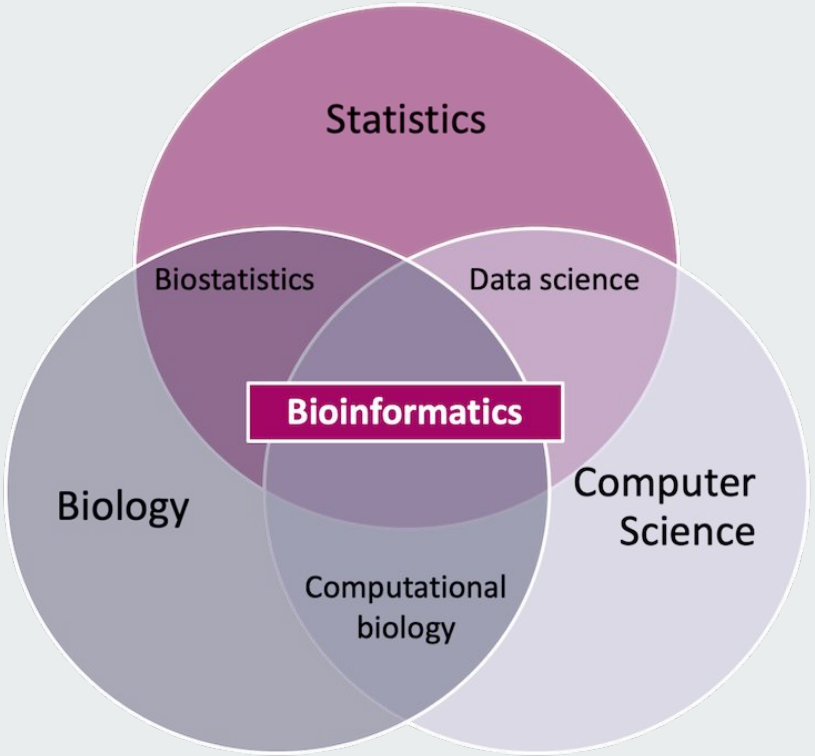
# Project 3 Bioinformatics and Data Science Reddits




