Text Analysis Report

Data:

We use data collected from interviews and online resources like Reddit regarding recycling as the central theme. 20 interview texts without the takeaway section, around 110 reddit posts, and text from an article about recycling from online repositories was used. An important omission is a substantial twitter dataset about recycling and sustainability which could not be used due to twitter api being no longer open source.

Data Preprocessing:

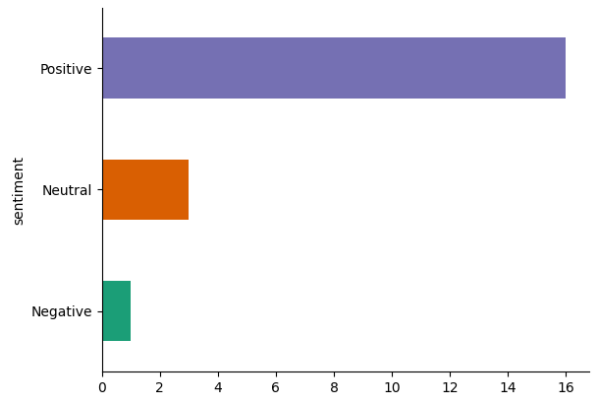
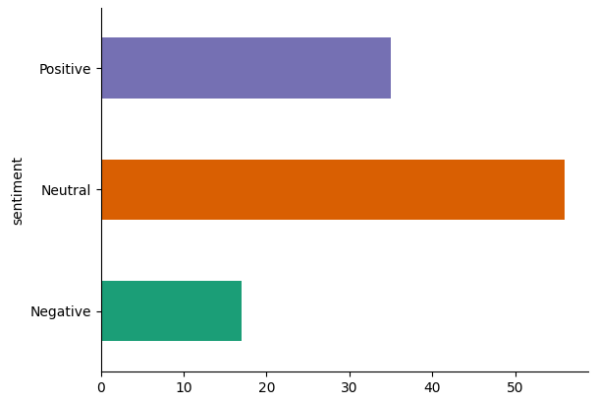
All the data collected is stored as raw text in json format. Interview data and web scrapped data are loaded individually and the following preprocessing steps are performed.

- Removal to URL, hyperlinks, emails, new line characters, quotations
- Conversion to bigram and trigrams
- Lemmatization and POS tagging
- Stop word removal

Sentiment Analysis:

After preprocessing the text data, we use ChatGPT to perform sentiment analysis. Particularly, we use the following prompt to get the sentiment and the associated text which invokes that sentiment:

"You will be provided with a text, and your task is to identify their sentiments about recycling as positive, neutral, or negative and what part of text corresponds to it."

Interview	Web
	

As it can be seen, in case of interviews, there were mostly positive sentiments about recycling and only 1 case of overall negative sentiment. This may be attributed to the fact that we were inquiring about their personal attitudes towards recycling and thus invoking a positive attitude and sense of self awareness. Whereas, in case of web data, the majority of posts have a neutral stance and a substantial amount of posts even have negative sentiment about recycling. This can be attributed to anonymity, since you can express yourself without fear of identification.

a. Positive Sentiments: The following texts from interview reflect positive statement according to ChatGPT
"because she personally thinks that recycling works and cares about the environment"
“Jay is proud of Brooklyn's community-centric approach to recycling and often takes part in local awareness campaigns.”

b. Negative Sentiments:

"kaylee says she hates taking out the trash because the trash room is disgusting and smells horrible"

c. Neutral Sentiments:

"honestly with our busy lives recycling is the least of our worries"
"she mentioned her less consistent recycling habits primarily due to confusion regarding labels and the inconvenience associated with the recycling process"

Topic Modeling:

We use gensim based Latent Dirichlet Allocation (LDA) model for topic modeling. Considering that we have restricted ourselves to the theme of recycling, and limited data points, we use 4 topics for interviews and 5 topics for web data to explore. The topics extracted by the LDA model are presented in the table below.

