Executive Summary

This Project uses a total set of 111,219 records (54,606 Honda comments and 54,013 Tesla comments) gathered from Reddit in order to compare how people feel about the car manufacturers in their prospective communities by conducting a sentiment analysis.

Main points:

- Both subreddits have concentrations of sentiment scores very close to zero.

- Both subreddits have around 2,000 more records with positive comments than neutral comments and positive comments more than double negative comments. They follow a similar distribution of sentiment categories.

- Honda's average sentiment score is around 30% higher than Tesla's sentiment score.

- there is a significant difference in sentiment score between the Tesla subreddit and the Honda subreddit in 2022. This Suggests that people overall post more positive things in the Honda subreddit.

- One of Tesla's most common words in negative comments is "fsd" which stands for the full self driving feature. This could indicate that people may be unhappy with this feature. Although this appears in the 100 common word word chart for positive and neutral sentiments as well so this may not indicate anything abnormal.

- One of Hondas's most common words in negative comments is "oil". This could indicate that people may be unhappy with something to do with oil related car processes. Although this appears in the 100 common word word chart for positive and neutral sentiments as well so this may not indicate anything abnormal.

- Common words may not show any insightful patterns because the concentrations of sentiment scores are very close to zero.

- Both subreddits have an average sentiment score that was higher in pre-covid years and became lower in post covid years. It might be possible that the pandemic had an effect on overall sentiment.

Next Steps

The next steps following this analysis would be to gather more data greater than what was used here, develop a custom made sentiment analysis algorithm designed specifically for subreddit comments, and compare Tesla's subreddit to other manufacturer's subreddit communities than just Honda. Further, other subreddit communities unrelated to automobile manufacturers should be looked at see if average sentiment also drops in post covid comments compared to pre covid comments.

Introduction

In only a few years Tesla's stock rose from being very small to the highest valued automobile manufacturer in the world and it has developed a huge following lately. At the same time, many might say that the company is overvalued and does not compare to current longstanding manufacturers. One way to see how a business may be doing is to see what people feel about it, especially in communities surrounding the business. In this project I wanted to compare sentiments of comments in an online tesla community and compare them to another automobile manufacturer online community. I decided to use Honda as the other automobile manufacturer because it has a loyal fan base with some of the most reliable cars in the business.

In order to examine this, a sentiment analysis on comments data gathered from 2 subreddits, a tesla community and a Honda community was done to gauge general feelings of redditors commenting in these communities. Data from 2018 to 2022 was gathered but, only the 2022 data is used for most objects in this analysis because the focus here is to test and compare recent sentiments in these communities.

I believed that we would see higher sentiment scores for comments in the Tesla subreddit than the Honda subreddit because of all the hype and attention Tesla has gotten these past 2 years.

Data Gathering

Reddit is vast website with millions of users, and numerous communities within. There is a ton of text data to be found here and luckily python has 2 APIs to do this. I chose to use Pushshift API because this API gives access to historical data and does not have any foreseeable limits on

how much data can be gathered in one time. Praw only allows for 900 comments at a time and the user does not have any control to access historical data.

For this analysis I wanted to gather data from 2018 – 2022 to compare the years with each other. In order to make sure they would be all from relatively the same time of year, each year contains comments dated between January and April 25th. 11,500 comments were gathered for each year in each subreddit, giving room for dropping blank rows during pre-processing to reach the target of 10,000 records for each year.

Data Cleaning and Pre-processing

Comments data from reddit can be very messy. There is a lot of noise that needs to be removed such as punctuation, emojis, and contractions. Plus, the words need to be put in a simpler form in order to give better results in a sentiment analysis. This is where lemmatization comes in. Lematization groups together inflected forms of a word so they can be analyzed as single item. For example: *ate* and *eating* would both be changed to *eat*.

To achieve this, I created a cleaning function in python that first changes all text to lower case, then it uses regular expressions to target unwanted characters in the data and substitute them with nothing, thereby deleting them from each comment. This was used to get rid of punctuation, website urls, newlines, the apostrophe for contractions, and using an encoding function to ignore anything that is not ascii encoded to delete emojis or other unwanted text. When dealing with contractions, I chose to only delete the apostrophes and keep the words as is. This is because the apostrophes would split the words in tokenization into 2 words making the apostrophe it's own word. Stop words were also omitted by using a list comprehension in the clean function that

splits every word in a comment, checks to see if it is a stopword, and if it is not, it is joined back together with the text. If it is a stopword, it does not exist anymore. Lastly, in the function, words are split again separately and lemmatized and then joined together again. Finally, I dropped any records from the data set that were null values or had no value.

```
Tesla Record Counts
 2019    11062
2020    11010
2021    10947
2018    10736
2022    10258
Name: date, dtype: int64


Honda Record Counts
 2019    11039
2022    10960
2018    10911
2021    10849
2020    10847
Name: date, dtype: int64
```

For the final record counts, we see in figure 1 that for each year, records are above 10,000 and the number of records between subreddits remains close with differences all being less than 1,000 records.

	year	tesla_comments
0	2018	Bear with me while I finish the article
1	2018	It really angers me to no end that people have...
2	2018	Do you have an Android phone? For whatever rea...
3	2018	My cars covered with bugs at the moment hiding...
4	2018	To charge at 48A you need a 60A breaker.
...	...	...
57450	2022	Thank you. The lawyer is going to be expensive...
57451	2022	I mean, I *guess*... better than just not gett...
57452	2022	This is an apple move tbh
57453	2022	>Yeah but that isn't a moat, just a lead. A...
57454	2022	BD DOES NOT use AI!! How many times do we need...

57455 rows × 2 columns

	year	tesla_comments
0	2018	bear finish article
1	2018	really anger end people audacity pull youre re...
2	2018	android phone whatever reason seems iphones re...
3	2018	car covered bug moment hiding nasty chip winds...
4	2018	charge need breaker
...	...	...
57450	2022	thank lawyer going expensive feel obligated fi...
57451	2022	mean guess better getting one warning
57452	2022	apple move tbh
57453	2022	gtyeah isnt moat lead moat implies preventing ...
57454	2022	bd use ai many time need say educare bit prepr...

54013 rows × 2 columns

Figure 1 (Above) shows the all the record year counts after cleaning the data. Figure 2 (Left) shows an example of comments before cleaning and Figure 3 (Right) shows those same comments after cleaning.

Exploratory Data Analysis

To get started before any sentiment analysis we look at generally the 25 most common words in each subreddit overall within our data sets. This shows us the 25 most common words from 2018 -2022.