by Tatiana Patrusheva
DSIR-1128

image from link

# Find the model to differentiate Bioinformatics from Data Science Subreddits

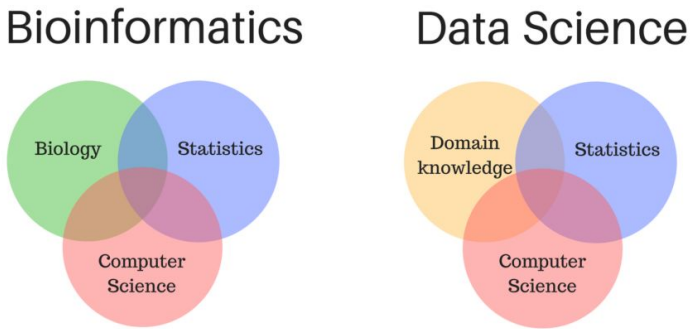
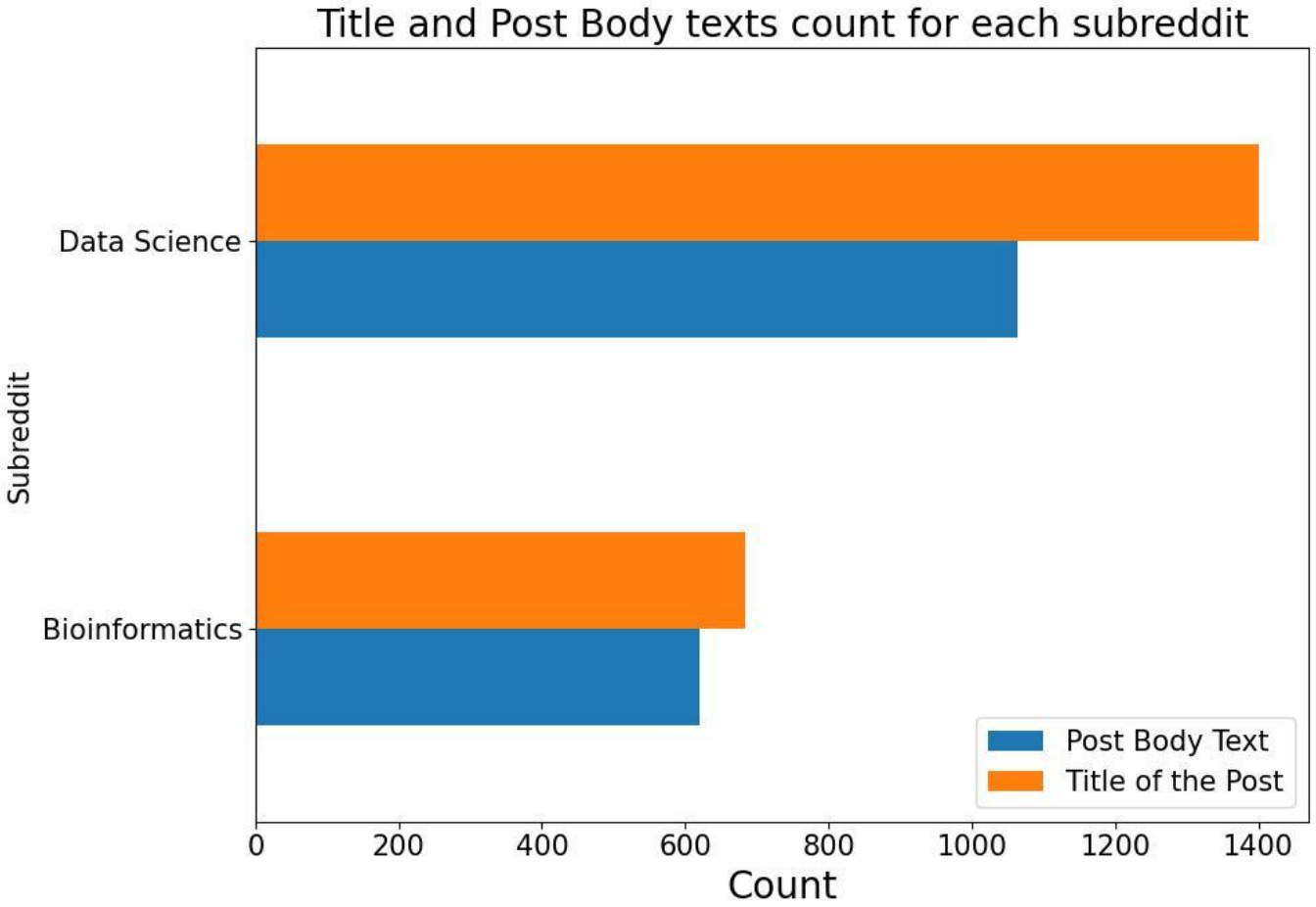
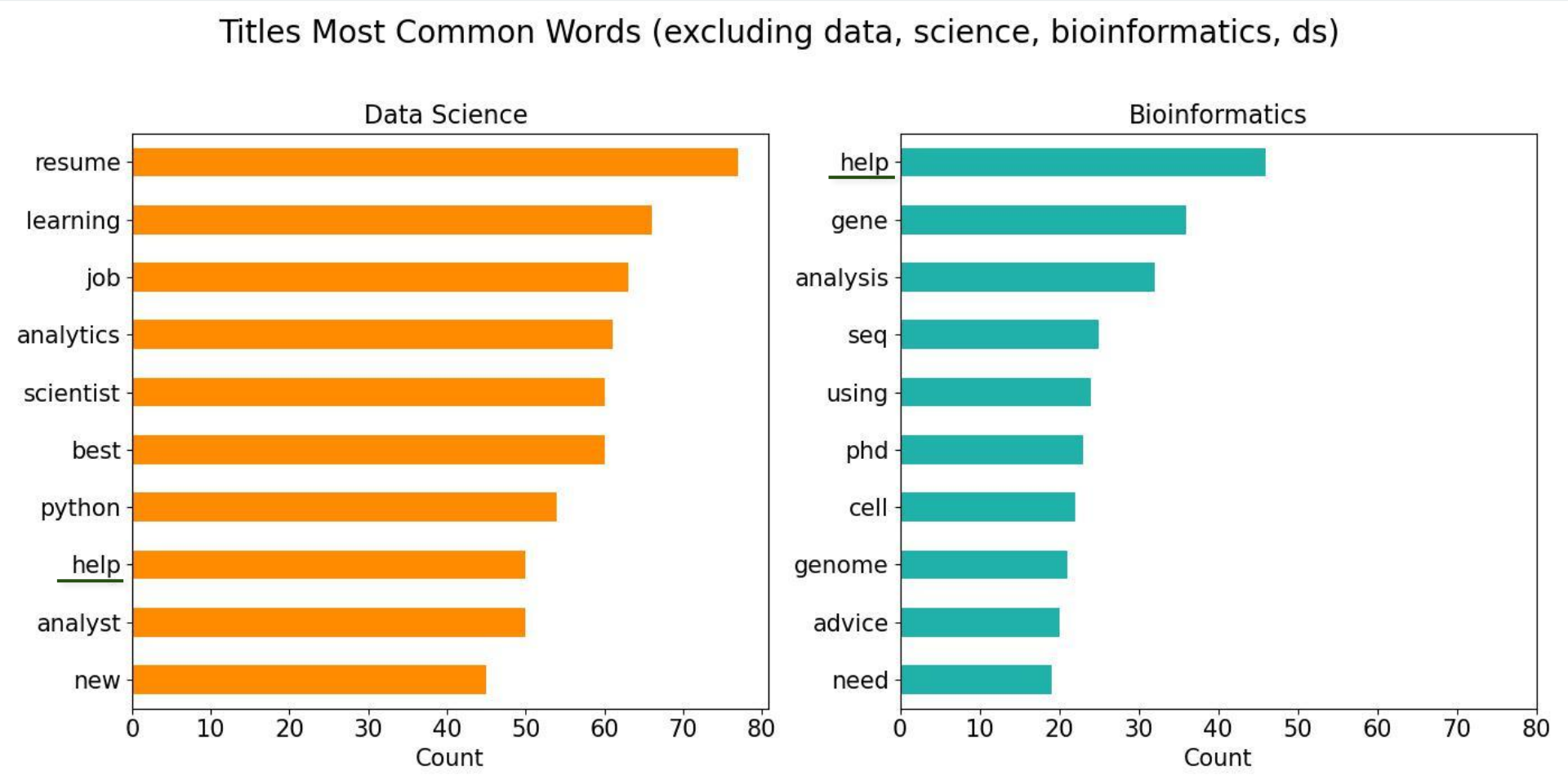
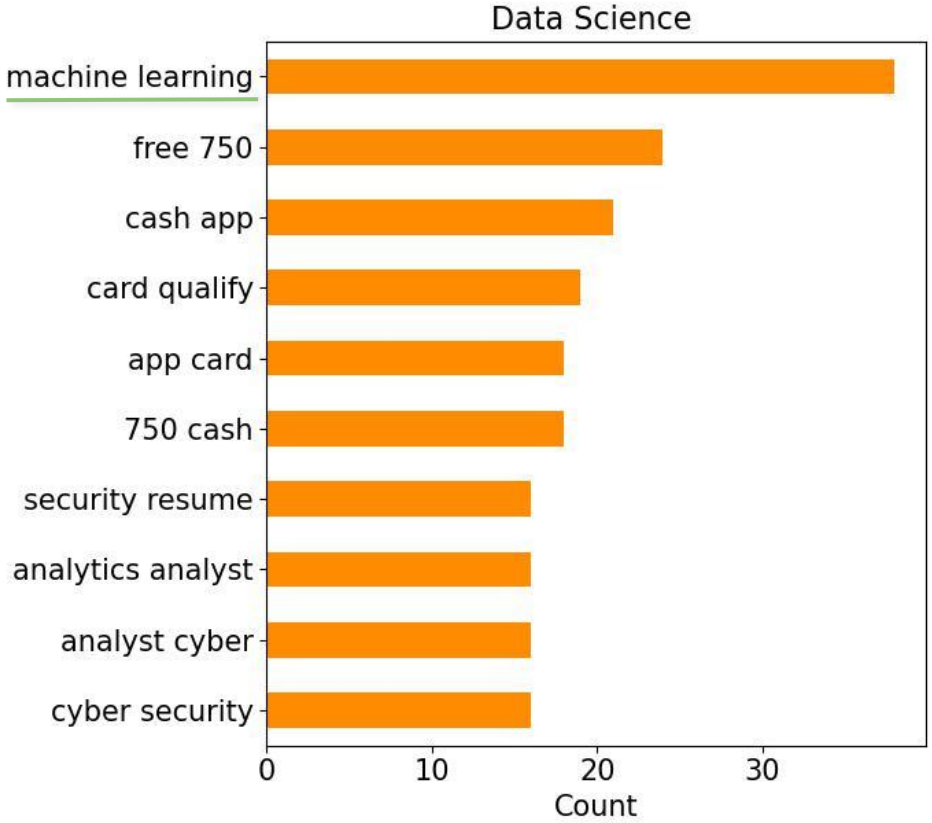