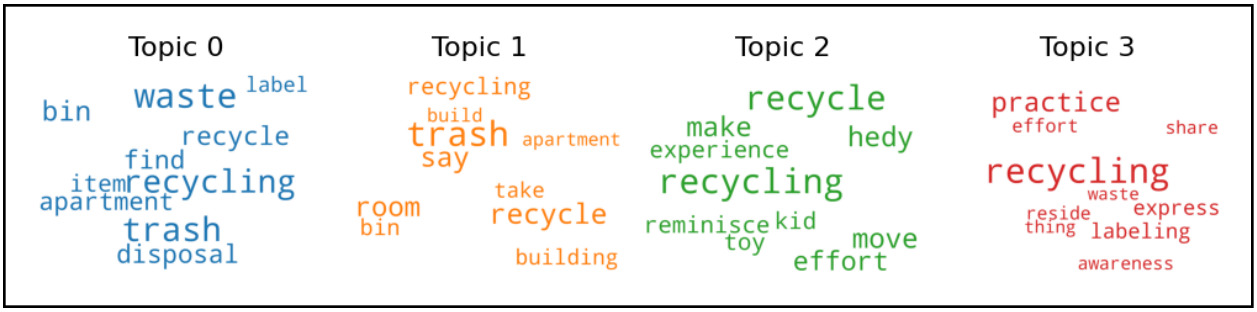
a. Identify Main Topics

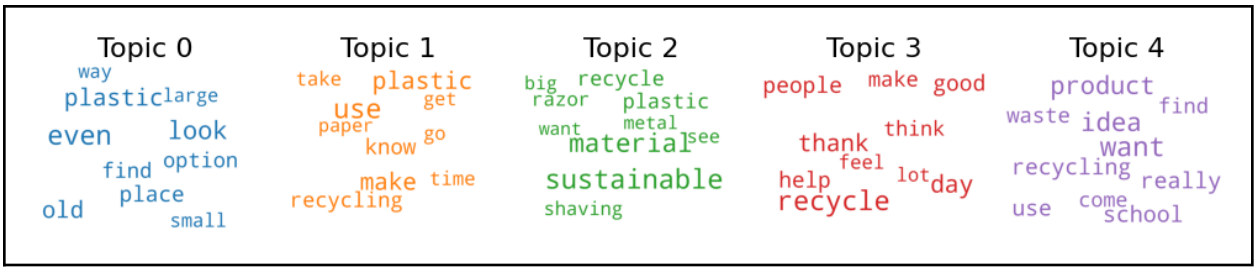
The topics and associated keywords along with the representative sentence (the sentence where contribution of the topic is maximum) are presented in the table. We can see topics like waste disposal bins, trash room in apartment, awareness effort, sustainable materials, plastic recycling.

	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0	0.9932	waste, trash, recycling, bin, recycle, disposal, find, apartment, item, label	[dispose, trash, use, designate, bin, landfill, recycle, compost, find, apartment, building, sid...
1	1	0.9946	trash, recycle, say, room, recycling, bin, building, take, build, apartment	[recent, graduate, live, year, partner, work, weekend, lot, cleaning, leave, fact, morning, inte...
2	2	0.9906	recycling, recycle, move, hedy, make, effort, kid, toy, experience, reminisce	[move, blend, cultural, experience, influence, perspective, recycle, chat, speak, journey, under...
3	3	0.9893	recycling, practice, labeling, express, effort, reside, awareness, share, thing, waste	[reside, live, year, witness, change, waste, management, practice, recently, family, day, commun...

	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0	0.9329	even, look, plastic, old, place, option, find, small, way, large	[look, way, tell, unable, blade, attach, also, unable, remove, make, plastic]
1	1	0.9851	use, plastic, make, know, recycling, time, take, go, get, paper	[partner, compost, long, time, recently, move, town, drop, point, available, membership, local, ...
2	2	0.9381	sustainable, material, recycle, plastic, razor, shaving, see, metal, big, want	[energy, require, recycle, limit, amount, want, metal, melt, huge, amount, energy, waste]
3	3	0.9464	recycle, day, thank, people, good, help, make, think, feel, lot	[experience, improved, recycle, make, efficient, stop, recycle, home, public, research, conduct,...
4	4	0.8850	want, idea, product, really, school, use, recycling, waste, find, come	[keen, country, specific, vape, recycling, program]