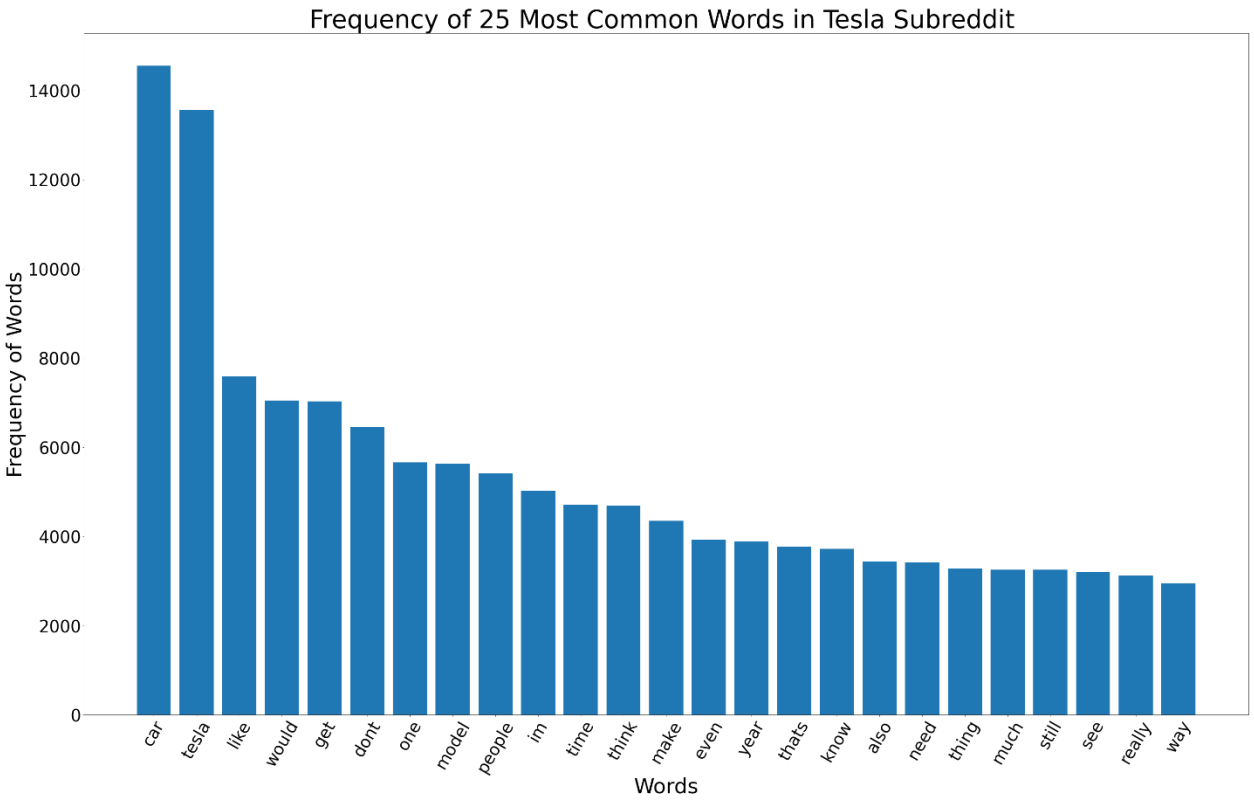


Figure 4 shows a bar graph of the 25 most common words in general for all 2018 – 2022 Tesla subreddit comments.

We see here that car is the most common word with just over 14,000 occurrences in the data set with car coming in a close second with a little less than 14,000 occurrences. Those two words would definitely be expected to be used a lot considering this is a car manufacturer community and Tesla is the manufacturer's name. Model is also a very common car enthusiast word to say.

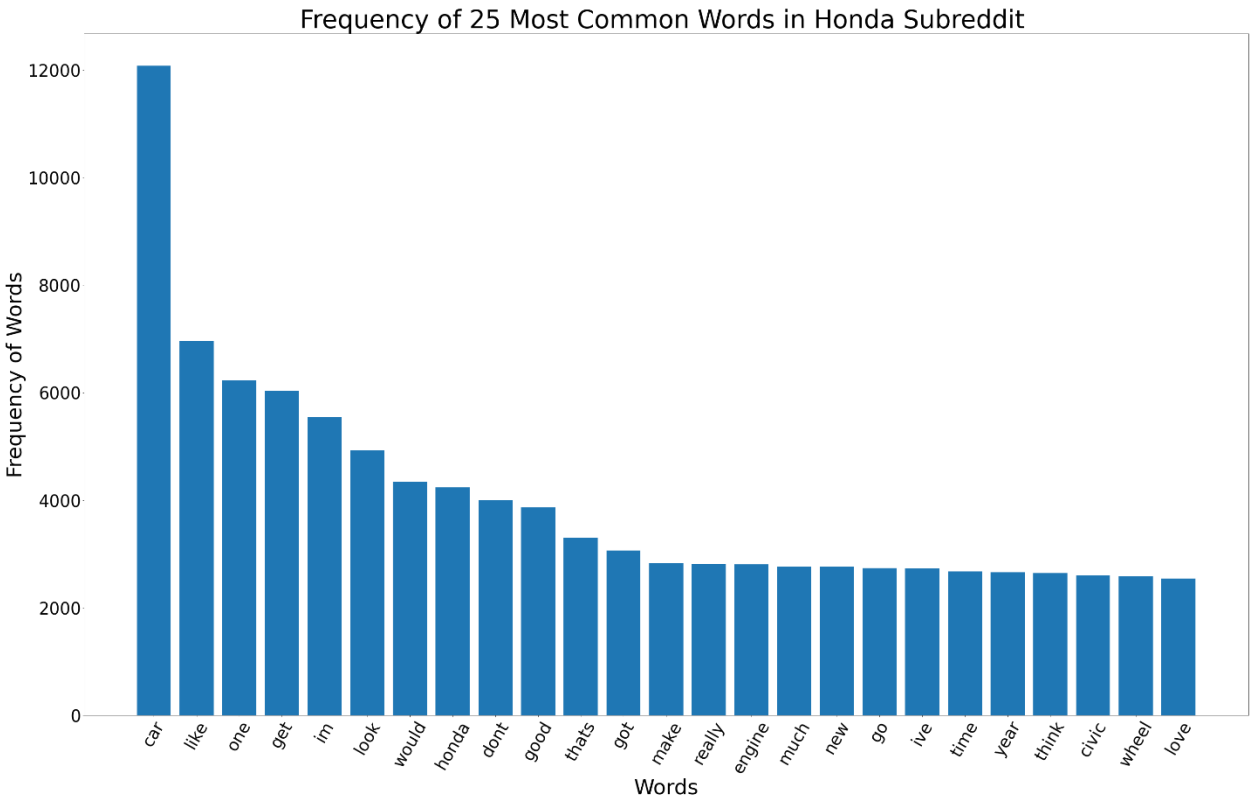


Figure 5 shows a bar graph of the 25 most common words in general for all 2018 – 2022 Honda subreddit comments.