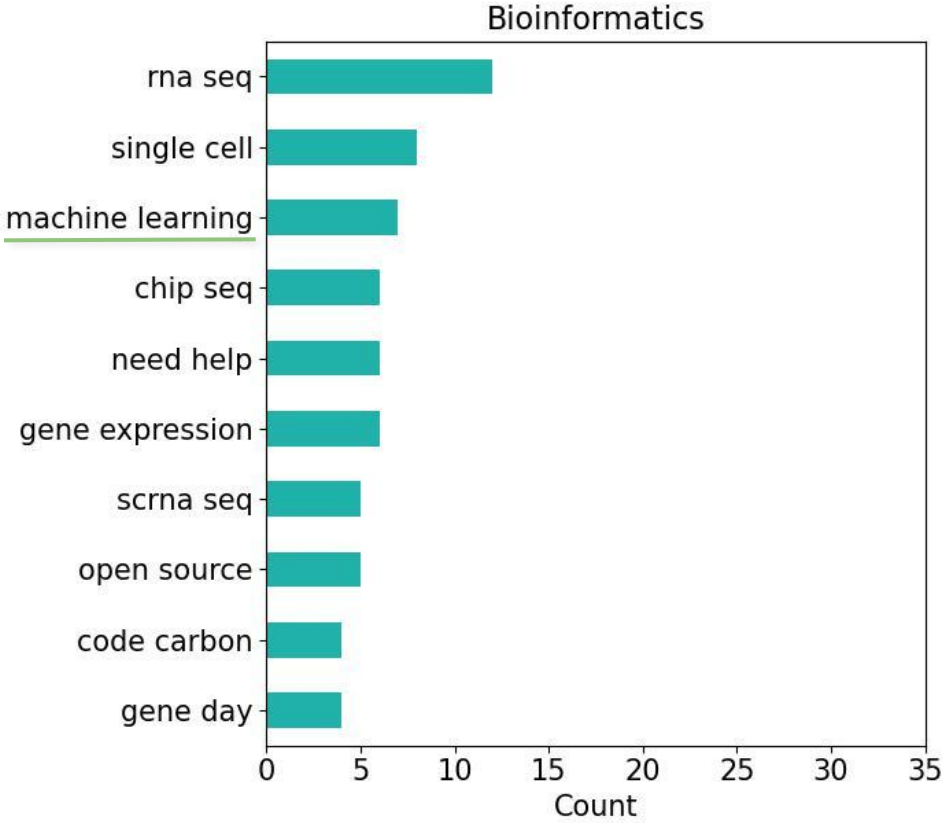

image from link

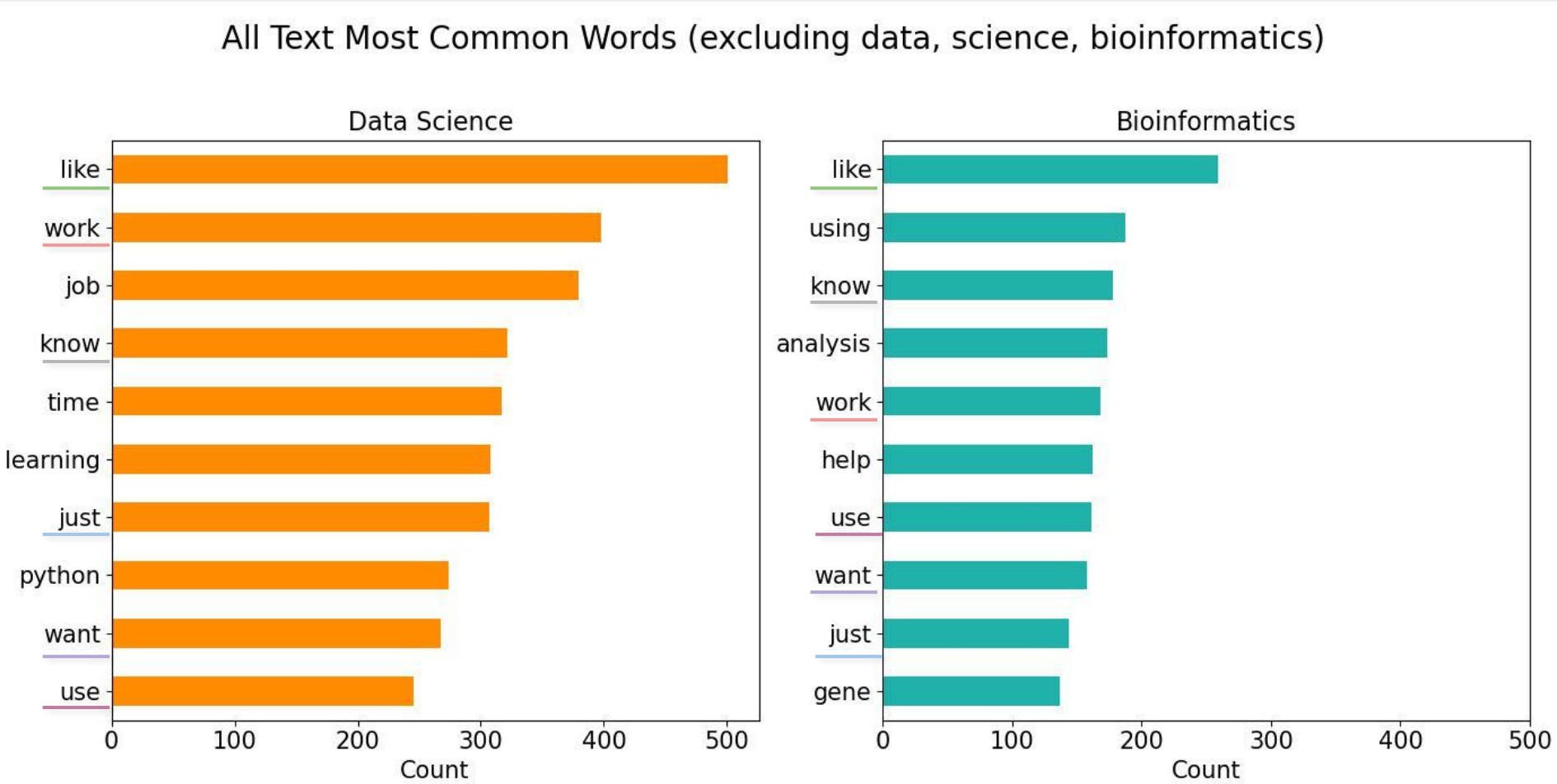

Title and Post Body texts count for each subreddit
Data Science
Bioinformatics
Subreddit
0
200
400
600
800
1000
1200
1400
Count
Post Body Text
Title of the Post


Titles Most Common Words (excluding data, science, bioinformatics, ds)
Data Science
resume
learning
job
analytics
scientist
best
python
help
analyst
new
0
10
20
30
40
50
60
70
80
Count
Bioinformatics
help
gene
analysis
seq
using
phd
cell
genome
advice
need
0
10
20
30
40
50
60
70
80
Count