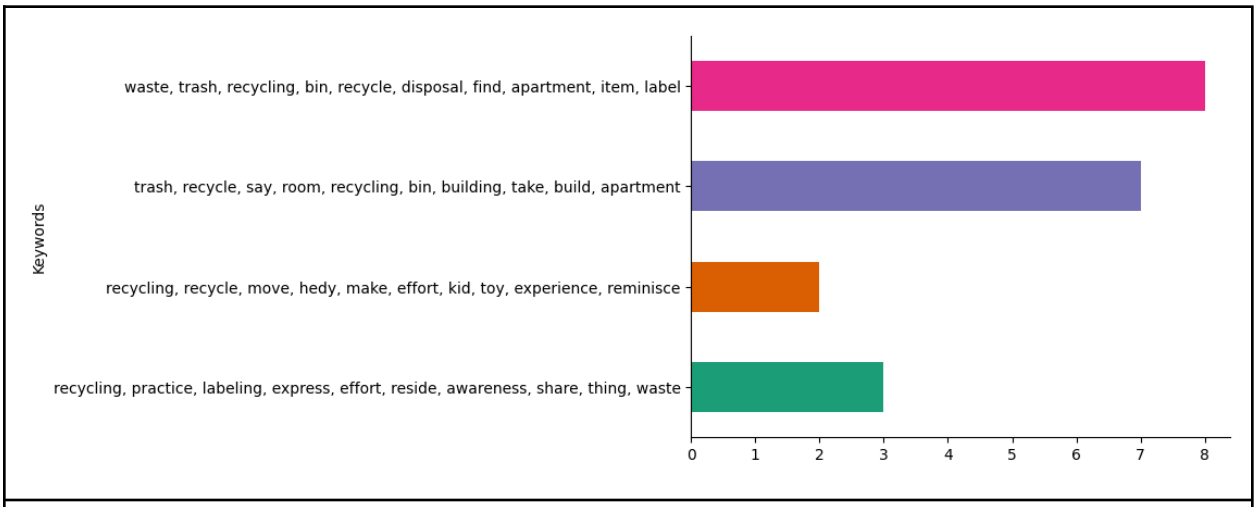
Along with that, the word cloud corresponding to the keywords for each topic is presented below.

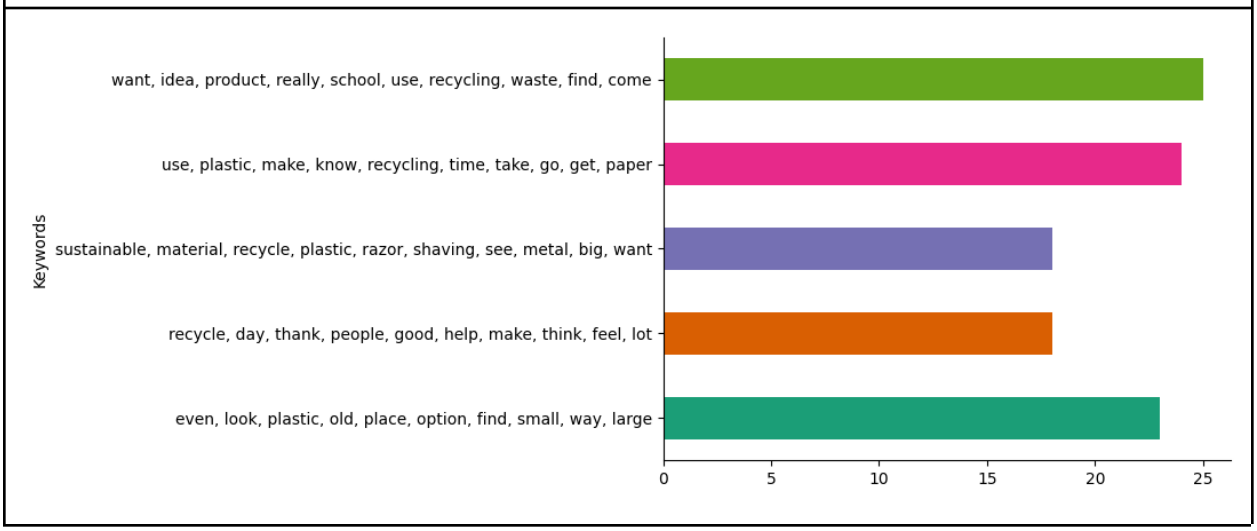




b. Topic Prevalence

We present the distribution of topics in the interview and web data as histograms. As it can be seen below, topics like waste disposal, trash rooms in apartments are discussed more while topics like awareness efforts, sustainable materials are discussed less.

