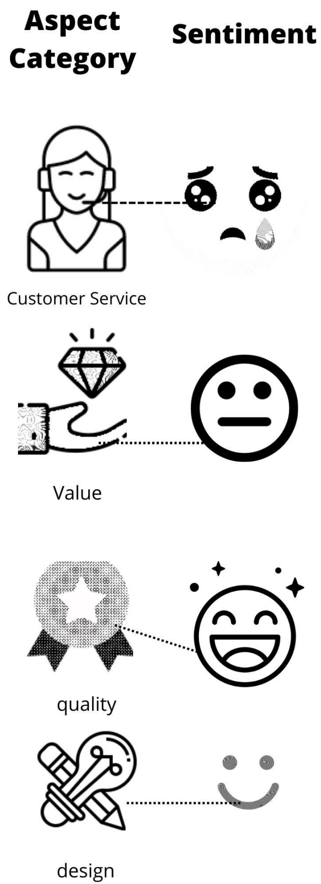
Aspect Based Sentiment Analysis



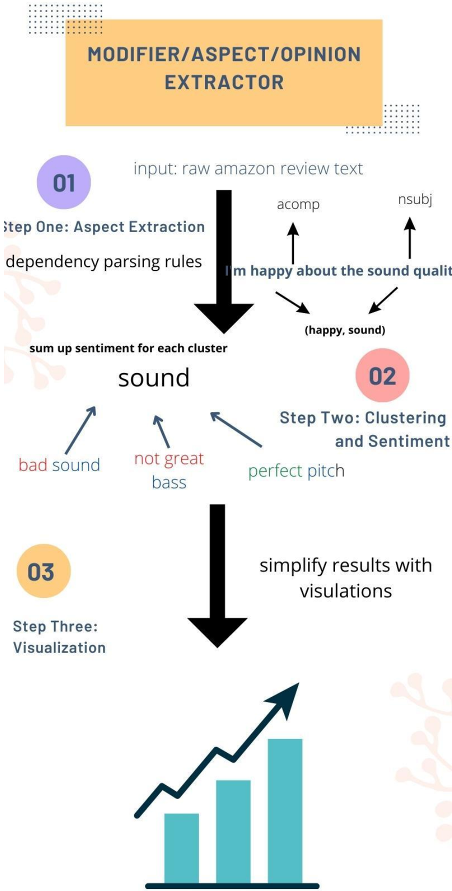
Making sense of messy amazon text data with a simple tool

