Save money, or save the earth?

Differentiating between frugal folks and environmentalist

Background

Increased awareness of climate change and its urgency has motivated governments and corporations to adopt greener policies. (e.g. the Straw-Free movement*)

Riding on this wave, there is a growing market for *green* products such as shopping bags, reusable/glass straws, clothes made from recycled waste.

*https://www.nationalgeographic.com/environment/article/news-plastic-drinking-straw-history-ban

Problem Statement

You are an analyst at a social media company which uses targeted advertising on its platform. You have been tasked to create a classification model to identify users who are likely to been keen on green products based on their text based interactions.

Strategy

Scrape data from two subreddits to train the classifier

ZeroWaste

where people discuss how to reduce environmental impact through green ideas and minimizing waste

Frugal

where people discuss how to conserve time, money, resources

Reduce waste, Recycle, Reuse, Save

Strategy

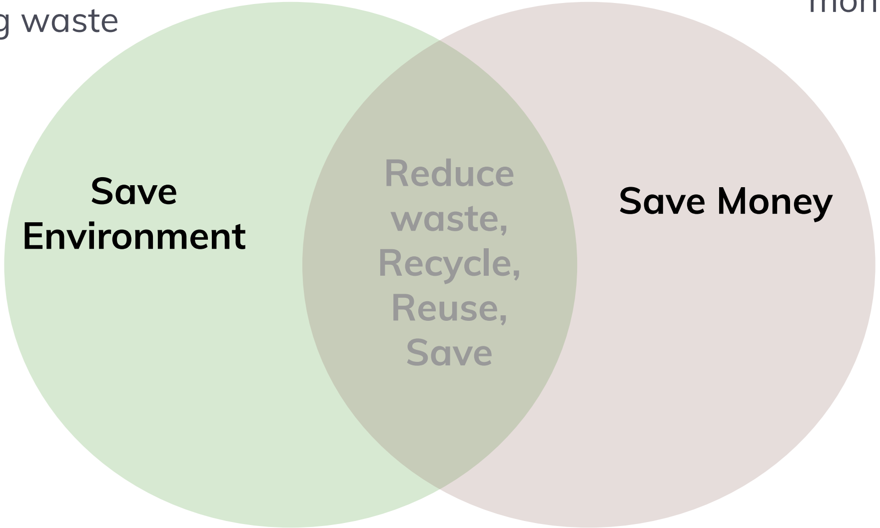
Scrape data from two subreddits to train the classifier

ZeroWaste

where people discuss how to reduce environmental impact through green ideas and minimizing waste

Frugal

where people discuss how to conserve time, money, resources



Problem Statement

Goal:

ACCURATELY distinguish green from frugal.

Sensitivity and Overall Accuracy are the primary metrics of success.

Contents

Data Preparation

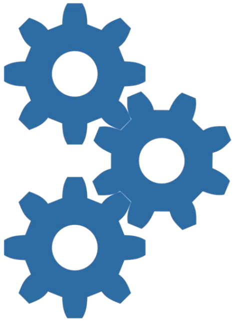
Data Extraction

10,000 text-based reddit posts for each subreddit were scraped using the pushshift API

Data Cleaned

Removed posts that were

- [removed]
- [deleted]
- have no description
- spam(e.g. “Testing”, “andfgasg”)



Exploratory Data Analysis

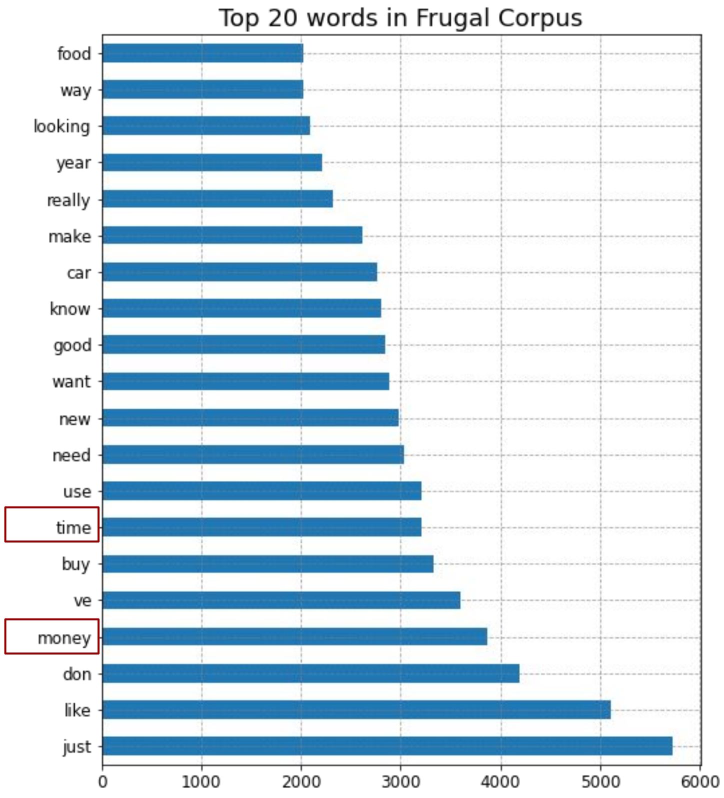
Unigrams (Top 20)

