reddit

Subreddit Posts Classification

Sibel Tanoglu

Problem Statement

I will use Reddit's data to build a classification model, predicting which subreddit a particular post belongs to, in order to aid several environmental organizations to create more influential campaigns.

Overview

- I selected **“r/Anticonsumption”** and **“r/minimalism”** subreddits for this project. Used NLP (Natural Language Processing) techniques to train a binary classification model for most accurate predictions.

- My model will provide an automation to read large volume of textual data, interpret it, measure sentiment, and determine the user’s lifestyle choices.

- It’s important to distinguish people’s language and sentiment in both communities to create more influential campaigns that promote consuming consciously and appreciating simplicity. This will lead to a more sustainable future for the planet.

Data Collection

- The dataset contains information (post titles, post texts, total upvotes, total comments, create date) on newest submissions collected through Reddit's Web API.
- 7009 documents (rows) and 5 features (columns) of said info.

Sample Posts

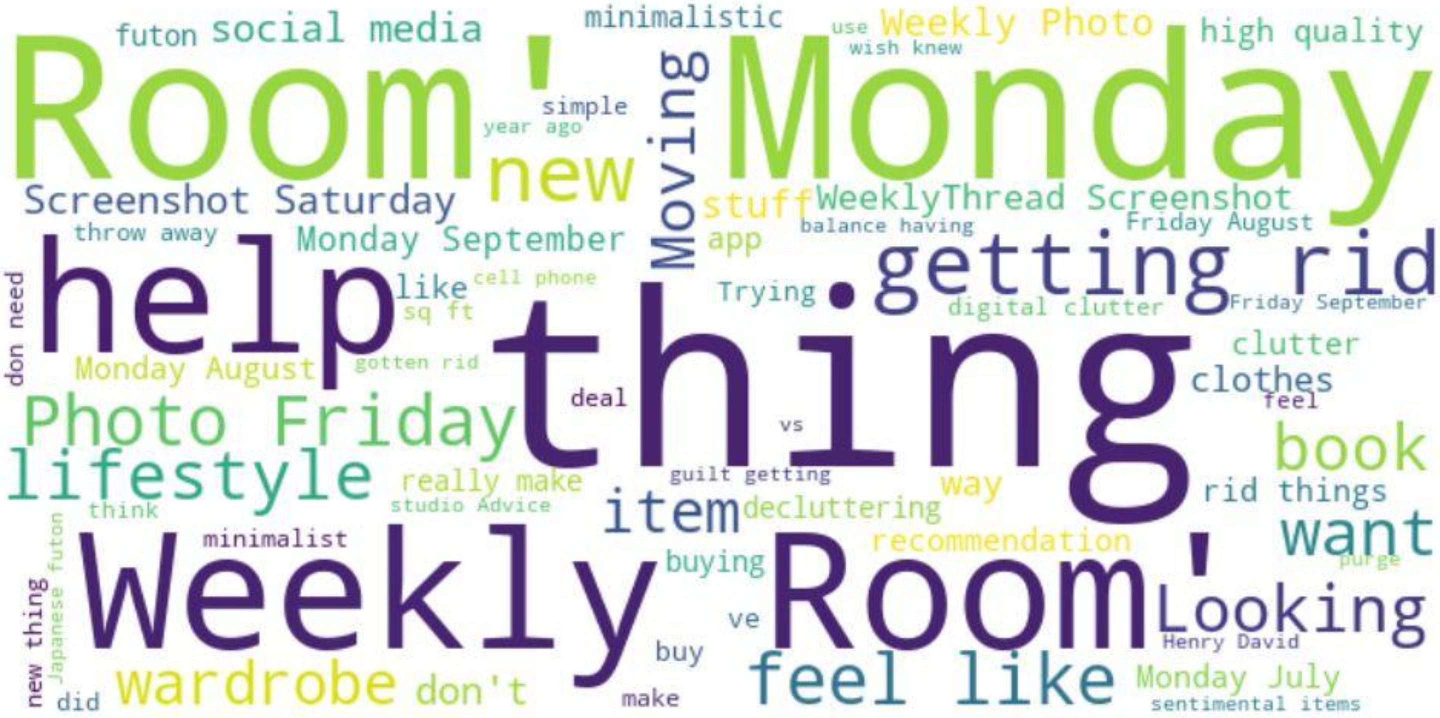
r/Anticonsumption

I just realized that I've been using and refilling this soap bottle for more than 20 years.

r/minimalism

Should I keep a pair of "nice" shoes that I never wear?