We see here that “car” also has the most frequency in the data set. Given that this is also a car manufacturer that makes sense. “honda” is also in here, along with “civic”, which is a model by the manufacturer. There are also many words in common here with the Tesla’s most common words. Those include: “like”, “one”, “get”, “Im”, “would”, “don’t”, “year”,”think”,”time”,”much”,”that’s” and “make”. This means the tesla and honda comments both share about 50% of their 25 most common words. This probably because both of these subreddits are centered around automobile manufacturers.

Sentiment Analysis

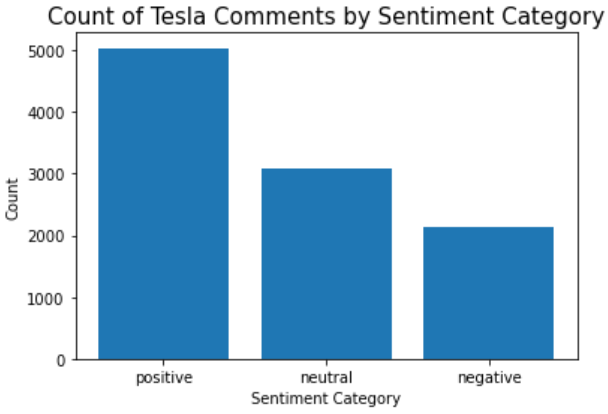
When doing sentiment analysis, there are many programs to choose from. In this project I chose to do an off the shelf, pre-trained model called textblob. The reason textblob was chosen for this is because it uses one scale of -1 (negative polarity) to 0 (neutral polarity) to 1 (positive polarity). This scale made it easy to compare sentiments with each other than other off the shelf methods. After getting the sentiment score, I categorized sentiment scores lower than 0 as negative, 0 as neutral, and higher than 0 as positive. Flair and NLTK's Vader already catgorizes the sentiment for the user and I wanted to do that part myself so I could compare them using one scale.

	year	tesla_comments	sentiment	sentiment_category
0	2018	bear finish article	0.000000	neutral
1	2018	really anger end people audacity pull youre re...	0.204762	positive
2	2018	android phone whatever reason seems iphones re...	0.000000	neutral
3	2018	car covered bug moment hiding nasty chip winds...	-0.250463	negative
4	2018	charge need breaker	0.000000	neutral
...	...	...	...	...
57450	2022	thank lawyer going expensive feel obligated fi...	-0.066667	negative
57451	2022	mean guess better getting one warning	0.093750	positive
57452	2022	apple move tbh	0.000000	neutral
57453	2022	gtyeah isnt moat lead moat implies preventing ...	0.142045	positive
57454	2022	bd use ai many time need say educare bit prepr...	0.175000	positive

54013 rows × 4 columns

First the counts of sentiment score categories are compared between recent 2022 comments of both subreddits:

```
positive    5035
neutral     3084
negative    2139
```




```
positive    5540
neutral     3521
negative    1899
```

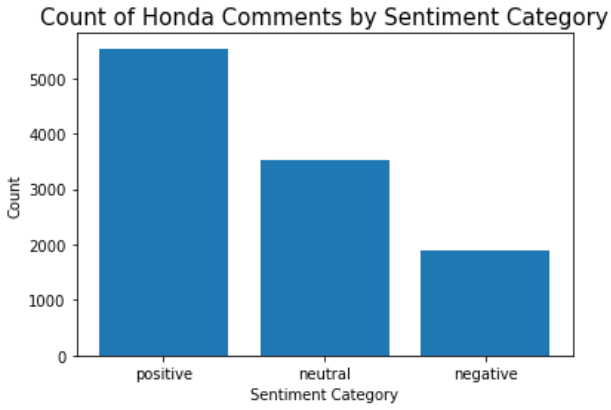


Figure 6 (Above) shows an example of what the data looks like after sentiment analysis. Figure 7 (Left) shows the count of Tesla comments by category with a bar graph and figure 8 (right) shows the same for Honda.

Both subreddits this year seem to follow a similar distribution. Both have around 2,000 more records with positive comments than neutral comments and positive comments more than double negative comments.

Next, a histogram of sentiment distrubtions is made to to see the frequeny of distrubitions across Tesla and Honda subreddits in 2022.