- (Frugal, ZeroWaste) x (CountVectorizer, TfidfVectorizer)

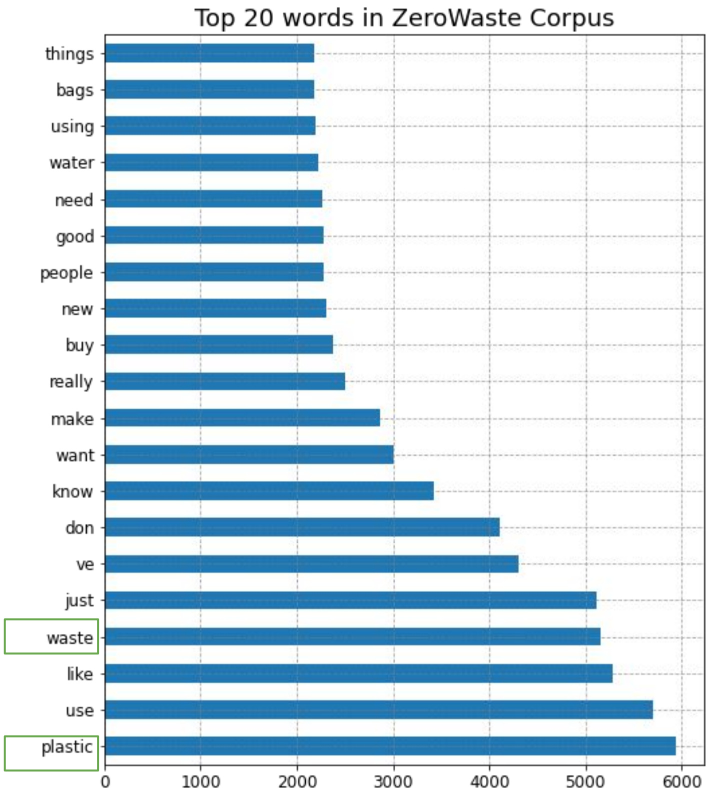
Bigrams (Top 20)

- (Frugal, ZeroWaste) x (CountVectorizer, TfidfVectorizer)