Dylan Dey

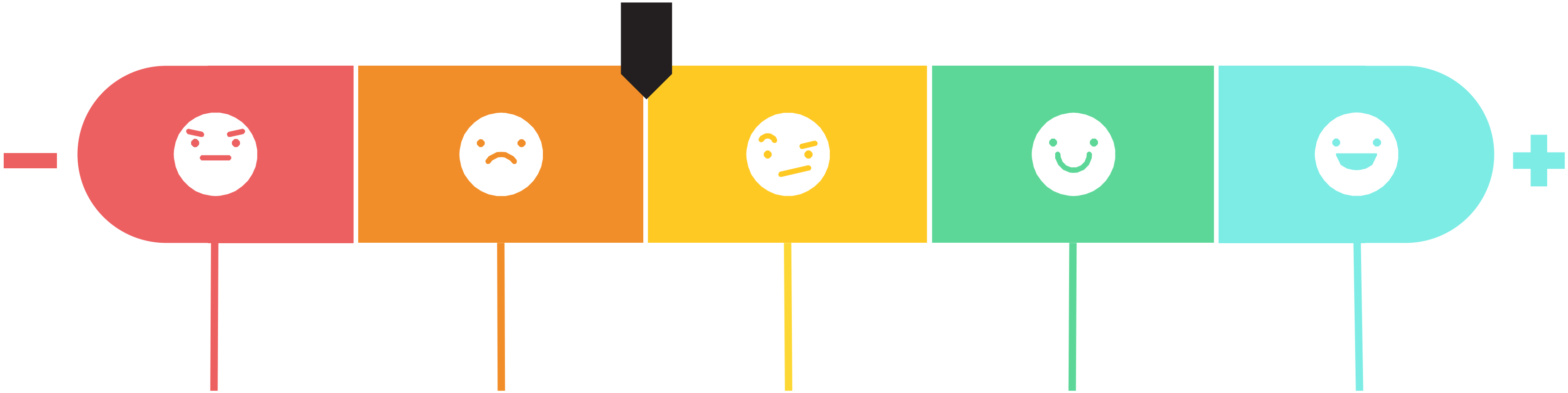
Before we begin…

Summary

- Creating an out of box aspect/opinion/sentiment triplet extractor using spaCy large english model and parsing logic/pos tagging capabilities
    - Clustering aspects to simple categories with unsupervised machine learning
    - Visualizing total sum of sentiment in clustered aspect graphs
- Verifying model performance with crowdsourced labels from Amazon Turk
- Verifying experimental setup

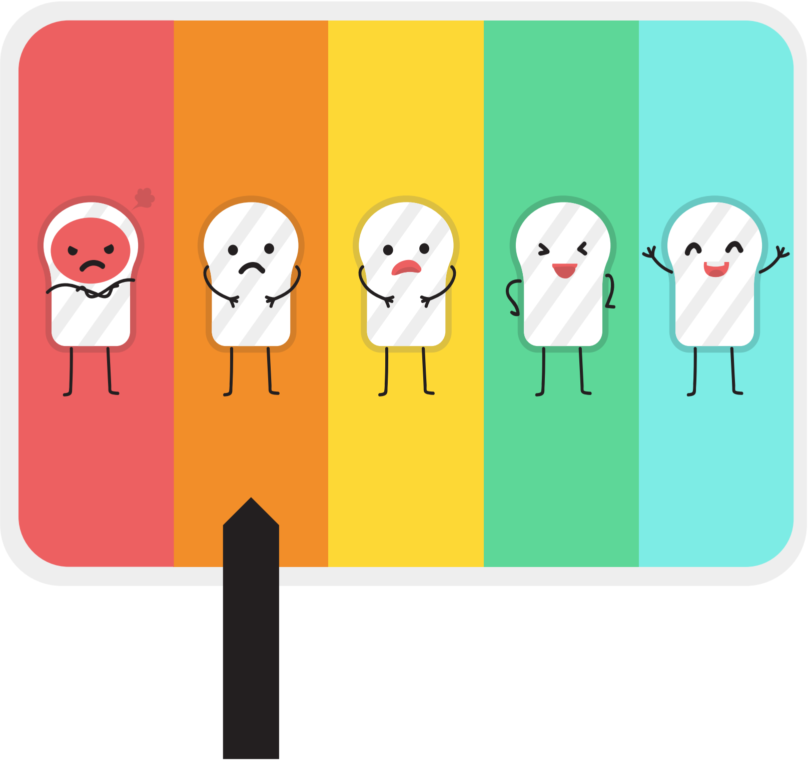


Outline



Business Problem	Data	Experimental Setup	Results	Discussion + Future Work
Why ABSA?	What type of data does the tool help with?	How did you verify the results of the tool?	Can we trust the performance of the tool?	Where do we go from here?

Why Aspect Based Sentiment Analysis (ABSA)



E-commerce

Today, most e-commerce website designs include a section where their customers can post reviews for products or services

Informal Reviews

There is potentially a disconnect from the amazon review ratings, and the overall sentiment of the body text explaining the review

Large Internet Text Data

It is often difficult to efficiently get useful data from a large collection of text data

Simplified

Transformation of informal review data can help make informative decisions

The Data

The Amazon Customer Reviews (Product Reviews) contains over 130+ million customer reviews available to researchers in TSV files in the amazon-reviews-pds S3 bucket in AWS US East Region, as per the provided readme file. The reviews were collected from 1995 to 2015. See the provided link for associated metadata. This project focuses on the dataset given by pulling "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Electronics_v1_00.tsv.gz" from the S3 bucket.

