Subreddit Classification

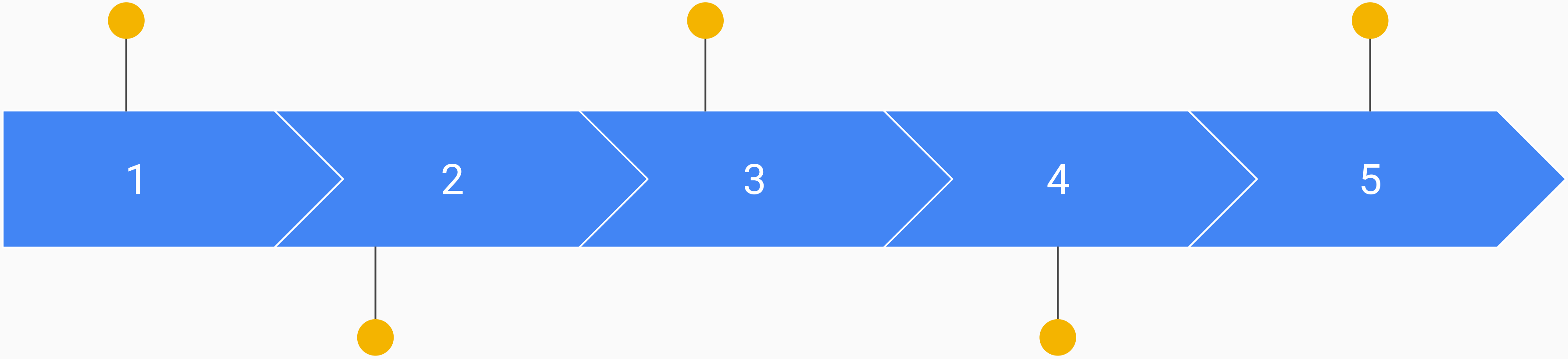
Manjila Moktan
March 21, 2025

BACKGROUND

- Declutter focuses on reducing clutter and organising spaces.
- Minimalism focuses on living with less.

Problem Statement :

This project aims to scrap through two subreddits(Declutter and Minimalism) and apply Natural language Processing (NLP) techniques to build a supervised ML model that predicts if the post belongs to subreddit Declutter or Minimalism with high accuracy, despite their similarities and differences.



Data Collection:

reddit Praw (.new , .top)

Data Preprocessing:

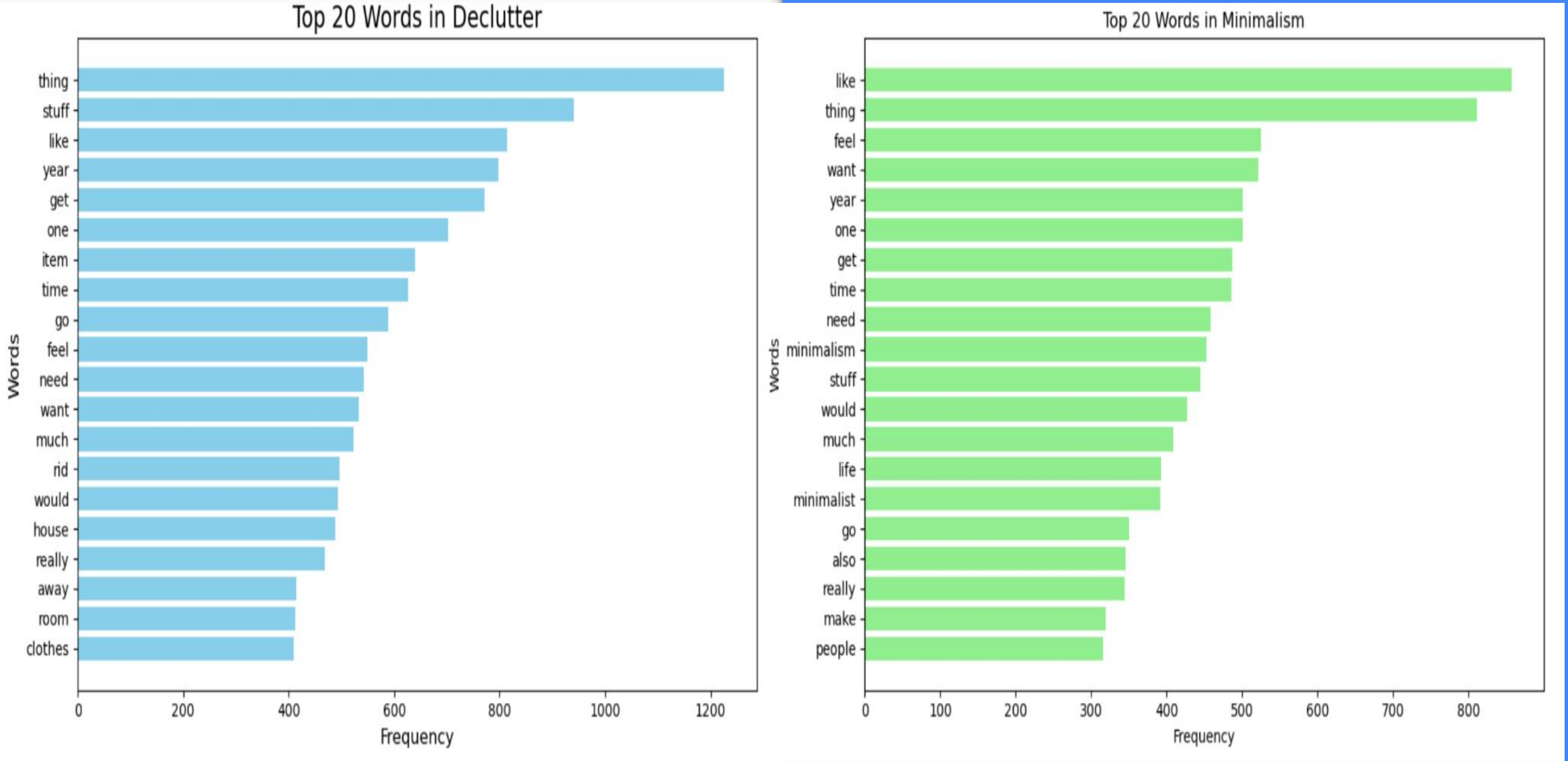
- Binarize the target
- top 10 words
- WordNetLemmatizer

Feature Extraction

-TF-IDF Vectorizer

Model Building
Evaluation

Conclusion

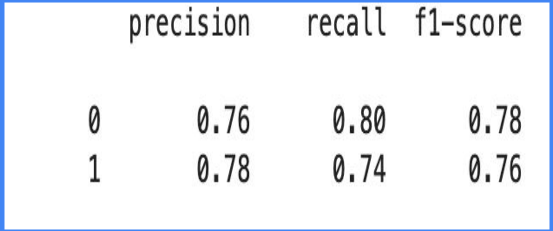


Among the top 20 words from each subreddits in text; total of 40, only 26 words are unique!!!