Unigrams

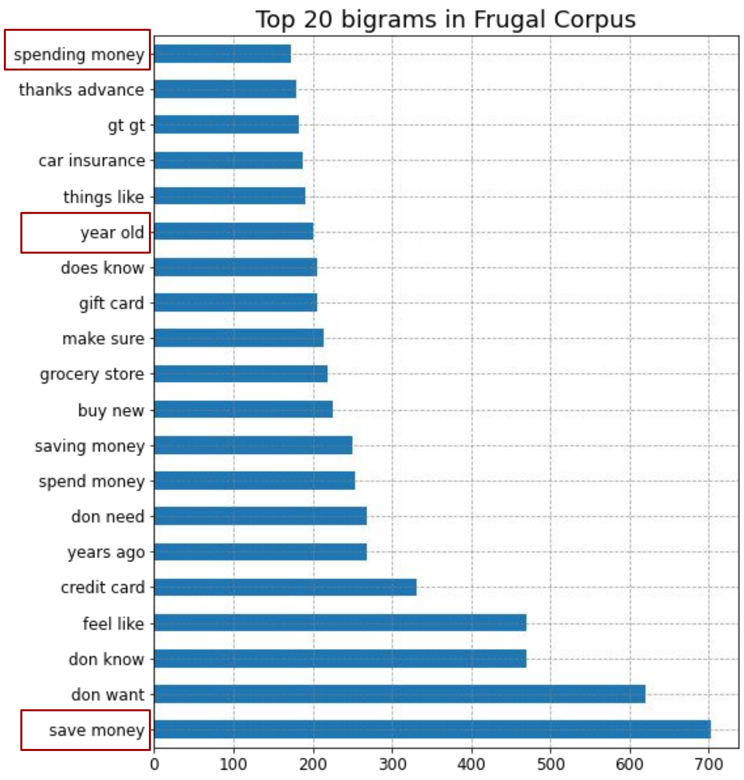


Top 20 words in Frugal Corpus
food
way
looking
year
really
make
car
know
good
want
new
need
use
time
buy
ve
money
don
like
just
0
1000
2000
3000
4000
5000
6000

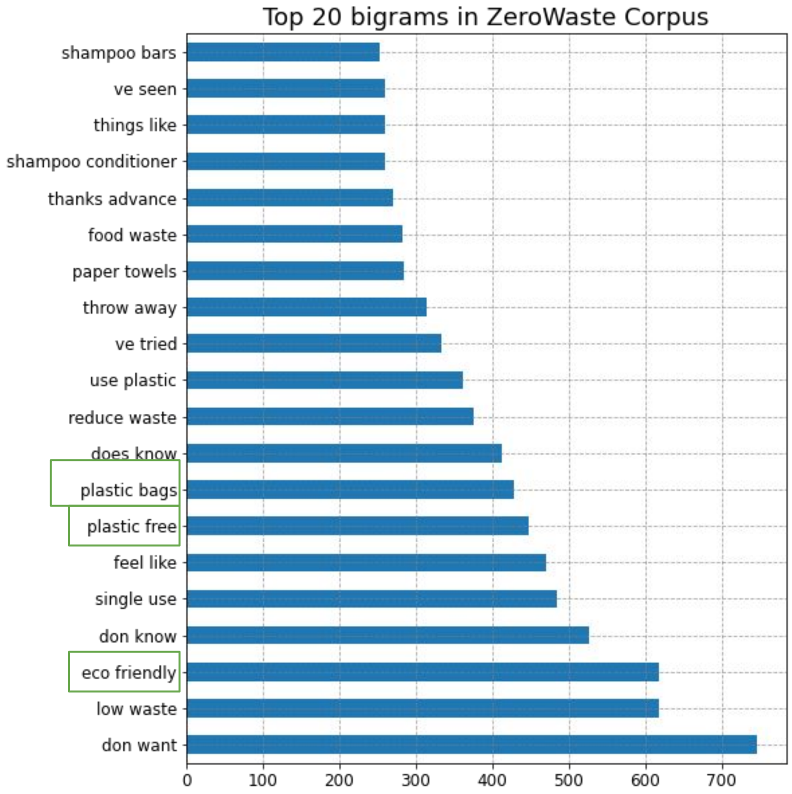


Top 20 words in ZeroWaste Corpus
things
bags
using
water
need
good
people
new
buy
really
make
want
know
don
ve
just
waste
like
use
plastic
0
1000
2000
3000
4000
5000
6000

BIGRAMS



Top 20 bigrams in Frugal Corpus
spending money
thanks advance
gt gt
car insurance
things like
year old
does know
gift card
make sure
grocery store
buy new
saving money
spend money
don need
years ago
credit card
feel like
don know
don want
save money
0
100
200
300
400
500
600
700



Top 20 bigrams in ZeroWaste Corpus
shampoo bars
ve seen
things like
shampoo conditioner
thanks advance
food waste
paper towels
throw away
ve tried
use plastic
reduce waste
does know
plastic bags
plastic free
feel like
single use
don know
eco friendly
low waste
don want
0
100
200
300
400
500
600
700

Analysis plan

Model Exploration (GridSearch)

Combinations of:

1. Model
    - K Nearest Neighbor
    - Multinomial Naive Bayes
    - Random Forest
2. Vectorizer
    - CountVectorizer
    - TfidfVectorizer
3. Hyper Parameters
    - Max_features
    - Min_df
    - Min_sample_leaf
    - etc.