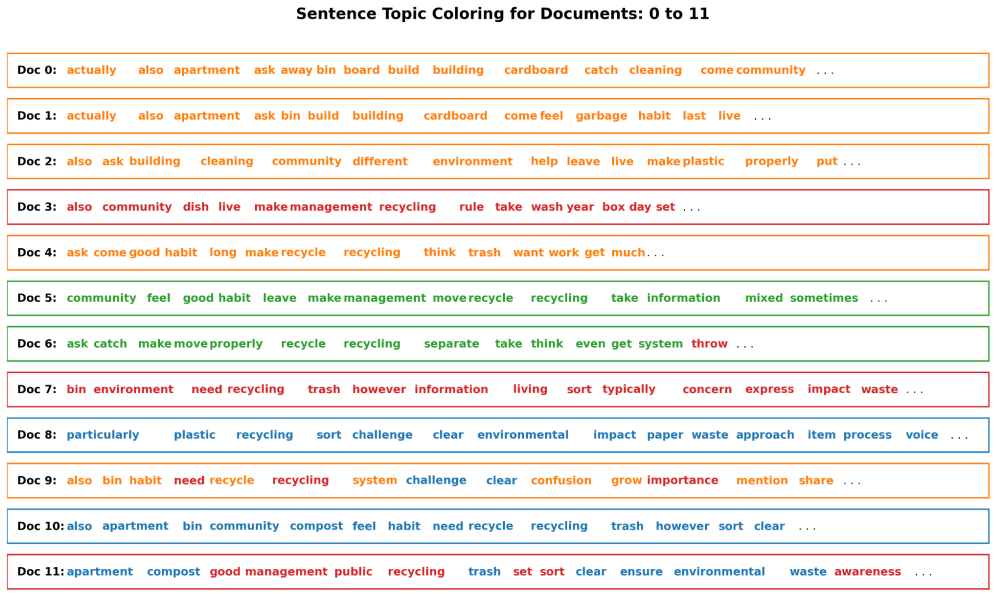


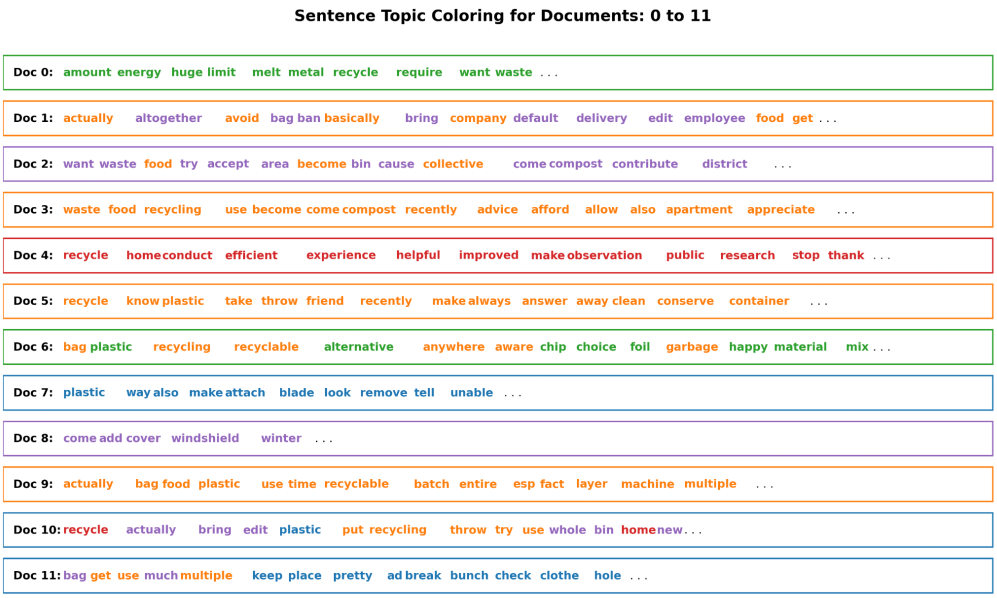


c. Topic Relationships

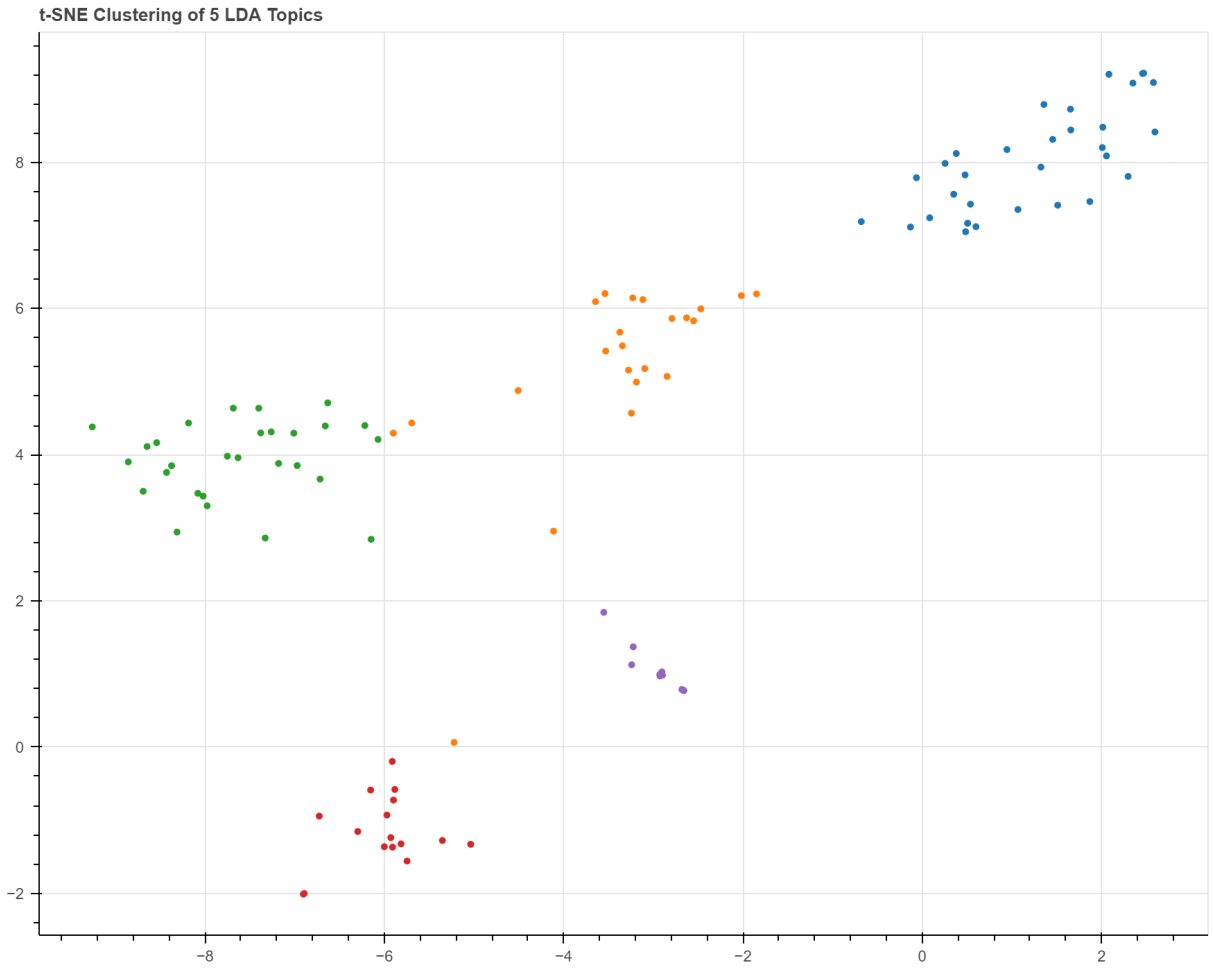
In order to identify relationships among the topics, we use two approaches:

1) Sentence topic coloring: In this approach, we color the entire sentence in its representative topic and color each word in the sentence according to the topic which it most likely belongs. This gives us a qualitative understanding of how topics are distributed among the sentences or document texts. As it can be seen, certain sentences are pure i.e. it mostly contains all words from same topic, whereas other sentences have a mix of different topic words.