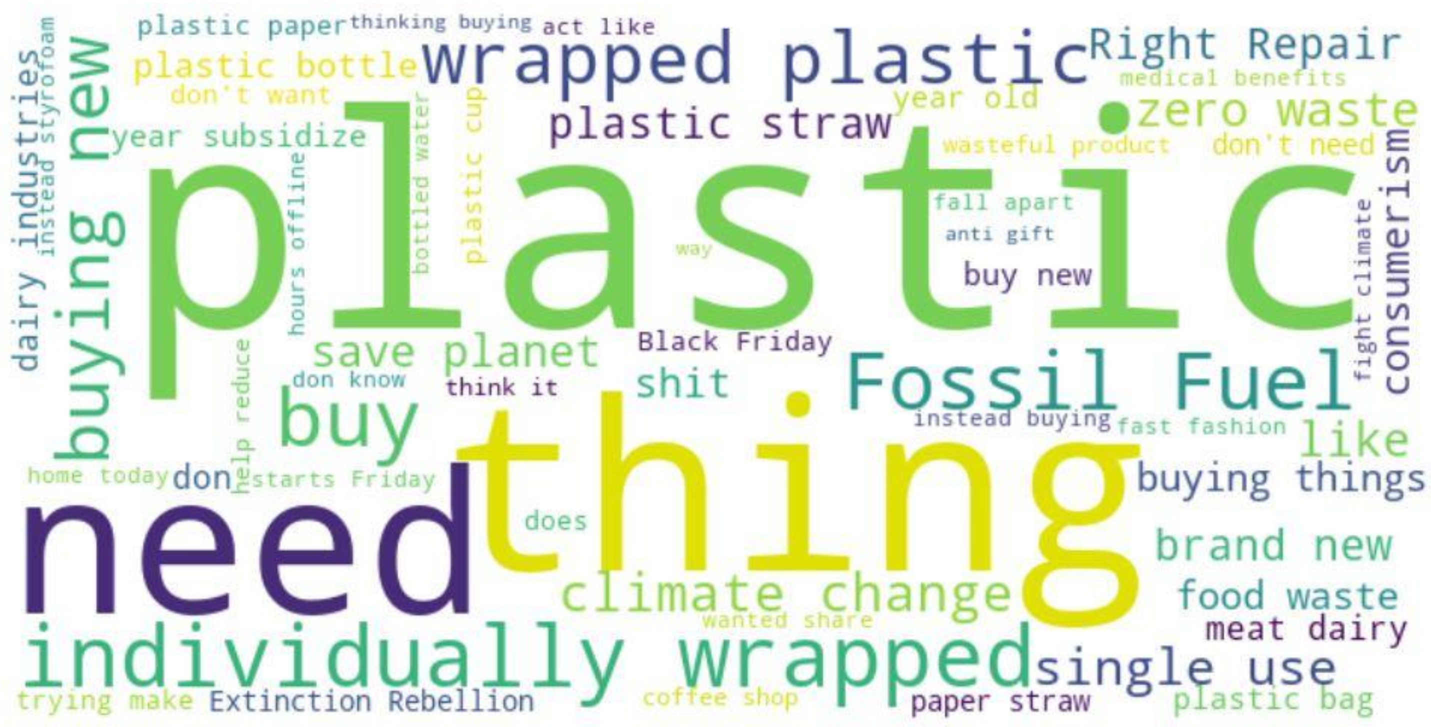
Words / r/minimalism



- thing
- item
- stuff
- decluttering
- room
- Monday
- wardrobe
- clothes
- simple
- throw away
- high quality
- sentimental
- lifestyle
- Henry David
- etc.

<u>Disclaimer:</u> Word clouds do not represent any statistical inference, use them responsibly!

Words / r/Anticonsumption



- plastic
- thing
- wrapped
- fossil fuel
- buying
- zero waste
- single use
- climate change
- save
- planet
- plastic straw
- sh.t
- extinction rebellion
- right repair etc.

<u>Disclaimer:</u> **Word clouds do not represent any statistical inference, use them responsibly!**