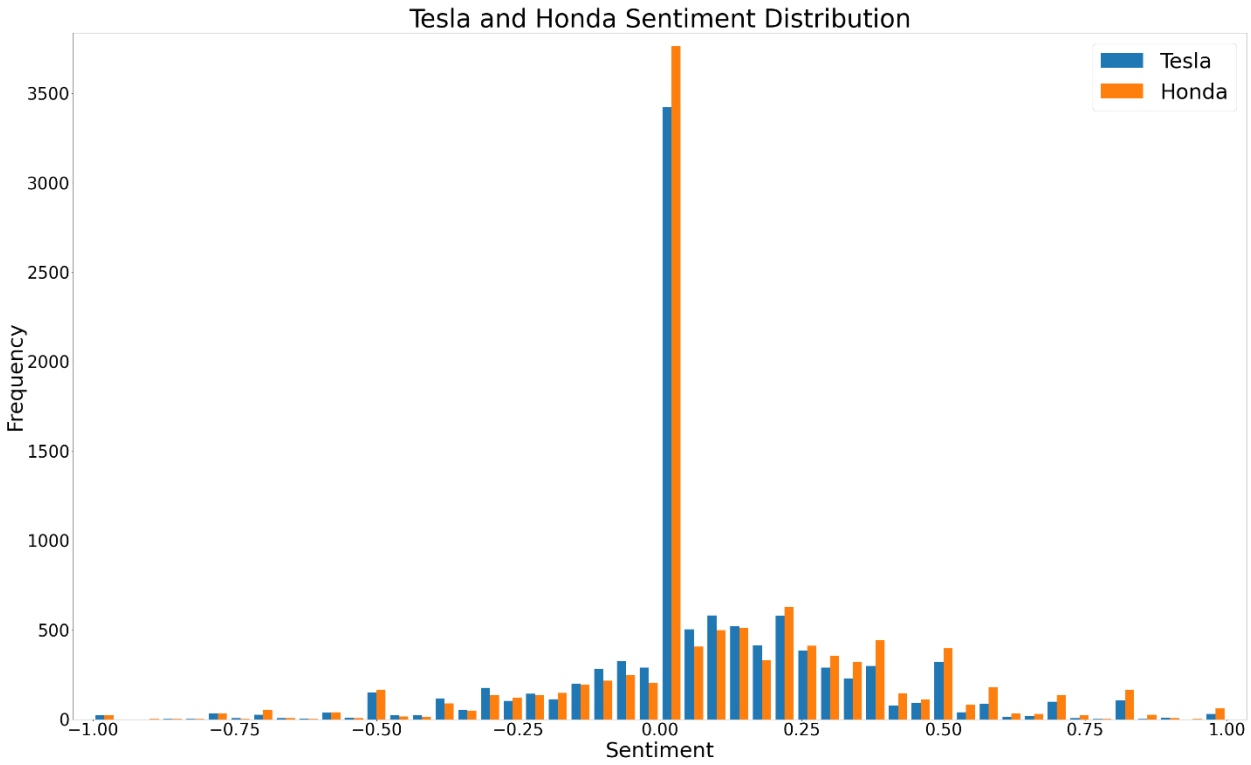


Figure 9 shows a grouped histogram of the sentiment distributions of both the Tesla and Honda subreddits for 2022.

Both subreddits have concentrations of sentiment scores close to 0 with over 3,000 records each After that, frequencies go down to a thousand and under. The farther away from 0, the less

concentrated they are. There are barely on comments with sentiment scores past .50 and -.50 frequencies.

Next the means of sentiment scores of both of these subreddits are compared for 2022.

```
Tesla's mean of sentiment scores is:  0.07716348227173293
 Honda's mean of sentiment scores is:  0.11001539351256129
```

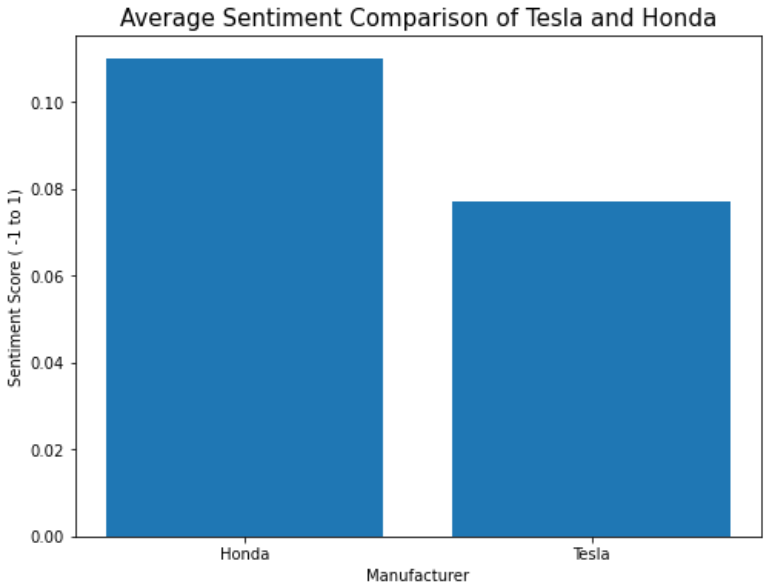


Figure 10 shows the means of sentiment scores for both subreddits in 2022 along with a bar graph of the same.

The means of Honda's and Tesla's sentiment scores are both positive but, also very close to 0. In other words, they are both only slightly positive. Honda's average sentiment score is around 30% higher than Tesla's sentiment score.

To test for statistical significance on this data I used the Mann-Whitney U test which is a non-parametric test that compares the medians of 2 independent samples. I chose to use a non-parametric test because it assumes very little about the population distribution.

```
Statistics=59778230.500, p=0.000
Different distribution (reject H0)
```

