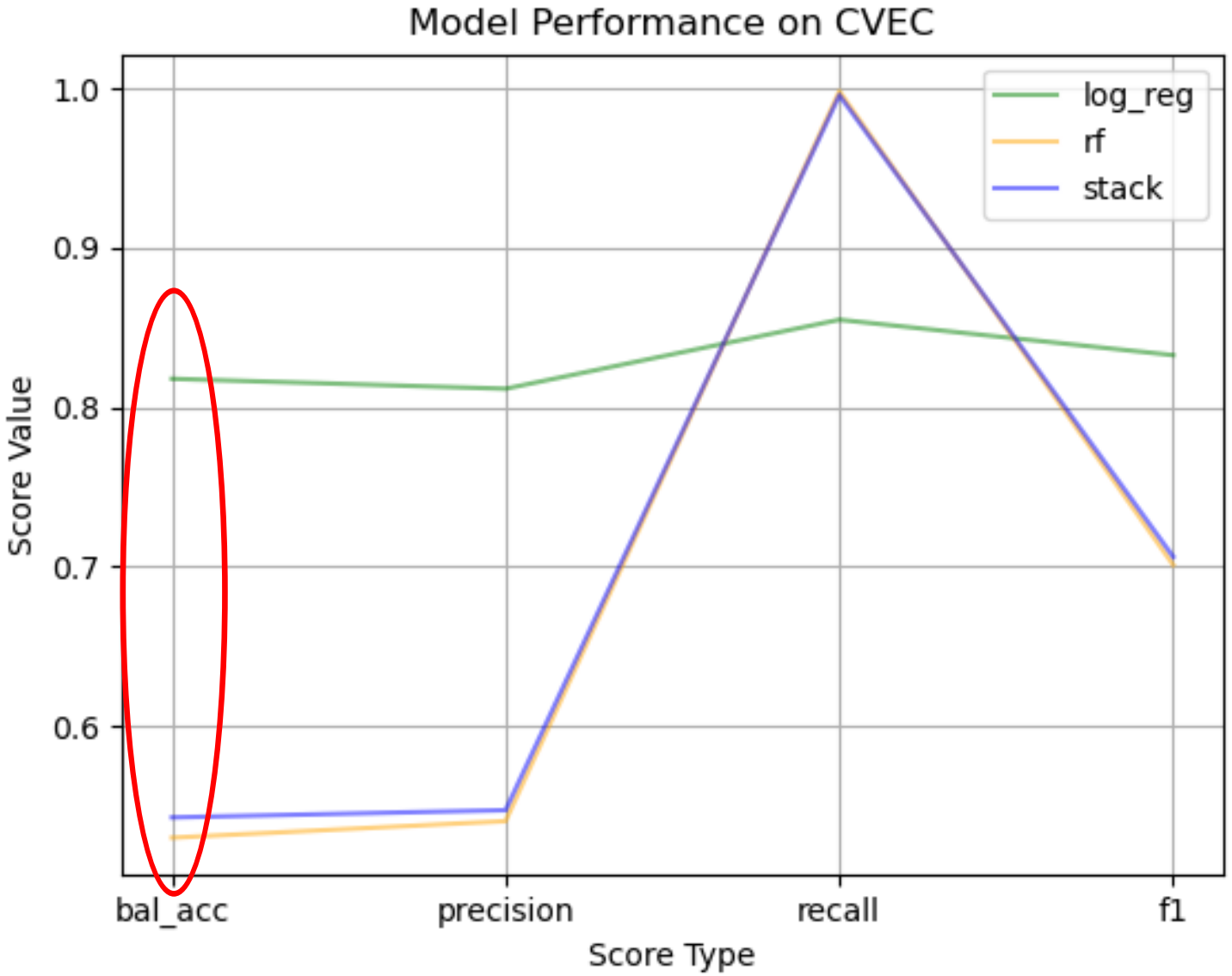

Top 30 Most Common 3-Grams in askreddit
thing youve seen
thing thats happened
worst thing youve
thing youve witnessed
weirdest thing youve
people dont know
dumbest thing youve
interesting thing youve
story youve heard
craziest thing youve
original crisp audio
things youve seen
advice youve received
modern gomorrah millionaire
long term relationship
10 years ago
gomorrah millionaire murders
funniest thing youve
men women reddit
way make money
got fired replaced
want original crisp
ask salary increase
conspiracy theory youve
referred ent tube
theory youve heard
lesson youve learned
piece advice youve
song youve heard
ent tube surgery
Count
0
5
10
15
20
25
30
Tokens