Data Cleaning & EDA

✓ Merged 6 scraped data files, removed duplicates

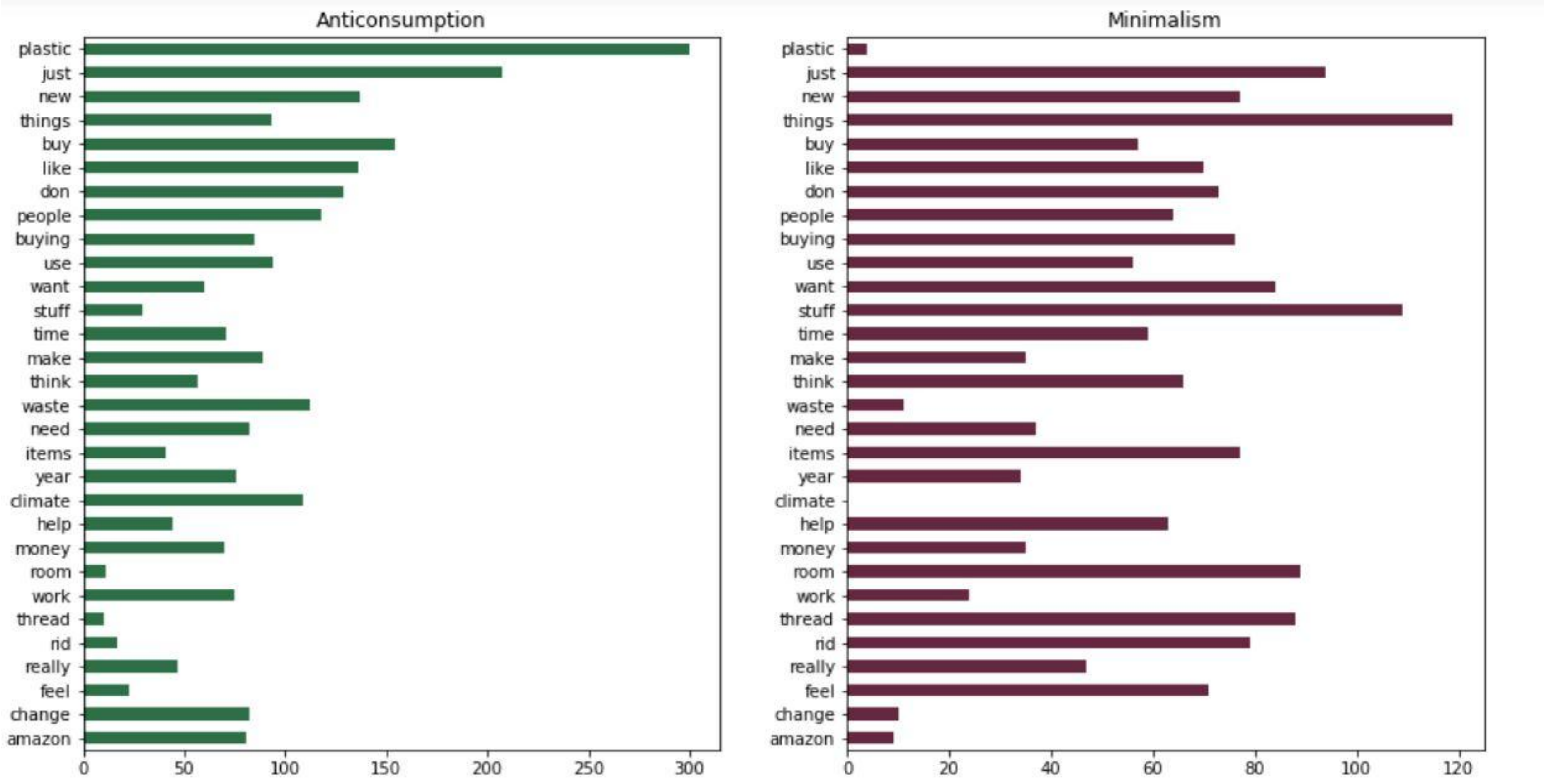
✓ Checked null values and decided to analyze only Title columns