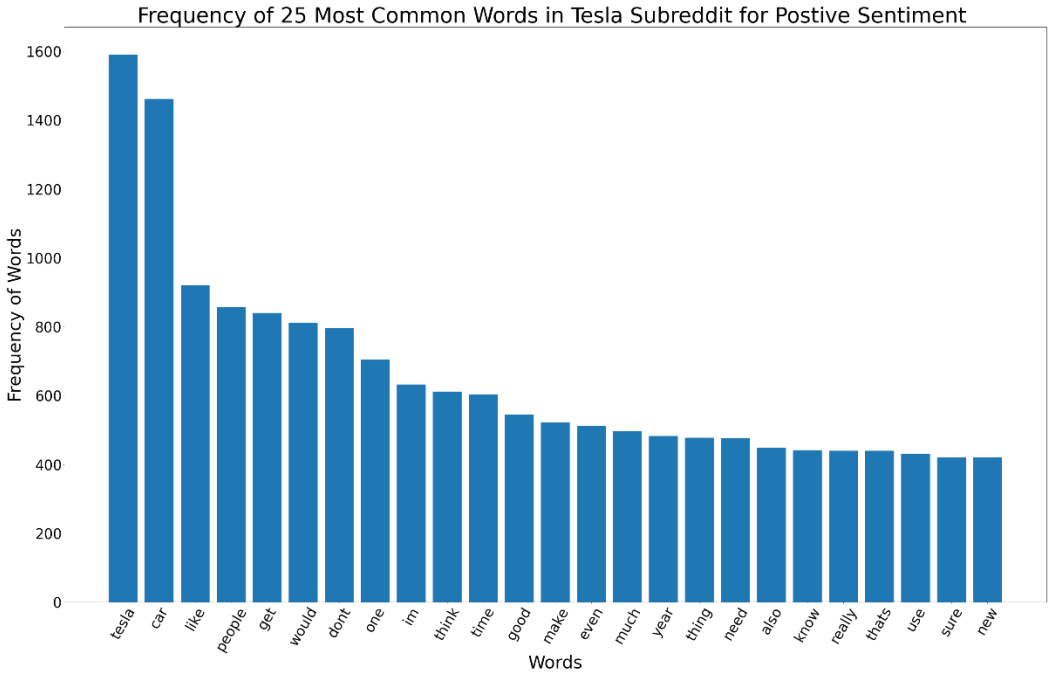
```
p= 4.742877798868163e-16
```

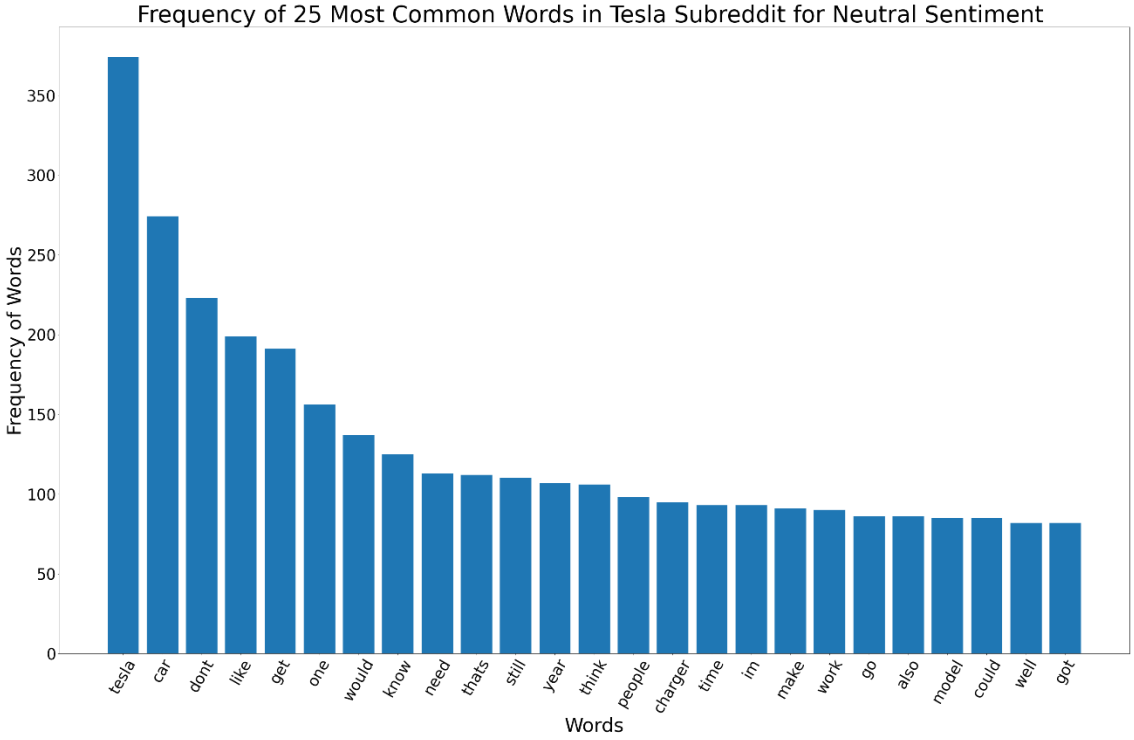
Figure 11 shows the output of the Mann-Whitney U test. In the test the p value was listed as 0 so the bottom output is the result of printing that same p value by itself in python.

We reject the null hypothesis because the p value is lower than .05. This means that there is a significant difference in sentiment score between the Tesla subreddit and the Honda subreddit in 2022. Honda has a significantly higher sentiment score. This may indicate that people comment more positive things in Honda compared to Tesla.

Common Words for Sentiment Groups

Next I examined what the most common words were for each sentiment category.