# Title Most Common Bigrams (excluding data, science, bioinformatics)



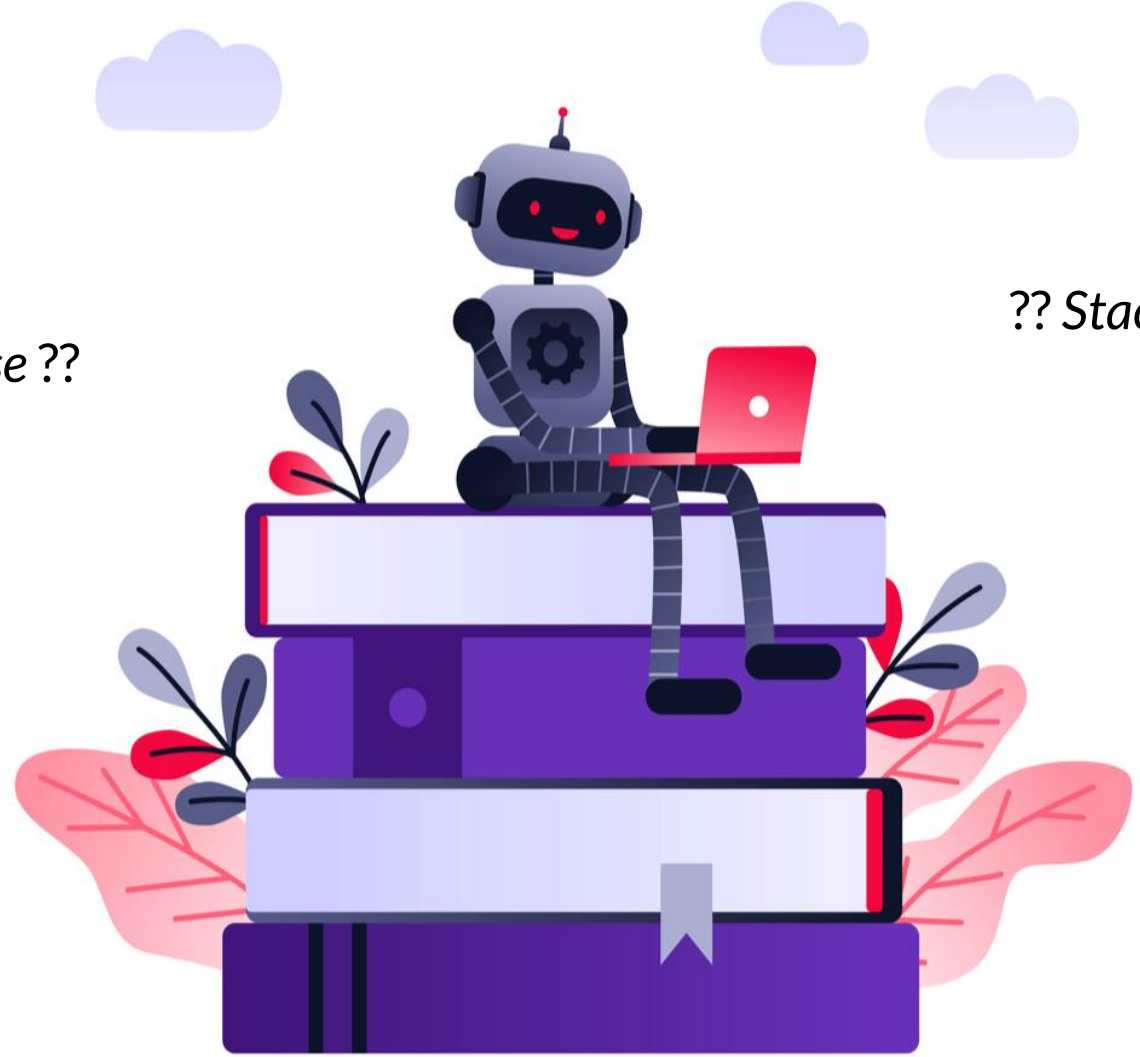




All Text Most Common Words (excluding data, science, bioinformatics)
Data Science
like
work
job
know
time
learning
just
python
want
use
0
100
200
300
400
500
Count
Bioinformatics
like
using
know
analysis
work
help
use
want
just
gene
0
100
200
300
400
500
Count

?? *XGBoost* ??

?? *Naive Base* ??

?? *Stacking* ??