✓ Vectorized text, identified most correlated words

✓ Plotted words, upvotes, comments distributions

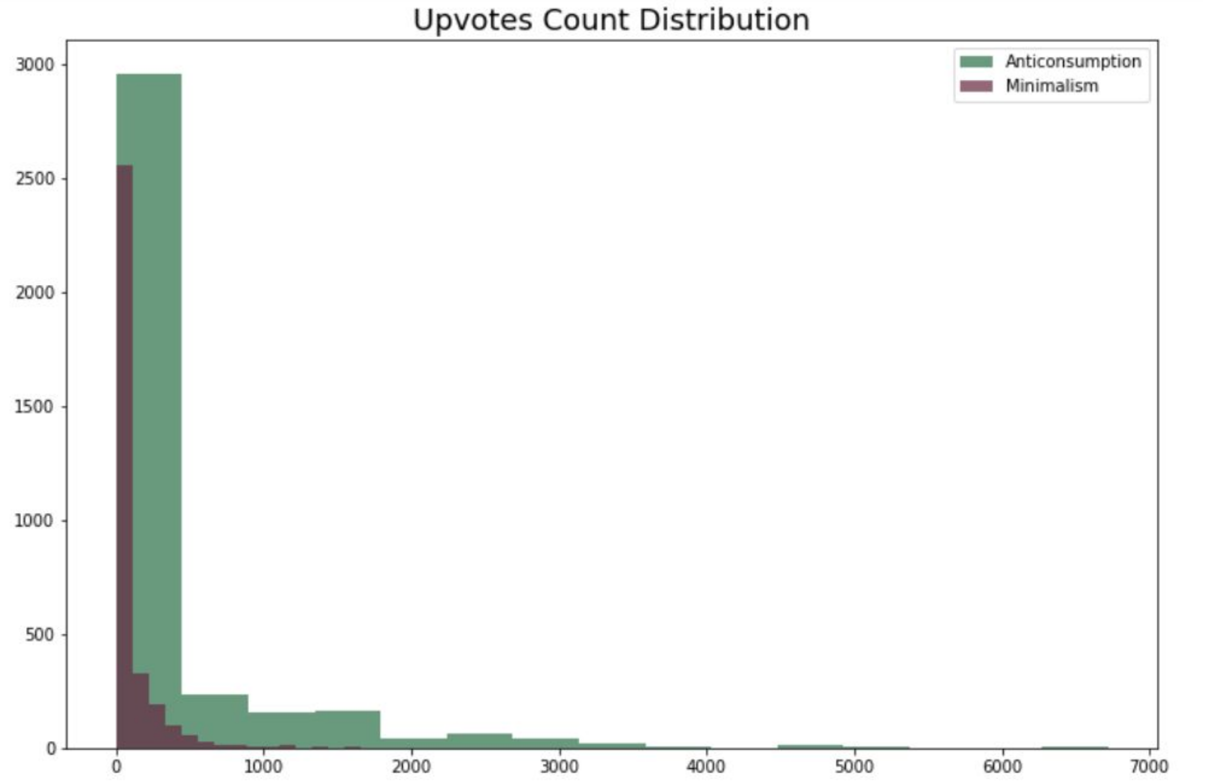
✓ Added new column for composite sentiment score