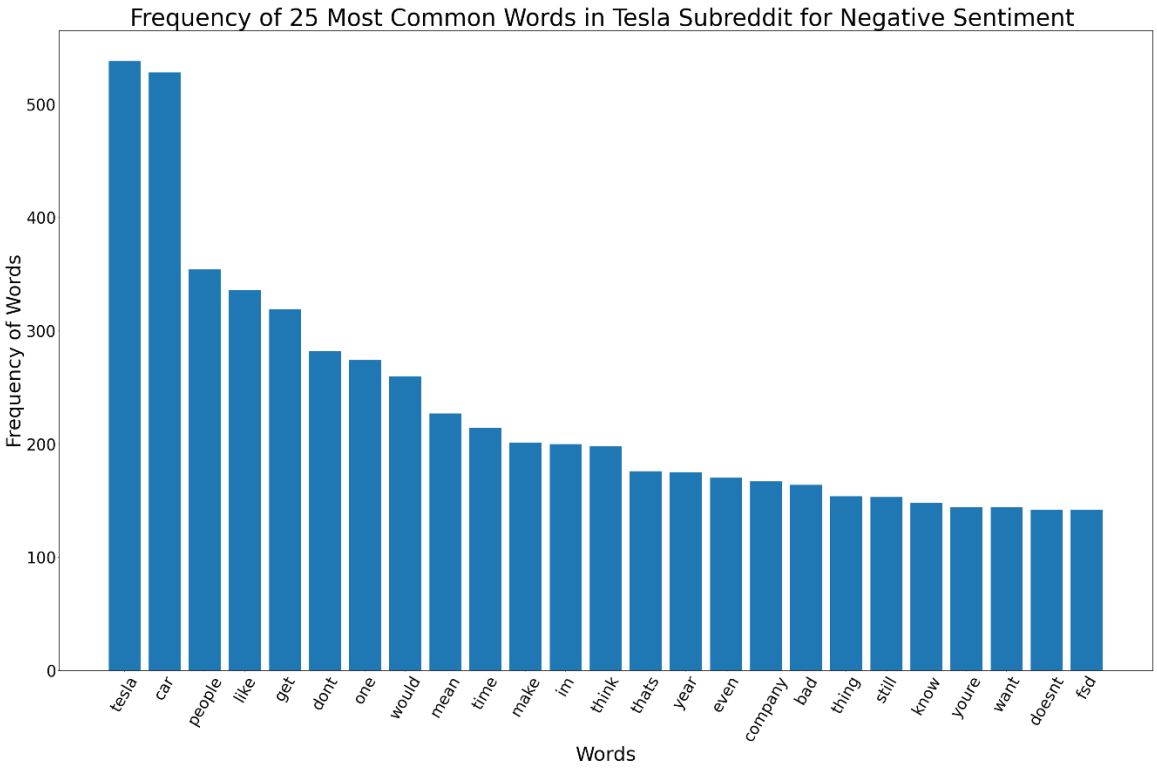
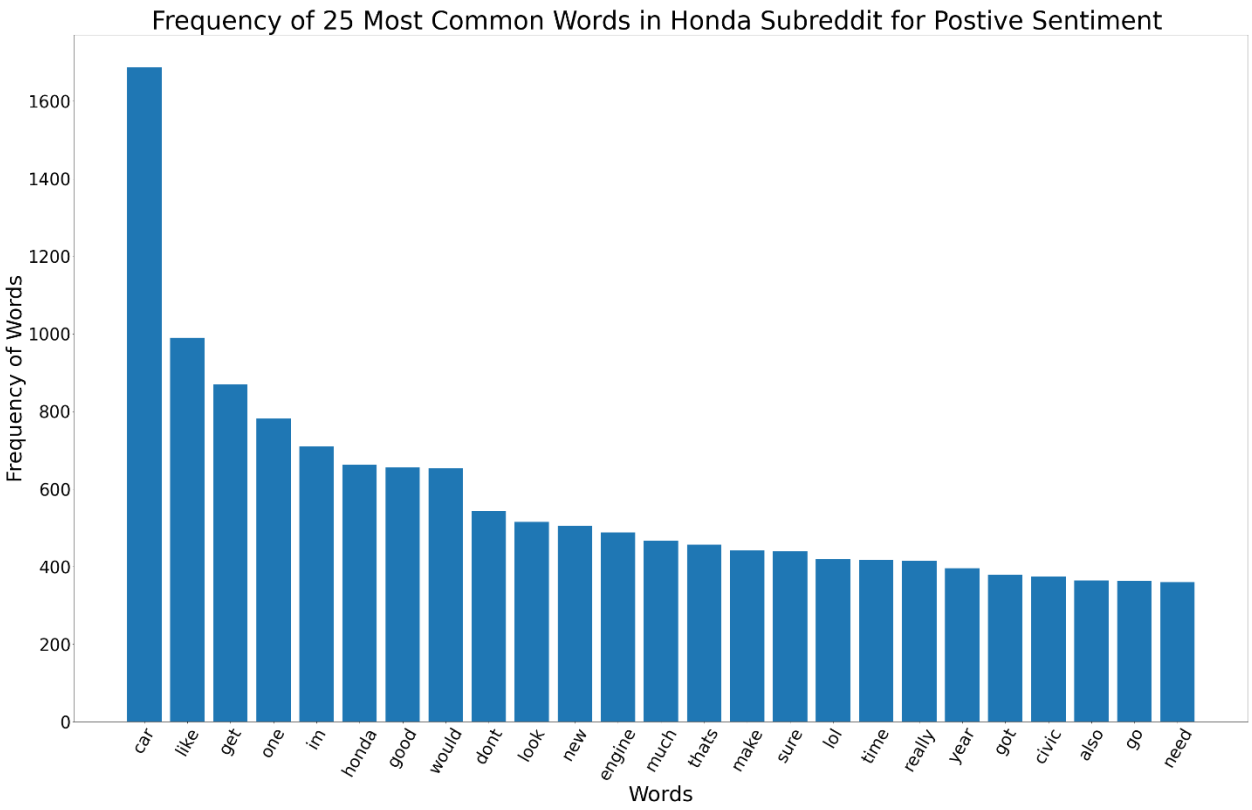
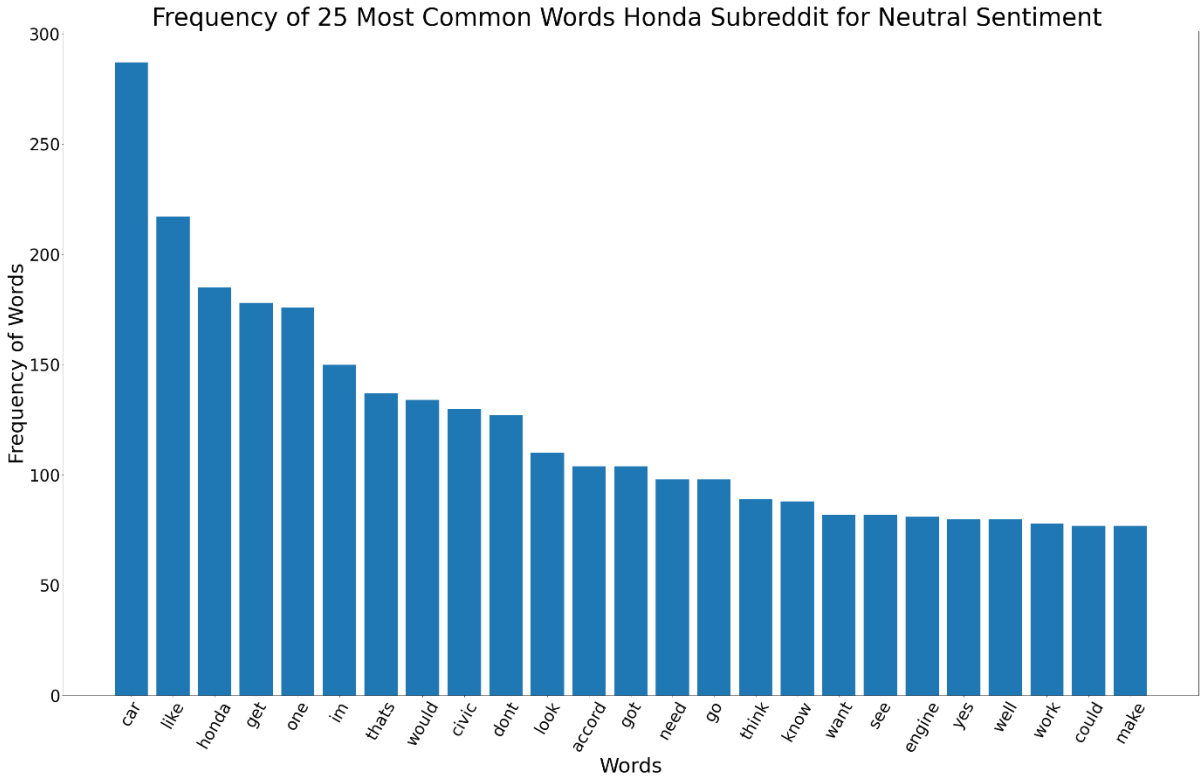


Figure 12 shows bar graphs of the 25 most common words in the tesla subreddit for each category (positive,neutral,negative) for 2022 comments.

In the positive words list 19 of them were also in the list of the most common words overall. This means almost all of them are present. Most of the words are the same across sentiment categories. “tesla” and “car” remained top common words despite sentiment category. Comments in the positive category had “new” included in the top 25 words list. This may indicate that people may feel positive of new things from Tesla such as new features coming out. Some unique words in the negative category include “company”, “bad”, and “fsd”. Fsd stands for Full Self Driving. This is a feature that has been talked about for years and still has yet to be accomplished with Tesla cars. Fsd being one of the most common words used in negative sentiment comments could indicate that people could be unhappy about this feature. Maybe it could be taking too long to be released for people.