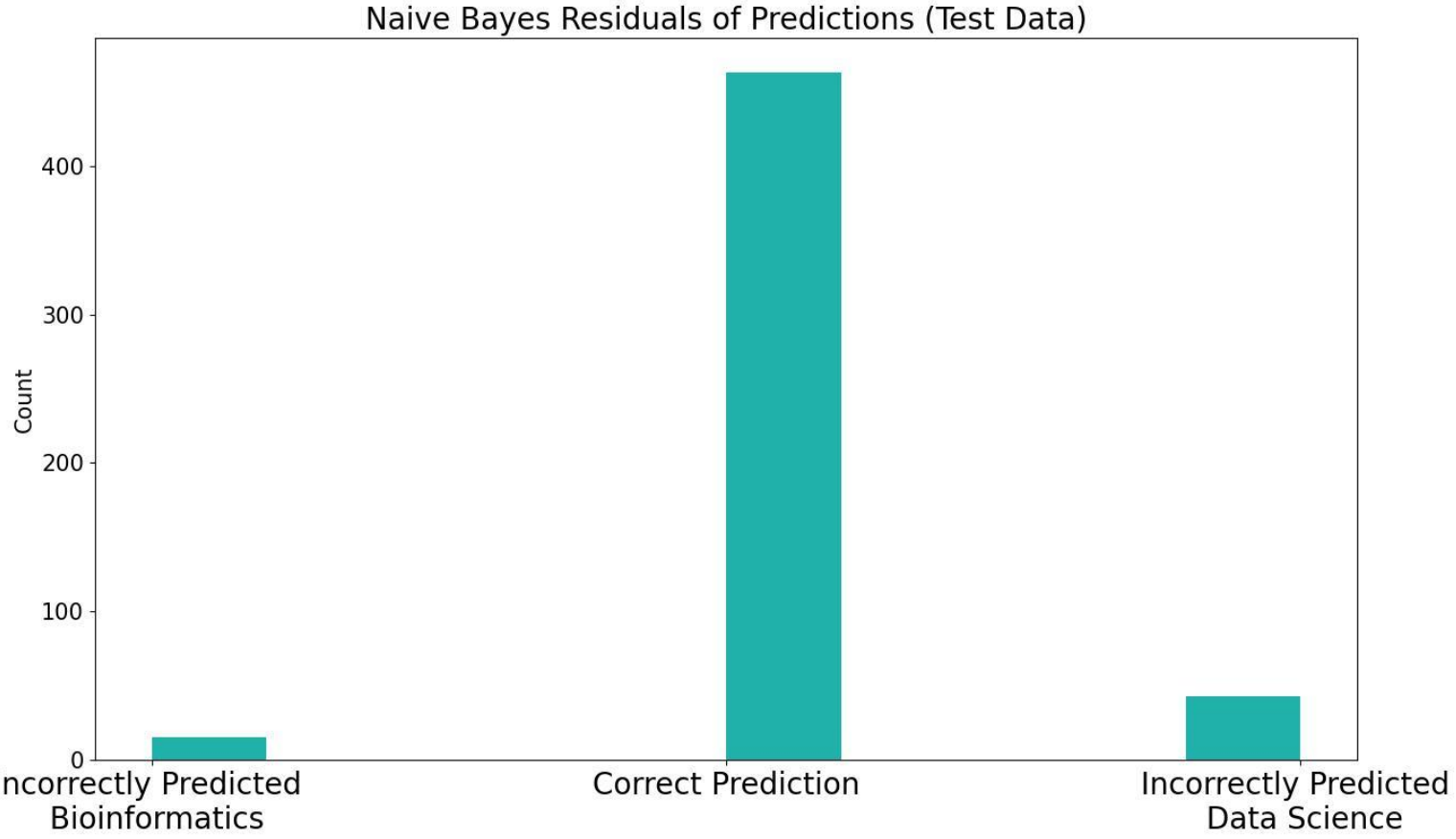


My baseline
is 67%

# Naive Bayes best model

F1 score:
0.82

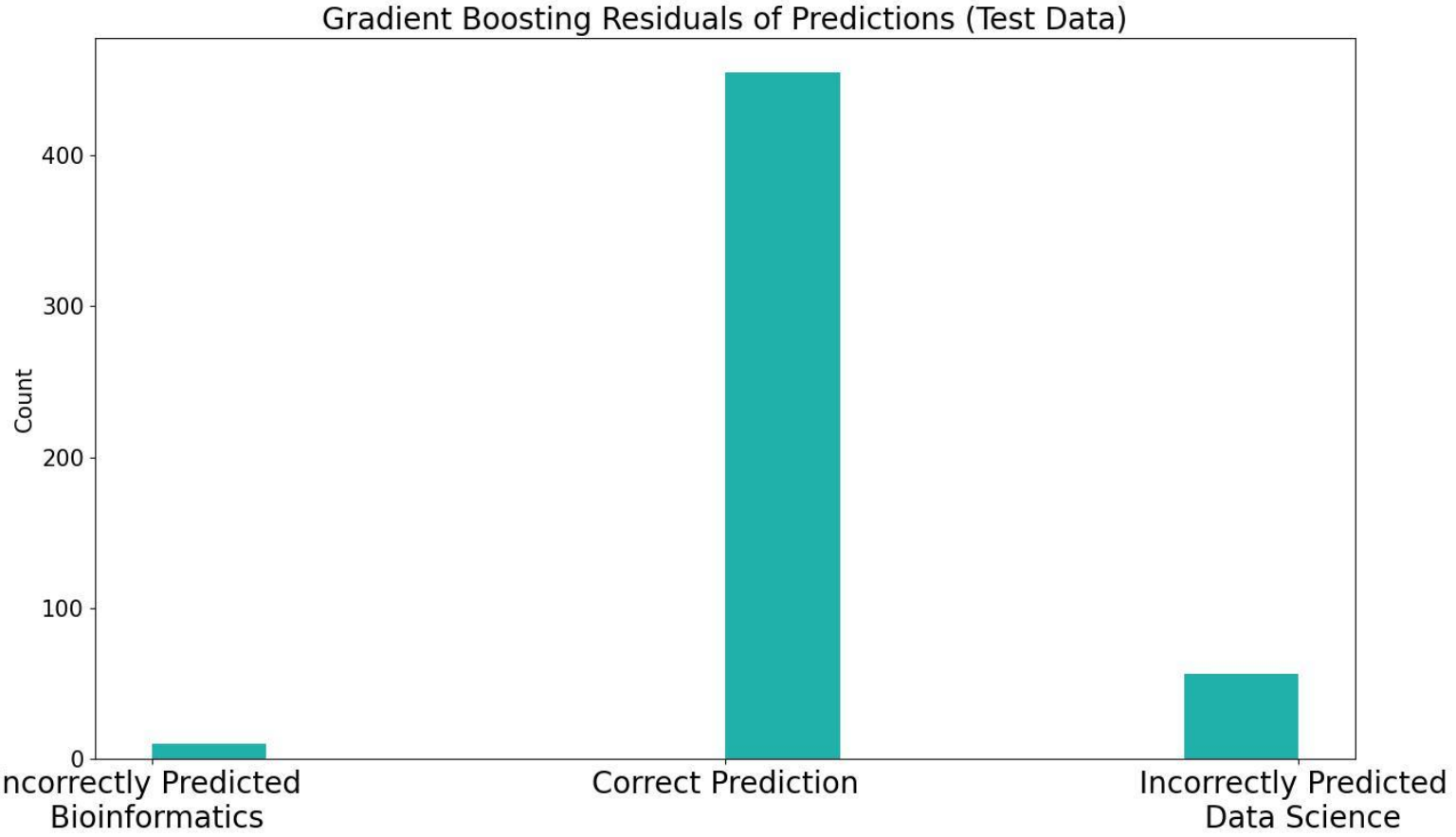
Test accuracy:
89%