Top 30 Words

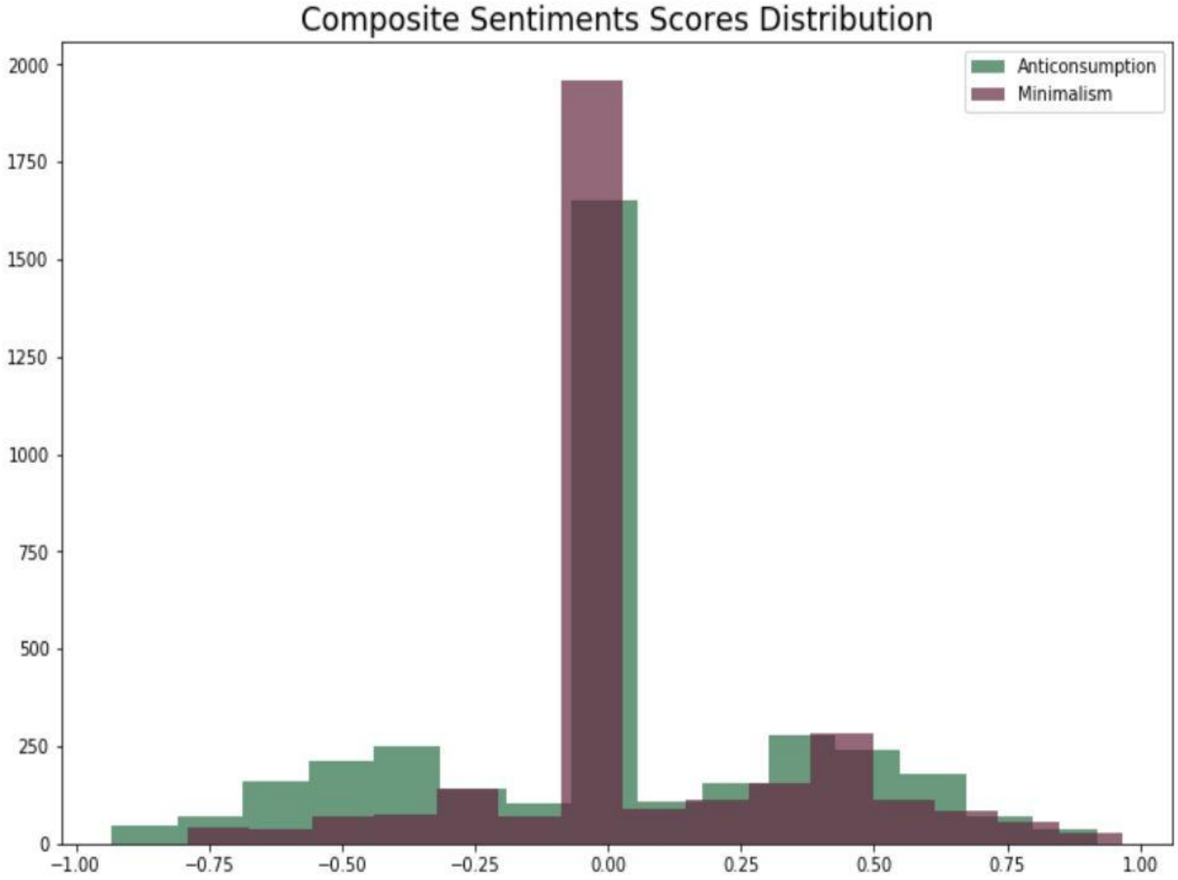


Anticonsumption
Minimalism
plastic
just
new
things
buy
like
don
people
buying
use
want
stuff
time
make
think
waste
need
items
year
climate
help
money
room
work
thread
rid
really
feel
change
amazon
0
50
100
150
200
250
300
0
20
40
60
80
100
120

Upvotes


Upvotes Count Distribution
Anticonsumption
Minimalism
3000
2500
2000
1500
1000
500
0
0
1000
2000
3000
4000
5000
6000
7000

Sentimental Analysis



- As composite sentiment score increases by 1, the post is about .45 times (%45) more likely to belong to r/minimalism subreddit.
- Minimalists are relatively more optimistic.
- Helping the environment is harder than helping yourself.

Preprocessing (NLP)

✓ Removed English stop words from documents

✓ Removed custom stop words such as subreddit name, common words etc.

✓ Removed HTML tags

✓ Removed non-letter characters

× Lemmatized words with WordNet (reduced the accuracy score)