Frequency of 25 Most Common Words in Honda Subreddit for Postive Sentiment
Frequency of Words
0
200
400
600
800
1000
1200
1400
1600
car
like
get
one
im
honda
good
would
dont
look
new
engine
much
thats
make
sure
lol
time
really
year
got
civic
also
go
need
Words


Frequency of 25 Most Common Words Honda Subreddit for Neutral Sentiment
Frequency of Words
0
50
100
150
200
250
300
car
like
honda
get
one
im
thats
would
civic
dont
look
accord
got
need
go
think
know
want
see
engine
yes
well
work
could
make
Words

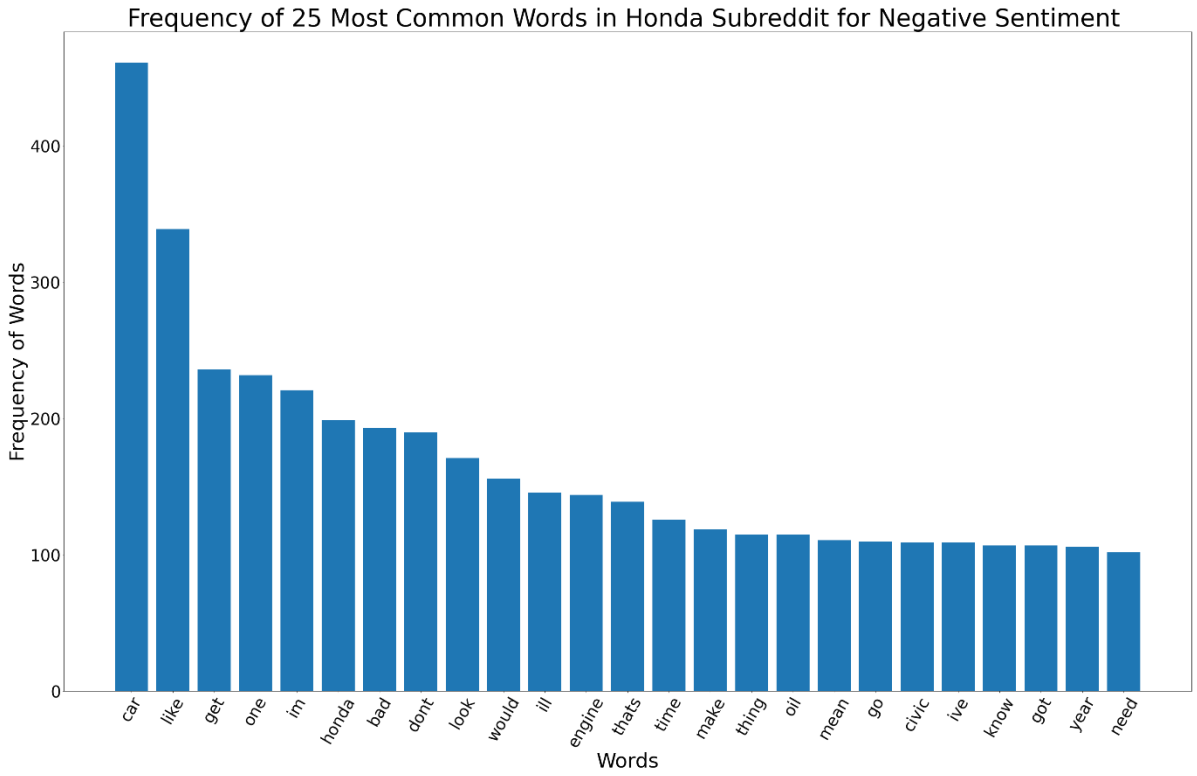


Figure 13 shows bar graphs of the 25 most common words in the Honda subreddit for each category (positive,neutral,negative) for 2022 comments.

The word “car” unsurprisingly dominates all other words in this subreddit despite sentiment category. One word that sticks out is “oil” here. It is not present in any other sentiment category, nor is it present in any of the Tesla words as well. It seems that oil is brought up quite frequently in negative leaning comments. This could suggest that Honda could be having oil issues in some of their cars. Another interesting unique word here is “ill” because that is not commonly used as a slang word anymore and being sick or ill does not seem to be car related. Maybe it could be covid related somehow.

There is very low indication that the 25 most common words may give any insight into what topics or areas that people may be giving positive or negative comments about since most words are shared across sentiments. This may be because the sentiment distribution of all comments is

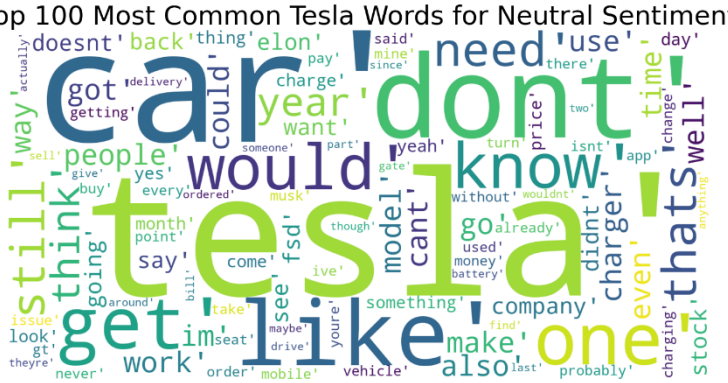
closer to zero meaning that they are all closer to neutral than to positive or negative. This could mean that there may be no common words that could point give any insight as to what topics or features about these manufacturers that may causing people to comment negatively or positively.

Word clouds off all the sentiment categories of both subreddits with the top 100 was create and compared.

Top 100 Most Common Tesla Words for Positive Sentiment

Top 100 Most Common Honda Words for Positive Sentiment

Top 100 Most Common Tesla Words for Neutral Sentiment



Top 100 Most Common Honda Words for Neutral Sentiment

Top 100 Most Common Tesla Words for Negative Sentiment

Top 100 Most Common Honda Words for Neutral Sentiment

Figure 14 shows word cloud charts for each subreddit (Tesla on the left, Honda on the right) and for each sentiment category (positive, neutral, negative) for 2022 comments.

In all of these word clouds we see all the most common words, the largest words in the charts to be mostly the same across sentiment categories and in both subreddits. “car” and the manufacturer name are unsurprisingly some of the largest words. What is interesting are the smaller words. For Tesla a form of “charger” type words are seen in every category. “fsd” is also seen in the positive category. In Honda, “oil” is seen in all categories. This may indicate that the previous results about these words indicating a pattern with positive/negative categories may not be accurate however, at the same time, it may be said that even though they are present in all categories, they are still more common to a certain category as top 25 words. For example, even though forms of “charger” are seen in all categories, it is still more prevalent in the positive category than the other words.