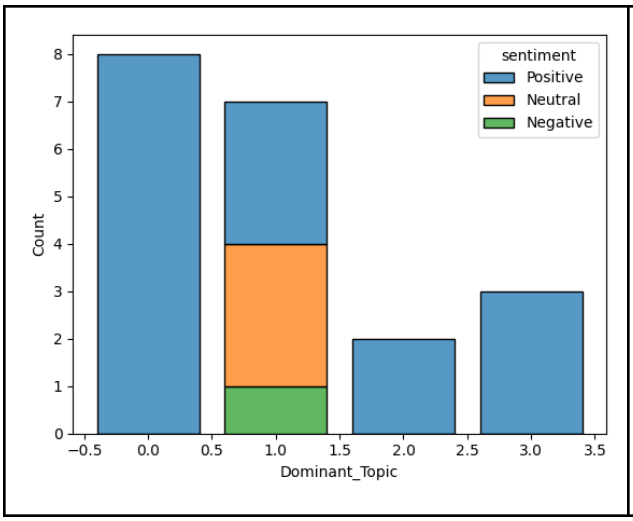


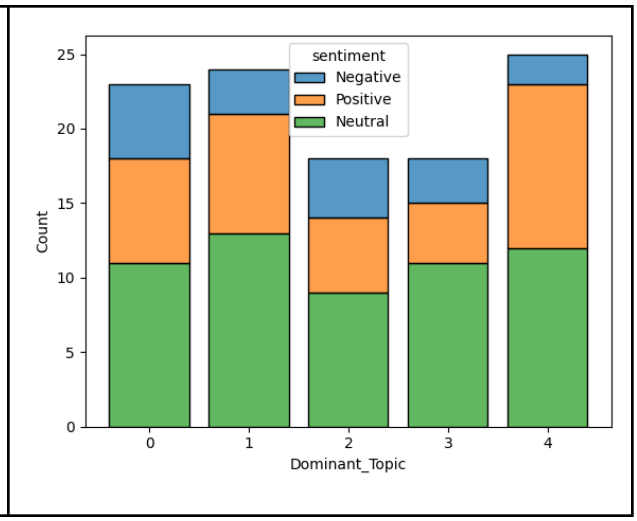
2) t-SNE plot of documents: We use t-SNE plot to cluster documents according to their topics to identify how varied the documents are in terms of topic separation. As it can be seen, certain topics are well separated, while others are marginally related.