To assess:

1. Accuracy
2. Sensitivity

Feature Analysis

To investigate:

1. Top performing features
2. Wrongly classified posts

Model Tuning

Stop Words

Model exploration

		Train			Test		
Model		**Accuracy**	**Sens**	**Spec**	**Accuracy**	**Sens**	**Spec**
Model 1a	KNN with CountVectorizer	0.82	0.81	0.83	0.73	0.72	0.73
Model 1b	KNN with TfidVectorizer	0.87	0.85	0.88	0.81	0.8	0.82
Model 2a	MNB with CountVectorizer	0.90	0.93	0.89	0.88	0.92	0.86
Model 2b	MNB with TfidVectorizer	0.91	0.93	0.89	0.88	0.92	0.86
Model 3a	RandomForest with CountVectorizer	0.92	0.92	0.92	0.89	0.90	0.89
Model 3b	RandomForest with TfidVectorizer	0.95	0.95	0.94	0.89	0.90	0.89

Model exploration

		Train			Test		
Model		**Accuracy**	**Sens**	**Spec**	**Accuracy**	**Sens**	**Spec**
Model 1a	KNN with CountVectorizer	0.82	0.81	0.83	0.73	0.72	0.73
Model 1b	KNN with TfidVectorizer	0.87	0.85	0.88	0.81	0.8	0.82
Model 2a	MNB with CountVectorizer	0.92	0.94	0.90	0.88	0.92	0.86
Model 2b	MNB with TfidVectorizer	0.92	0.94	0.91	0.88	0.92	0.86
Model 3a	RandomForest with CountVectorizer	0.92	0.92	0.92	0.89	0.90	0.89
Model 3b	RandomForest with TfidVectorizer	0.95	0.95	0.94	0.89	0.90	0.89

Feature analysis (top performing features in MNB and RF)

Assess top performing 120 features

Incorporate misleading features as StopWords (Set1)

MNB CVEC	MNB TVEC	RF CVEC	RF TVEC
ity	apps	product	hi
insurance	deodorant	bank	razor
shampoo bar	plastic containers	plastic bags	product
straws	phone plan	gas	bank
compostable	money	cardboard	plastic bags
car insurance	1000	15	gas
zw	conditioner	60	cardboard
reducing waste	reducing	metal	15
low waste	pill bottles	throw	60
ethique	300	monthly	metal

Feature analysis (top word count in Wrongly Classified posts)

Assess top 120 features

MNB		RF	
False Negative	False Positive	False Negative	False Positive
years	people	just	use
really	clothes	like	just
waste	way	don	like
got	food	buy	don
clothes	using	money	make
money	ideas	want	ve
going	does	new	water
year	products	ve	know
free	lot	know	buy
looking	plastic	people	used

Feature analysis (top word count in Wrongly Classified posts)

Assess top 120 features

Incorporate misleading features as StopWords (Set2)

MNB		RF	
False Negative	False Positive	False Negative	False Positive
years	people	just	use
really	clothes	like	just
waste	way	don	like
got	food	buy	don
clothes	using	money	make
money	ideas	want	ve
going	does	new	water
year	products	ve	know
free	lot	know	buy
looking	plastic	people	used

Model tuning

		Train			Test		
	Models	Accuracy	Sens	Spec	Accuracy	Sens	Sens
Model 2b	MNB original Stopwords	0.924	0.944	0.909	0.884	0.918	0.860
Model 2b_sw1	MNB Stopwords set1	0.924	0.941	0.910	0.883	0.916	0.859
Model 2b_sw2	MNB Stopwords set2	0.924	0.942	0.909	0.882	0.912	0.860
Model 3b	RF original Stopwords	0.948	0.954	0.943	0.894	0.902	0.888
Model 3b_sw1	RF Stopwords set1	0.948	0.953	0.943	0.894	0.903	0.886
Model 3b_sw2	RF Stopwords set2	0.947	0.952	0.942	0.896	0.905	0.889

Conclusions and Recommendations

The best performing model can distinguish between Green users and their close counterparts, the Frugals with an accuracy of 89%. Which means that while it may make errors in advertising 11% of the time it will likely perform much better against more *dissimilar* audiences.

Base on the current analysis, we would recommend to use the Random Forest model with stop word list 2.

Future directions

1) Combine predictions with other behaviours on the social media platform

2) Investigate wrongly classified posts in fulltext form

3) Refine list of stop words

4) Stemming and Lemmatizing

Thank you!