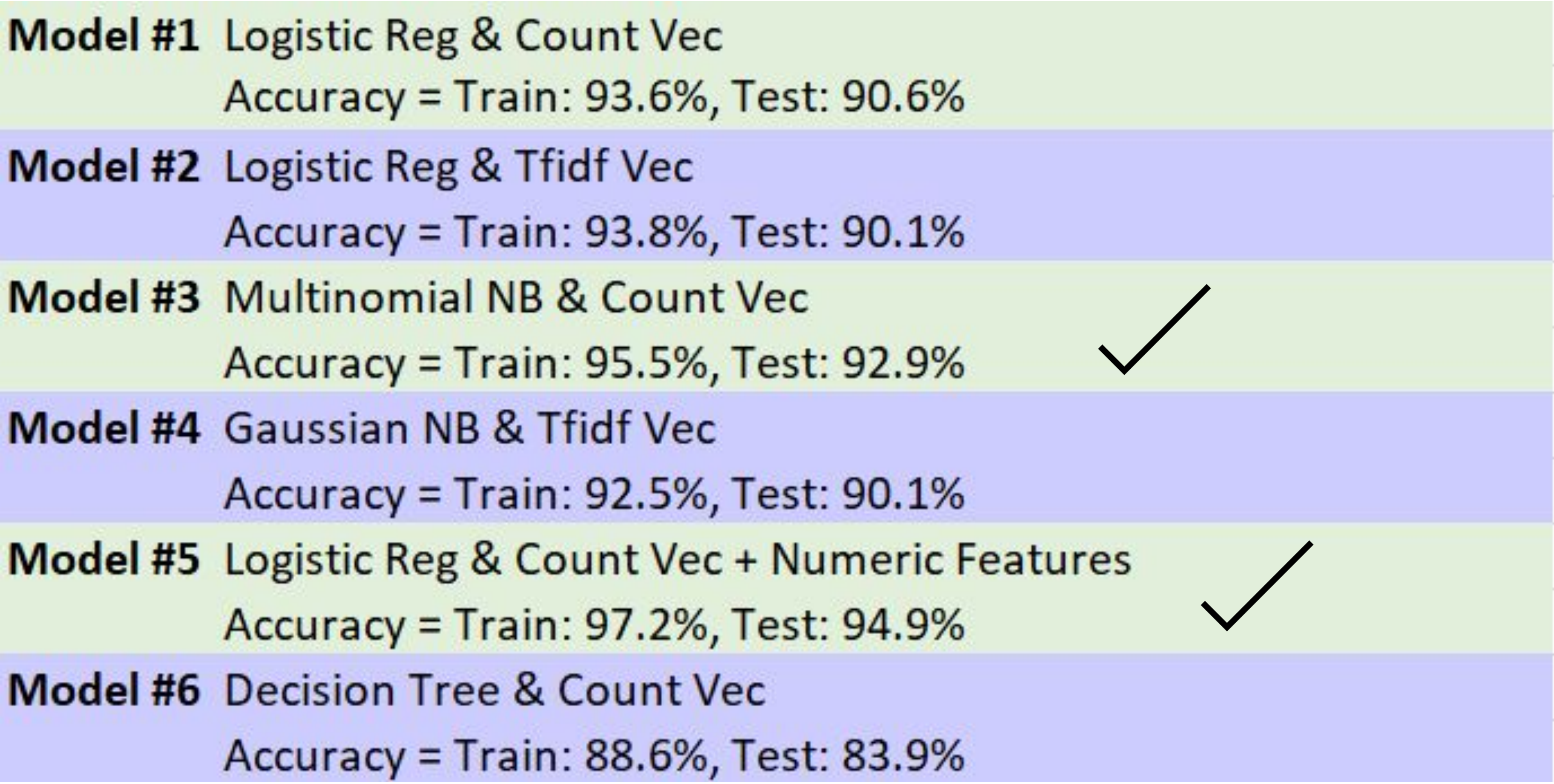
Classification Modeling

Try numerous combination of techniques until you reach the best prediction accuracy on unseen data

How?

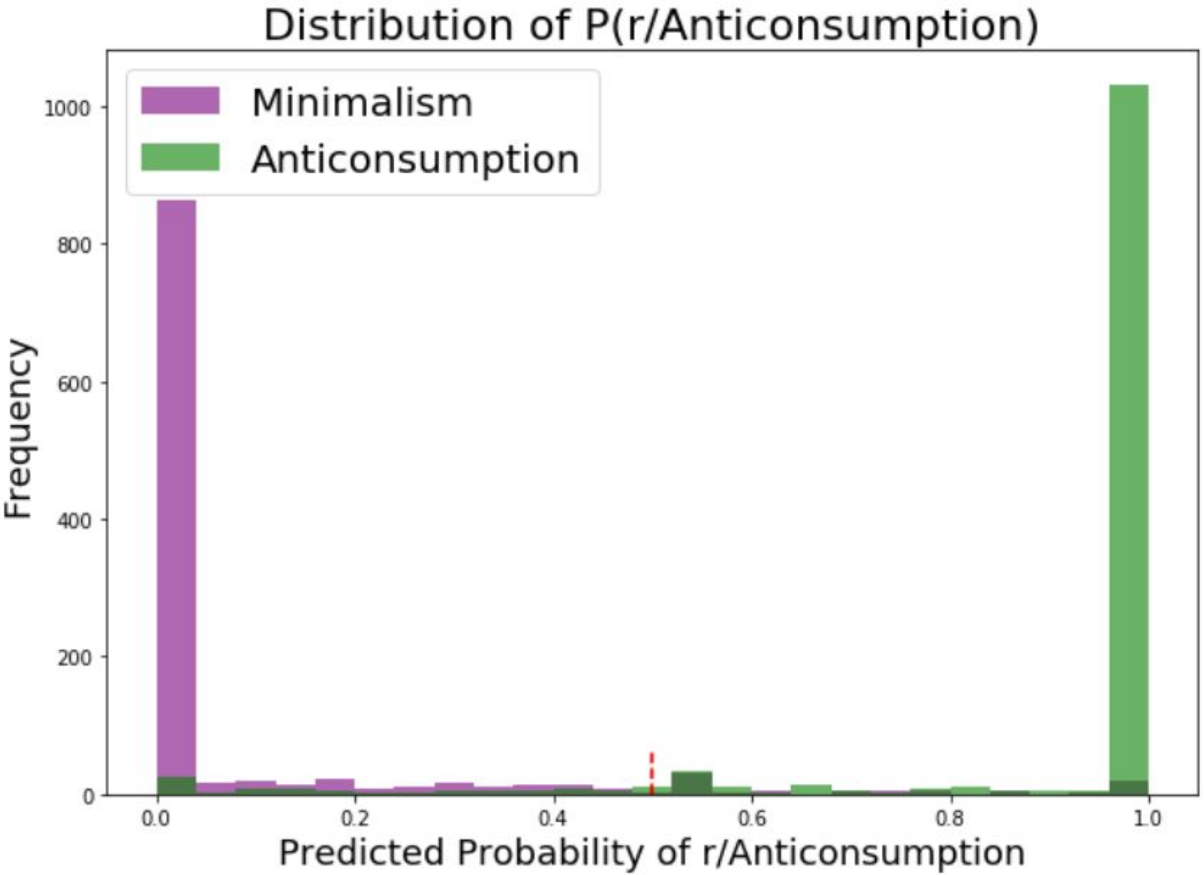
- Estimators: Logistic Regression, Multinomial NB, Gaussian NB, Decision Tree, and Random Forest
- Transformers: CountVectorizer, TfidfVectorizer
- Hyperparameters
- Gridsearch
- Pipeline