# Gradient Boosting best model

F1 score:
0.78

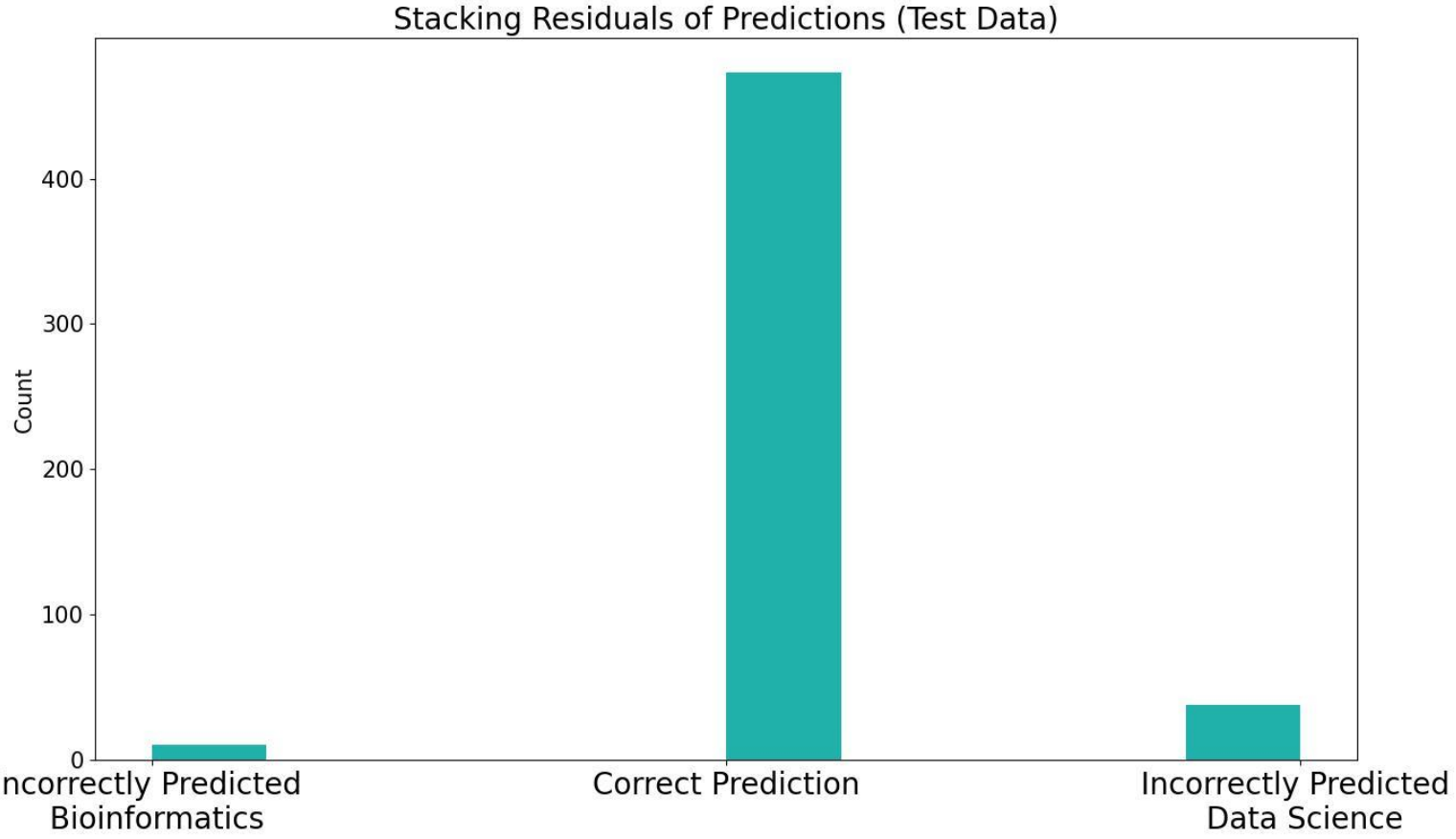
Test accuracy:
87%



# Stacking Best Model

F1 score:
0.85

Test accuracy:
91%



# Conclusions

## Best Model - Stacking with Tokenization

## Recommendations

- Data from different websites

- Balance between amount of time to fit and performance

Thank you!