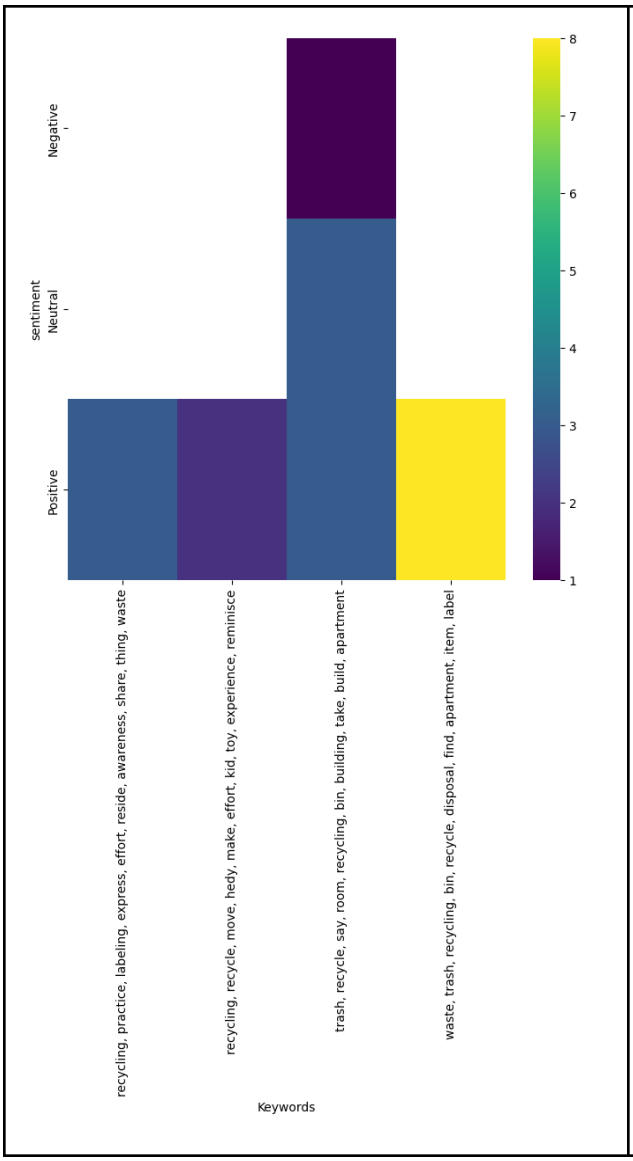


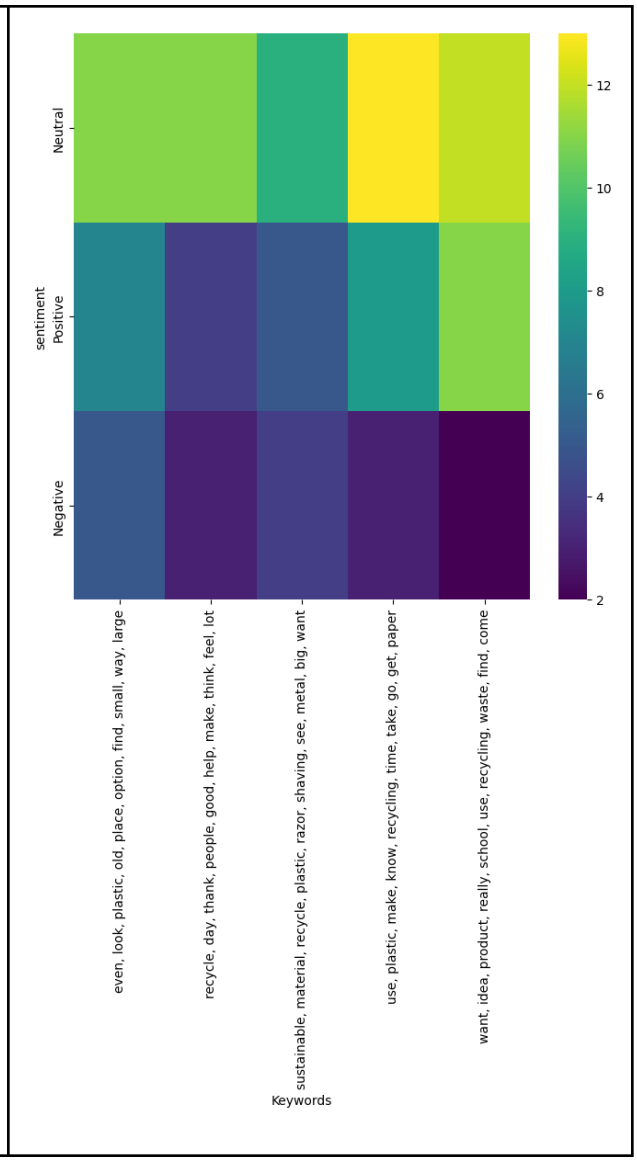
Cross-Referencing Sentiment and Topics

In order to analyze the distribution of sentiments across topics, we plot stacked histograms which are presented below. It can be seen that in case of interviews, only one topic had diversity in terms of sentiments. The topic related to trash rooms in the apartment had mixed sentiments, while other topics had only positive sentiments associated with them.
But in case of web data, the sentiments are equally distributed among the topics and each topic has a fair representation of positive, negative and neutral sentiments.
Furthermore, we also present correlation plots of topics and sentiments to get a more granular understanding of keywords and sentiments.


sentiment
Positive
Neutral
Negative
Count
Dominant_Topic


sentiment
Negative
Positive
Neutral
Count
Dominant_Topic


Negative
Neutral
Positive
sentiment
recycling, practice, labeling, express, effort, reside, awareness, share, thing, waste
recycling, recycle, move, hedy, make, effort, kid, toy, experience, reminisce
trash, recycle, say, room, recycling, bin, building, take, build, apartment
waste, trash, recycling, bin, recycle, disposal, find, apartment, item, label
Keywords


Neutral
Positive
Negative
sentiment
even, look, plastic, old, place, option, find, small, way, large
recycle, day, thank, people, good, help, make, think, feel, lot
sustainable, material, recycle, plastic, razor, shaving, see, metal, big, want
use, plastic, make, know, recycling, time, take, go, get, paper
want, idea, product, really, school, use, recycling, waste, find, come
Keywords

Limitation:

One limitation of our study is the interview data we had to work with. Because each of us conducted our own interviews, they were all of different lengths and not necessarily focused on the same topics. Additionally, because we didn't record our interviews word for word and instead summarized them, the language we used doesn't necessarily depict what the interviewees actually said, but rather how we chose to recount the interview. Therefore, our own biases may have affected our results. Additionally, we don't have a lot of data, so the trends we see are not strong. Furthermore, the data we collected was limited in quantity and scope, as public comments on recycling are meager on sites like Reddit. And while we did have twitter ids for data on recycling, due to the paid nature of twitter api, we couldn't collect that data. In order to improve our study, access to data from active forums like twitter along with broadening the scope could lead to substantial increase in depth of the insights.