Model Evaluation

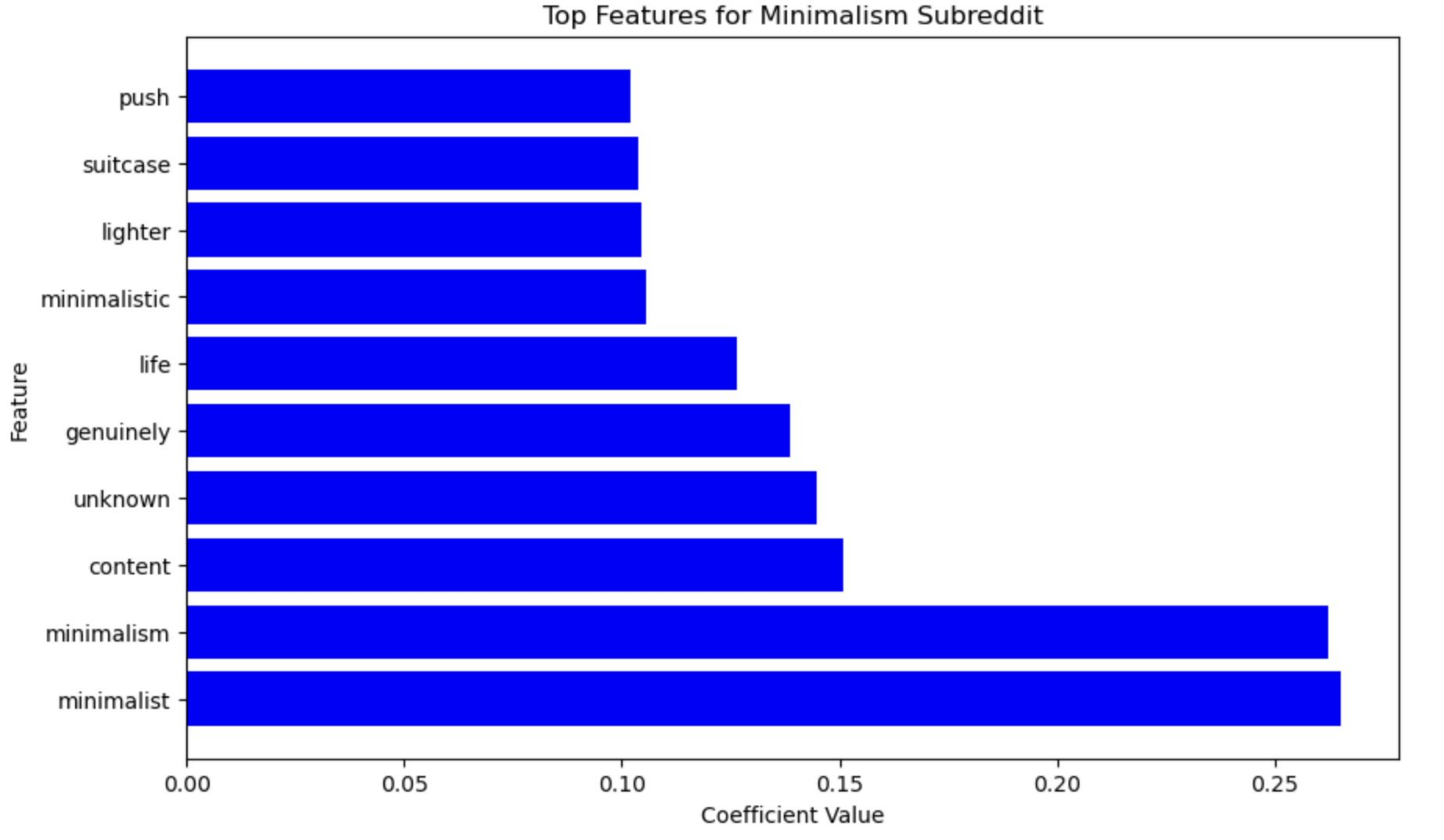
- Logistic Regression Model

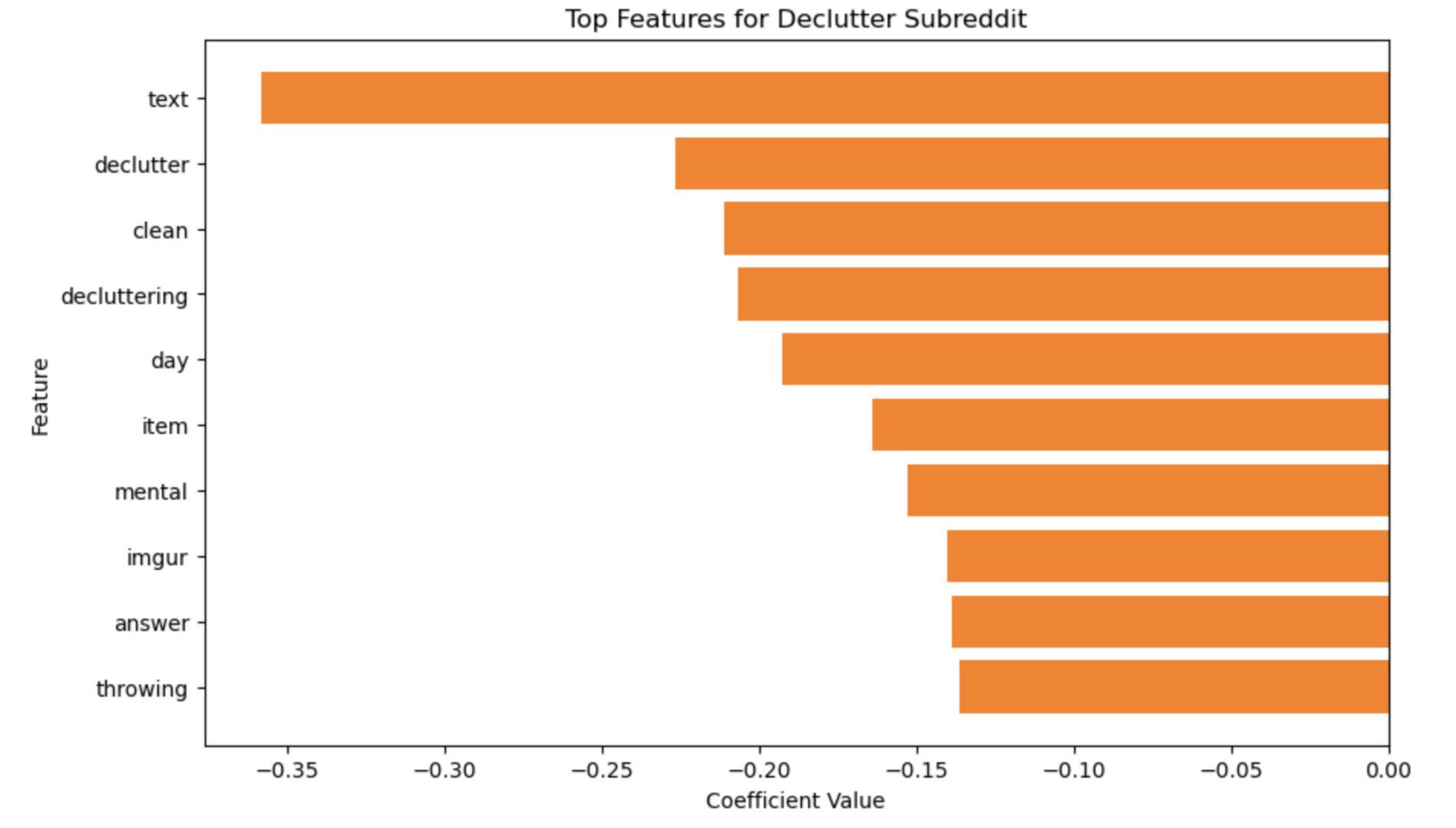
(Accuracy = 77%)



```
     precision    recall  f1-score

0         0.76      0.80      0.78
1         0.78      0.74      0.76
```


Top Features for Minimalism Subreddit
Feature
push
suitcase
lighter
minimalistic
life
genuinely
unknown
content
minimalism
minimalist
0.00
0.05
0.10
0.15
0.20
0.25
Coefficient Value


Top Features for Declutter Subreddit
Feature
text
declutter
clean
decluttering
day
item
mental
imgur
answer
throwing
−0.35
−0.30
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
Coefficient Value

Conclusion And Recommendation

- Use of Logistic Regression
- Use aggressive stemming eg. PorterStemmer instead
- Exploring other techniques