In [4]:
```
df = pd.read_csv('data/df_electronics.tsv', sep='\t')
df.groupby('product_id').count().sort_values(by='star_rating').tail(20)
```

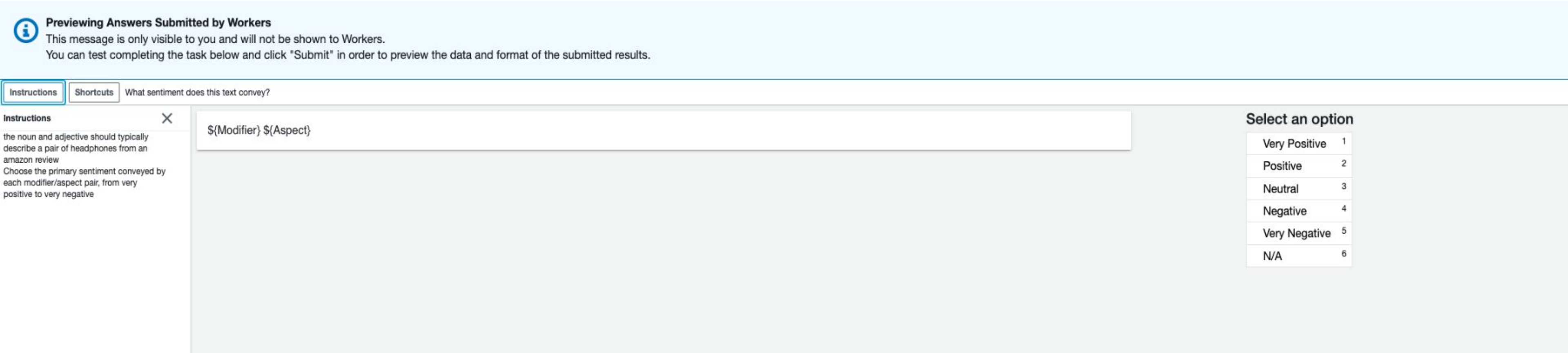
Out[4]:

product_id	marketplace	customer_id	review_id	product_parent	product_title	product_category	star_rating	helpful_votes	total_votes	vine	verified_purcl
B004LTEUDO	3997	3997	3997	3997	3997	3997	3997	3997	3997	3997	[illegible]
B004HHICKC	4213	4213	4213	4213	4213	4213	4213	4213	4213	4213	[illegible]
B0	click to scroll output; double click to hide	3	4773	4773	4773	4773	4773	4773	4773	4773	[illegible]
B001TH7GSW	4866	4866	4866	4866	4866	4866	4866	4866	4866	4866	[illegible]
B008KVUAGU	5015	5015	5015	5015	5015	5015	5015	5015	5015	5015	[illegible]
B003WGRUQQ	5072	5072	5072	5072	5072	5072	5072	5072	5072	5072	[illegible]
B002MAPT7U	5295	5295	5295	5295	5295	5295	5295	5295	5295	5295	[illegible]
B001GTT0VO	5580	5580	5580	5580	5580	5580	5580	5580	5580	5580	[illegible]
B0052SCU8U	5756	5756	5756	5756	5756	5756	5756	5756	5756	5756	[illegible]
B00316263Y	5813	5813	5813	5813	5813	5813	5813	5813	5813	5813	[illegible]
B00D5Q75RC	6062	6062	6062	6062	6062	6062	6062	6062	6062	6062	[illegible]
B004QK7HI8	6536	6536	6536	6536	6536	6536	6536	6536	6536	6536	[illegible]
B00F5NE2KG	6688	6688	6688	6688	6688	6688	6688	6688	6688	6688	[illegible]
B0019EHU8G	7586	7586	7586	7586	7586	7586	7586	7586	7586	7586	[illegible]
B000WYVBR0	7835	7835	7835	7835	7835	7835	7835	7835	7835	7835	[illegible]
B0001FTVEK	8793	8793	8793	8793	8793	8793	8793	8793	8793	8793	[illegible]
B0012S4APK	9359	9359	9359	9359	9359	9359	9359	9359	9359	9359	[illegible]
B003EM8008	9766	9766	9766	9766	9766	9766	9766	9766	9766	9766	[illegible]
B0002L5R78	11166	11166	11166	11166	11166	11166	11166	11166	11166	11166	[illegible]