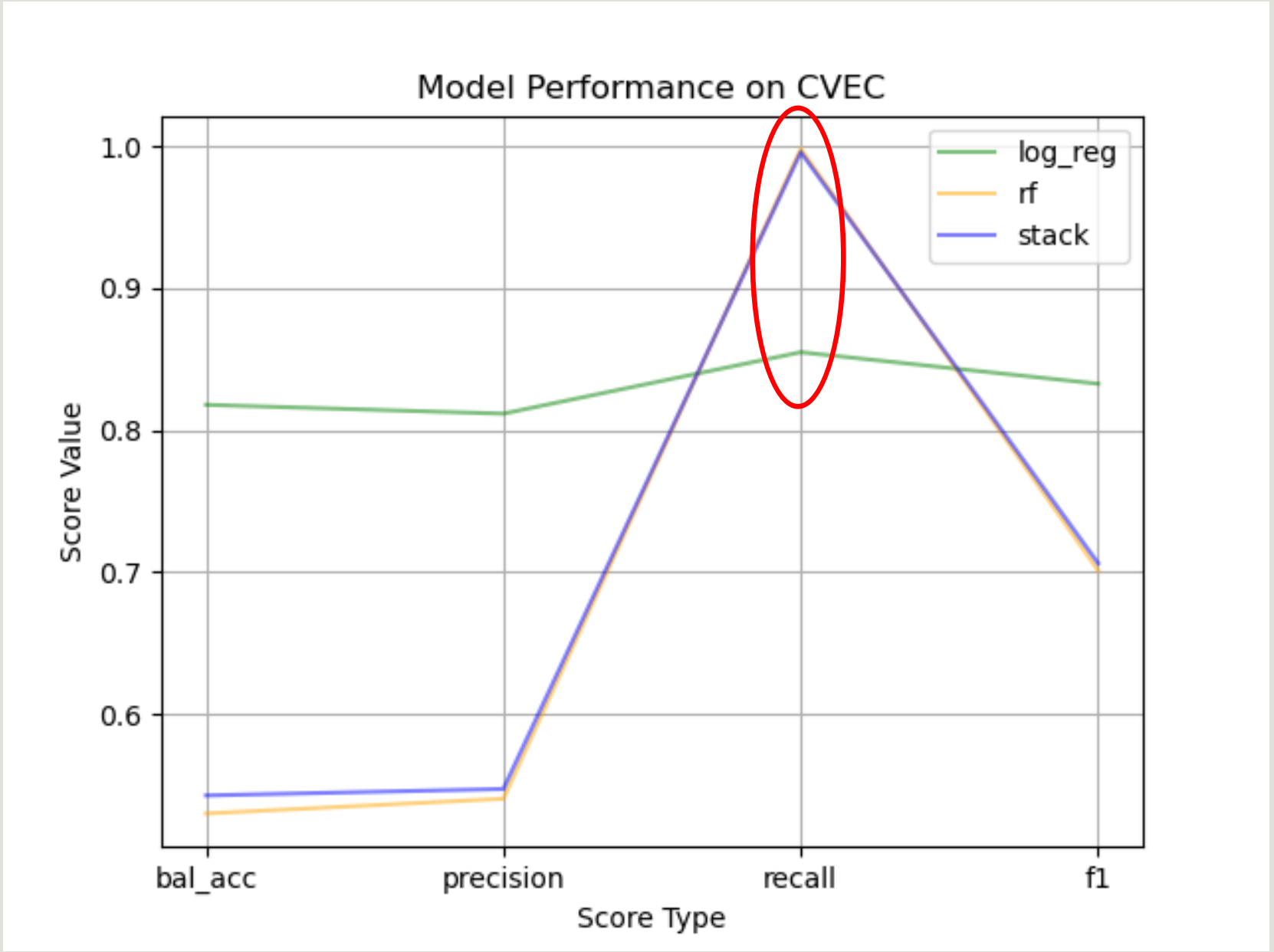

Top 30 Most Common 3-Grams in askreddit
thing youve seen
thing thats happened
worst thing youve
thing youve witnessed
weirdest thing youve
people dont know
dumbest thing youve
interesting thing youve
story youve heard
craziest thing youve
original crisp audio
things youve seen
advice youve received
modern gomorrah millionaire
long term relationship
10 years ago
gomorrah millionaire murders
funniest thing youve
men women reddit
way make money
got fired replaced
want original crisp
ask salary increase
conspiracy theory youve
referred ent tube
theory youve heard
lesson youve learned
piece advice youve
song youve heard
ent tube surgery
Count
0
5
10
15
20
25
30
Tokens

Solution

Pre-Trained Models


Model Performance on CVEC
log_reg
rf
stack
1.0
0.9
0.8
0.7
0.6
Score Value
bal_acc
precision
recall
f1
Score Type


Model Performance on CVEC
log_reg
rf
stack
1.0
0.9
0.8
0.7
0.6
Score Value
bal_acc
precision
recall
f1
Score Type


Model Performance on CVEC
log_reg
rf
stack
1.0
0.9
0.8
0.7
0.6
Score Value
bal_acc
precision
recall
f1
Score Type

1. Problem Statement
2. Methodology
3. Analysis
4. Conclusion

Recommendation

Recommendation

Logistic regression @ 0.82 balanced accuracy

Recommendation

Logistic regression @ 0.82 balanced accuracy
Random forest @ 0.98 recall

Future Considerations

Future Considerations

- Data scrapping script

Future Considerations

- Data scrapping script
- Cleaned data (25k)

Future Considerations

- Data scrapping script
- Cleaned data (25k)
- Hyperparam tuning flexibility

Future Considerations

- Data scrapping script
- Cleaned data (25k)
- Hyperparam tuning flexibility
- Score-preference flexibility

Executive Summary

This project can save you **time and money** by offering pre-trained models for your future NLP projects!

Thank you!

QUESTIONS?