When It comes to word commonalties, there are some words that stick out however, most of the words are the same across sentiment categories and many even across subreddits. This might be because the average sentiments in general of bot subreddits are very close to 0. This means that even though there area a lot of comments with positive and negative polarities, many may give close to neutral results, which in turn may have many of the same words in common that are seen in the graphs.

Lastly, we want to look at previous years to see if sentiment has changed over time. we compare means of sentiments for 2018 -2022 for both the Tesla subreddit and the Honda subreddit.

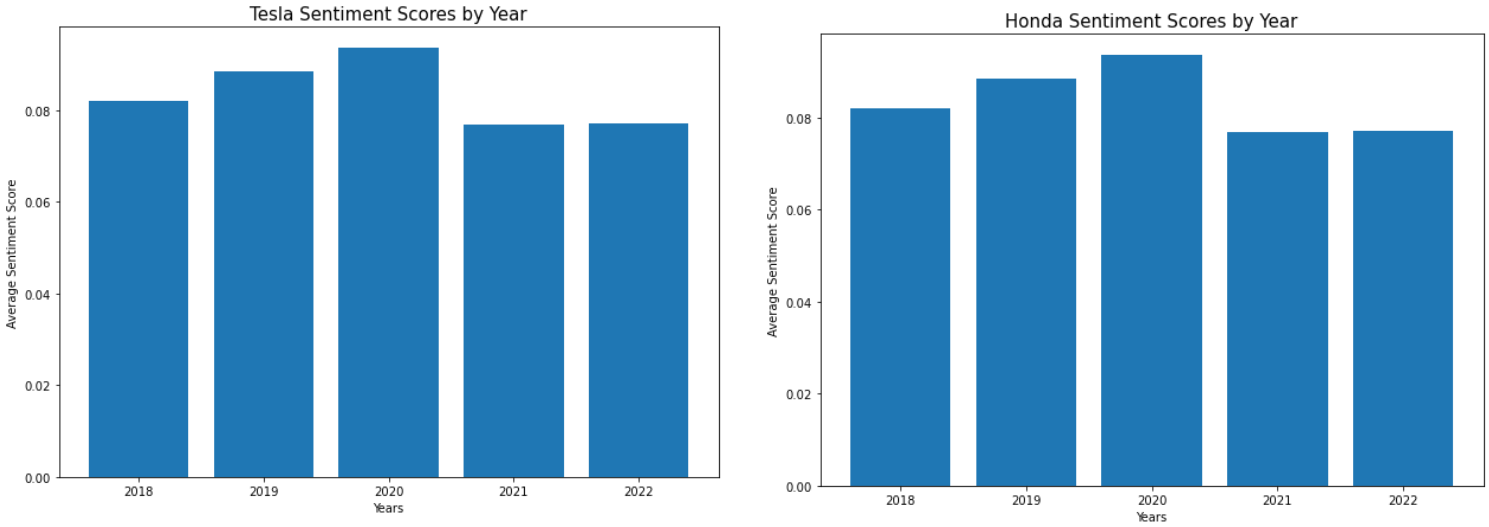


Figure 15 shows bar graphs of average sentiment scores by year for Tesla (left) and for Honda (right).

It looks like both subreddits have been following a similar pattern for the last 4 years. We see sentiment growing and peaking in 2020, then going down in 2021 and 2022. This pattern seems to align with when Covid-19 started to spread. For each year this data was gathered within the timeframe of January 1st to April 25th. Large Covid lockdowns and restrictions really started to take affect around March. This means that our 2020 data would mostly be from a pre-covid time period. This could mean that what we are seeing here for 2020 is the peak right before the pandemic occurred.

It may be a small difference but these graphs do illustrate a change in pattern starting with 2021 of sentiment dropping instead of following the trajectory with previous years.

Results

Originally, I predicted that the Tesla subreddit would have higher sentiment scores than Honda overall. According to this analysis, this is not what the data suggests. Honda's sentiment scores were significantly higher than Tesla's in 2022. This means that people generally commented more positively in Honda than in Tesla.

Despite the sentiment category it looks like most of the words remained the same in each subreddit, across categories. Both manufacturers had some words that stuck out for them in some categories. Tesla had "company", "bad", and "fsd" in the negative category, while having "new" be unique in the positive category. Honda had "oil" and "ill" in the negative category. Although the word chart revealed these words to be present in just about every other category as well but not in the 25 most common words groups. It is possible that with words like "fsd" and "oil" people could be saying negative things about features that they do not like. It is also possible that since the sentiment scores for both subreddits are closest around 0, we may be getting mostly neutral words without any discernible pattern present.

Looking at the average sentiment per year, both subreddits show a very similar pattern. Their average sentiments go up every year until 2021 in which they plummet and also stay down in 2022. Since the 2020 data was taken mostly form pre-covid months it might just be that we are seeing the peak sentiment before the pandemic hit and what might be seen from this data is that the sentiment was lower because of the covid-19 pandemic.

In conclusion, it looks like Tesla may not have as positive of a following as a long standing manufacturer such as Honda because Honda has a significantly higher sentiment score than tesla as of right now in 2022. It is hard to tell what reasons why these manufacturers have positive or

negative sentiments because most words associated with them are shared across sentiment categories. There is a drop in sentiment in post covid years that could indicate that the pandemic could have lowered people's sentiment. This could be a general drop and not related to the manufacturers themselves.

Conclusion

There are many ways that this experiment could be improved. For starters, I used an off the shelf method with text blob to do the sentiment analysis. The data that the textblob sentiment analysis model was probably trained on a different type of text data, not reddit manufacturer subreddit comments. A custom made algorithm trained on this exact kind of data would probably give more accurate results. There are also many more car manufacturers to be looked at and compared to Tesla. Only one other one was chosen for this experiment. Results could be different with an American manufacturer such as Chevy or Ford. Lastly, data cleaning could be improved. Text data can be very messy there are a lot different ways strategies to clean it up for analysis. For example, I handled contractions by simply taking out the apostrophe and keeping the word as is. Another strategy would have been to import a library to turn the contractions their full counterparts(wouldn't as would not). I chose not to do this because I felt that this might create extra words that were not truly there, but, it is certainly possible that it might give on better results.

I think another further avenue of research might be to see if there is a drop in sentiment between pre covid years to post covid years for subreddits unrelated to automobiles in order to see if this drop in sentiment could be generalized to reddit or social media as a whole.