Product_id "B0001FTVEK", or Sennheiser-RS120-Wireless-Headphones - was chosen to showcase the triplet extractor as it had a large amount of verified reviews and a pair of headphones seemed like a reasonable choice for aspect based sentiment analysis.

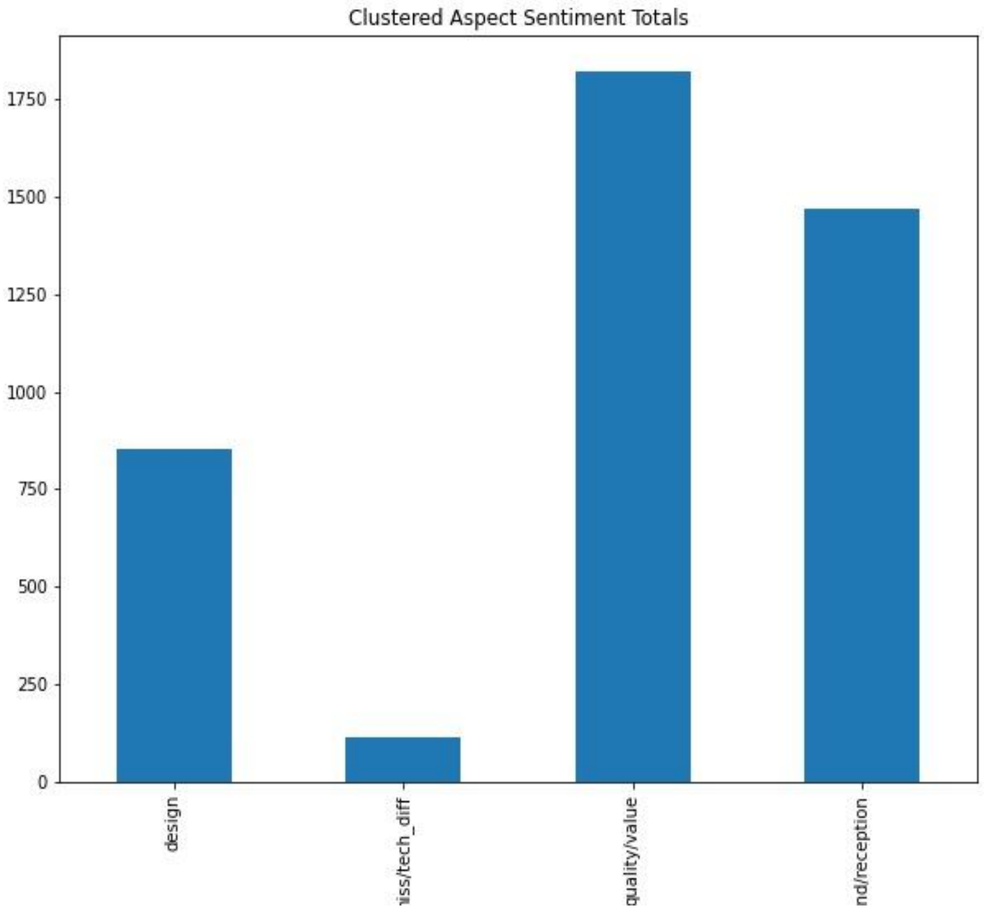
The Data/Experimental Setup

- The aspect/opinion pairs are created through unsupervised machine learning, more specifically a k-means clustering algorithm by scikit-learn with 4 clusters. This means that some “new” data was generated with this project. Below you can see an example of an html sheet a Turk Worker was assigned when labeling aspect/opinion pairs for this project.
- In total, 410 workers submitted 6107 non-null aspect/opinion pairs for sentiment intensity pertaining to 1438 unique aspects. Duplicate pairs of aspect/opinion pairs were included to inspect variance of submission from human labels and machine labels for each opinion pair. No qualifications or screening was put in place before the workers were chosen, but I did review sections of the data and accept or reject what seemed reasonable.