Accuracy Scores (Baseline: 52.7%)

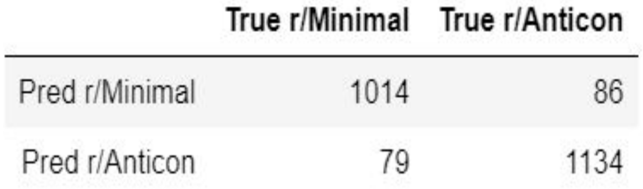


Model #1	Logistic Reg & Count Vec Accuracy = Train: 93.6%, Test: 90.6%	
Model #2	Logistic Reg & Tfidf Vec Accuracy = Train: 93.8%, Test: 90.1%	
Model #3	Multinomial NB & Count Vec Accuracy = Train: 95.5%, Test: 92.9%	✓
Model #4	Gaussian NB & Tfidf Vec Accuracy = Train: 92.5%, Test: 90.1%	
Model #5	Logistic Reg & Count Vec + Numeric Features Accuracy = Train: 97.2%, Test: 94.9%	✓
Model #6	Decision Tree & Count Vec Accuracy = Train: 88.6%, Test: 83.9%	

Predictions vs True Values

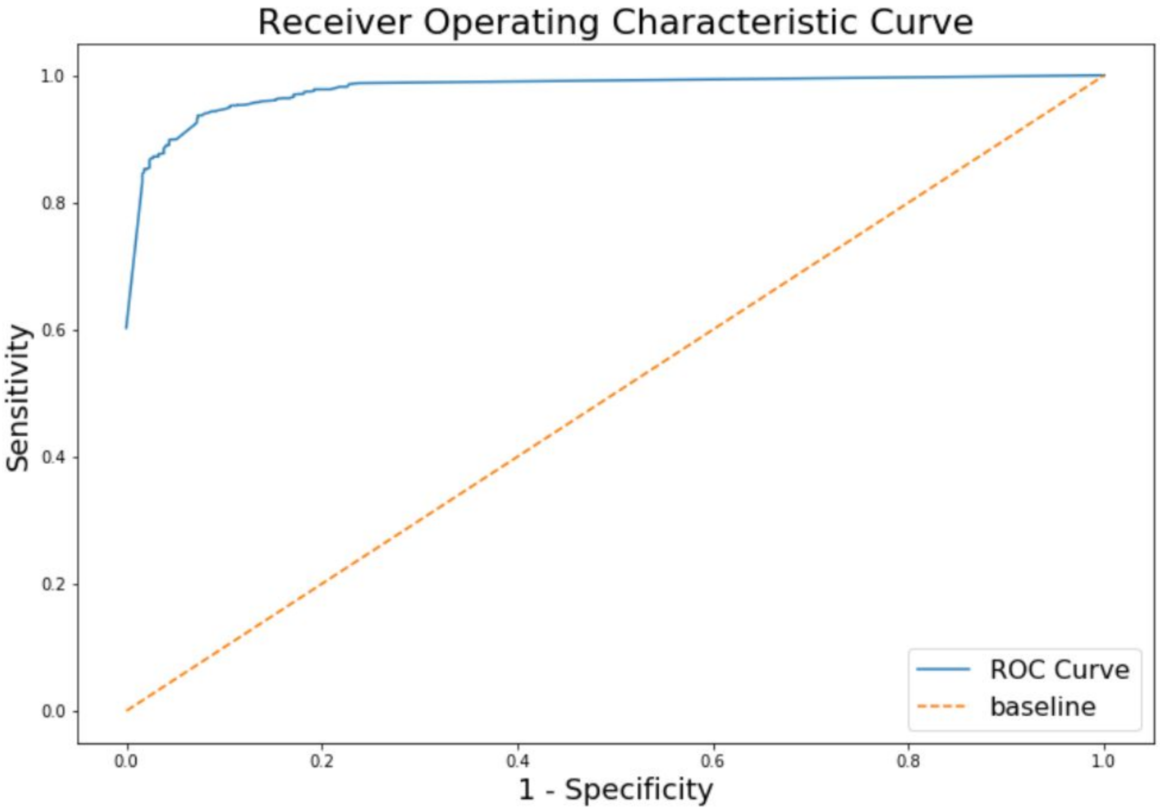


- Not much overlapping
- Errors (FP, FN) are balanced



	True r/Minimal	True r/Anticon
Pred r/Minimal	1014	86
Pred r/Anticon	79	1134

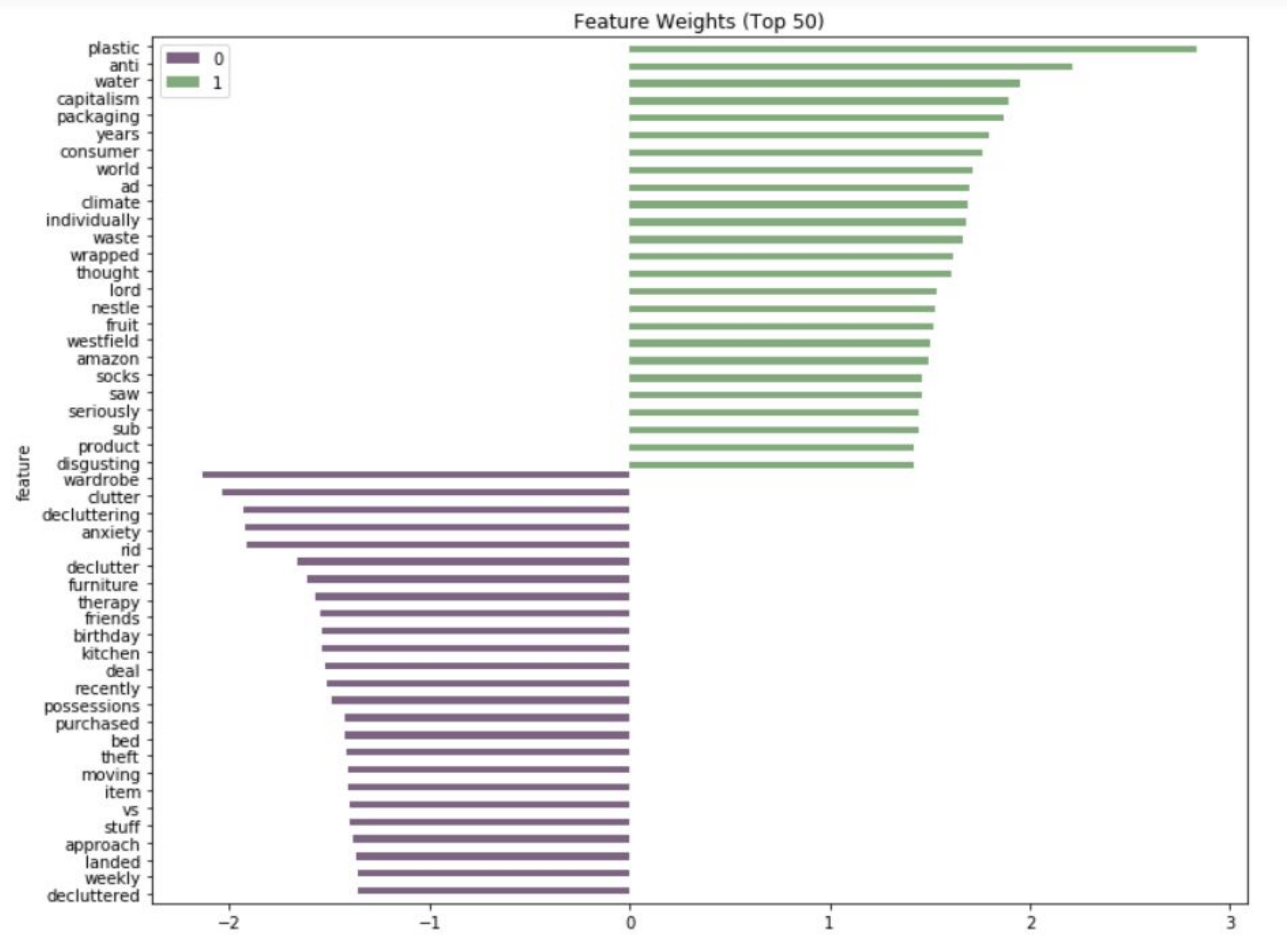
ROC Curve



Area under the curve represents accurate predictions

Weight of Words

0 = r/Minimalism
1 = r/Anticonsumption



Summary

- People’s language, sentiment, and behaviour from textual content can be classified by choosing right techniques

- As the model learns from new data, its performance improves

- Top words can be good resource for environmental campaigns to understand people’s needs and tendencies

Recommendations

- ❏ Collect more data
- ❏ Spend more time with current features
- ❏ Engineer new features such as comments, upvotes ratio etc.
- ❏ Research and implement additional lemmatizing techniques

Thank you for listening!

Any questions?

r/AskReddit

Sources

https://www.reddit.com/r/Anticonsumption/

https://www.reddit.com/r/minimalism/

https://www.reddit.com/