- All aspects/opinion pairs generated from Product_id “B0001FTVEK" from previous slide



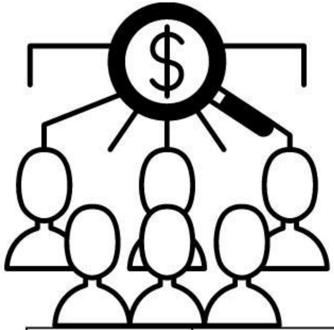
Results

- a lot of positive sentiment quality/value and sound/reception categories as compared to the design and hiss/tech_diff categories.
- The hiss category is low enough that it should be the major focus for the company to funnel resources in response to customer demand for improving their product.
- focus first on fixing the hiss
- Look at improving Design next such as batteries and cradle

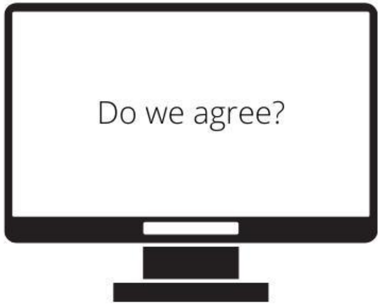


Results

Comparison of Turk Data to Extractor Data

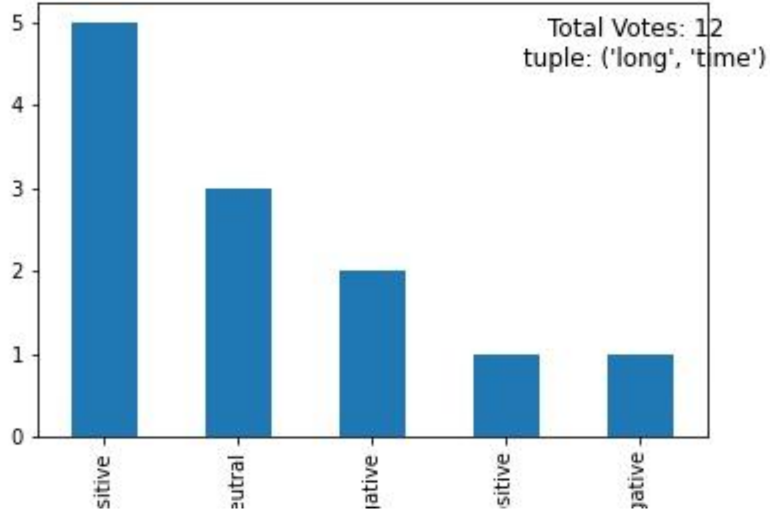


VS



	Neutral	Positive	Very Positive	Negative	Very Negative
Precision	36.20%	51.64%	35.76%	35.52%	23.86%
Accuracy	23.58%	7.76%	3.68%	2.78%	0.41%

Human Votes for Sentiment on Unique Aspect/Opinion Pair



Human labeled tuples	Sum_of_squares
('wireless', 'headphones')	232.75
('second', 'set')	4.75
('clear', 'sound')	26

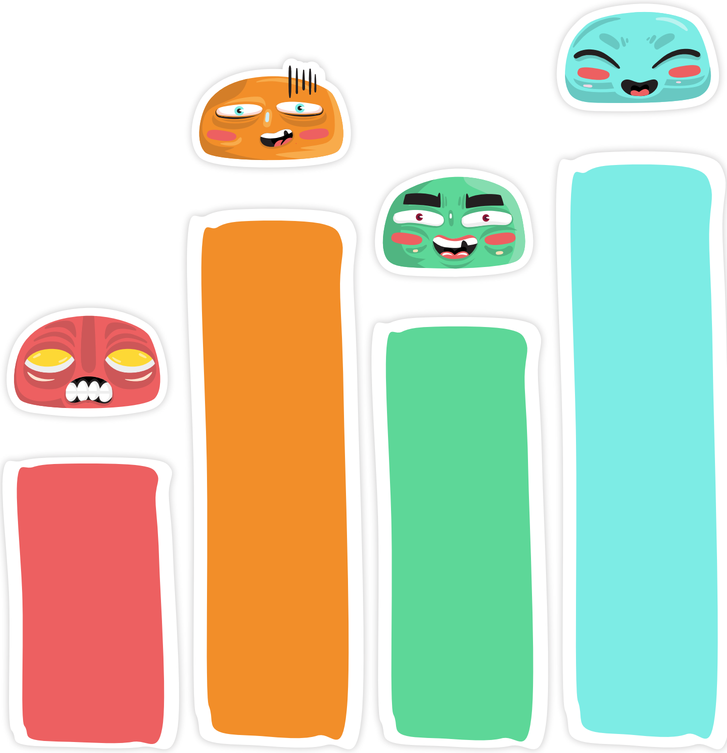
Discussion and Future Work

1) Due to a lack of funding from grants and some other issues in experimental design, I feel the most appropriate next step would be to repeat the experiment with more carefully collected labeled data. Increasing the reward per label and the requirements for submission (such as proving a proficiency in english) I believe can substantially improve the quality of the human labeled data and reduce the variance in this data significantly. Sourcing reliable test data is the number one priority in continuing future work for this project.

2) Integrate into AWS for scaling

3) integrate into a SQL server for data management

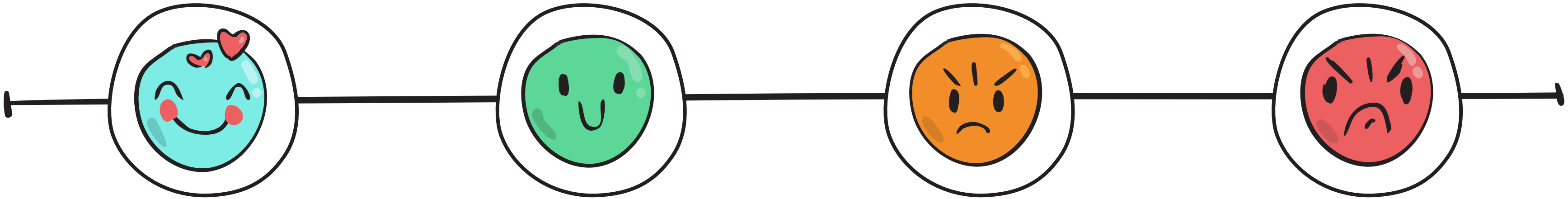
4) incorporate a flash based UI for viewing results at scale



Thank You!

Dylan Dey

Ddey2985@gmail.com



Results Appendix

Residual sum of squares for all human tuples that have more than ten votes. Every single aspect/opinion pair labeled by the extractor displayed no variance between identical tuple pairs, and all sum_of_square residual errors were zero.

Human labeled tuples	Sum_of_squares
('wireless', 'headphones')	232.75
('second', 'set')	4.75
('clear', 'sound')	26
('sound', 'quality')	870
('great', 'headphones')	19
('second', 'pair')	11.2
('good', 'sound')	60.75
('rechargeable', 'batteries')	4.666666667
('good', 'range')	16.66666667
('great', 'quality')	58.75
('Sound', 'quality')	4.5
('great', 'product')	34.75
('great', 'range')	18.66666667
('great', 'sound')	60.66666667
('good', 'quality')	83
('long', 'time')	11.2