VADER sentiment analysis

After identifying key features and issues with the work cloud analysis, sentiment analyses were performed on both the review titles and contents to further investigate and quantify sentiments expressed in the reviews. Valence Aware Dictionary and sEntiment Reasoner (VADER), a lexicon-based sentiment analysis tool developed specifically for social media, was utilized to analyze both the review titles and contents for the selected Echo devices. Examples of VADER analyses are shown in Figure 1. The analysis outputs 4 scores for both unidimensional and multidimensional sentiment evaluation:

- **Compound (Cpd) score.** This score, ranging between -1 and 1, provides a single unidimensional measurement of sentiment. A "positive" text will score 0.05 or above, while a "negative" text will score -0.05 or below. A score between -0.05 and 0.05 indicates a neutral sentiment.
- **Pos, Neu, Neg scores.** These scores represent the relative proportions of positive (Pos), neutral (Neu) and negative (Neg) text in the inputs. Each score ranges from 0 to 1, and the sum of the scores equals 1. These allow multidimensional measurement of sentiments expressed in the input text.

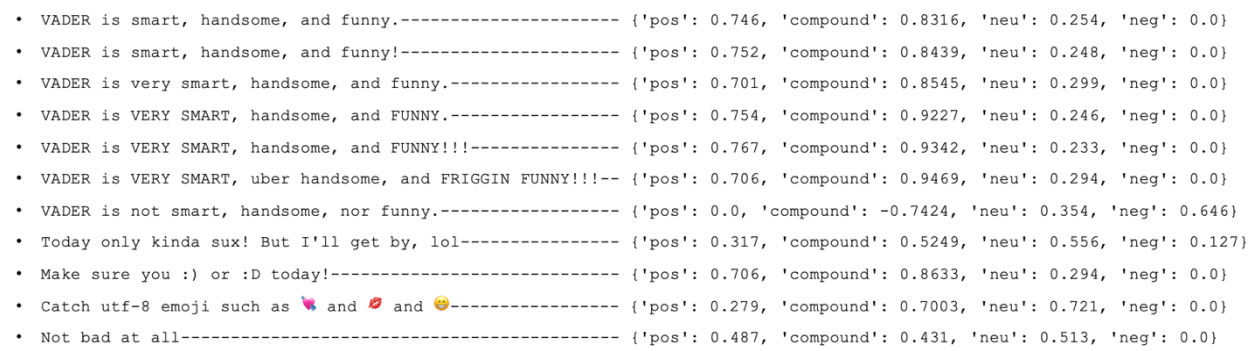

Figure 1. Examples of VADER sentiment analysis.

The VADER analysis of the review contents is summarized in Figure 2. For each product, the VADER scores are plotted against the star ratings in swamp plots, and all reviews are colored based on star rating. Overall, the distribution of the compound scores appear to be in agreement with the review's star ratings. The Cpd score plots show that, for all products, the majority of 5-star reviews scores above 0.05, reflecting the positive sentiments expressed in these reviews. As star rating decreases, more reviews are scored below 0, which indicates negative sentiments. The score distributions of 1- and 2-star reviews, however, do not mirror the 4- and 5-star reviews. Instead, at the other end of the spectrum, about half of the 1-star reviews score above 0 and the other half below 0. In addition, data points appear to be evenly distributed from -1 to 1 for most devices except those of Echo Show 1st Generation, which show an uneven distribution with more reviews scored at either extreme (-1 or 1) and less scored as neutral, i.e. around 0. In other words, many reviews with low star rating have a positive score, many of which score similarly as 1 just like a 5-star reviews.

The Pos, Neu, and Neg scores provide additional insights into how reviews are scored in this sentiment analysis. Similar to the Compound scores, the distributions of Pos and Neg scores appear to agree with the star ratings. Most 5-star reviews have Pos scores ranging from 0 and 0.8

and Neg scores ranging from 0 to 0.2, indicating a 5-star review could have up to 80% of positive text and 20% of negative text. On the other hand, most 1-star reviews have Pos scores ranging from 0 to 0.2 and Neg scores ranging from 0 and 0.2, reflecting a 1-star review could have up to 20% of positive text as well as 20% of negative text. The big difference in Pos scores and small difference in Neg scores suggest that reviews, regardless of star ratings, share similar relative amount of negative text, and the relative amount of the positive text increases as the star rating increases. Interestingly, many of the reviews, independent of star rating, have a Neu score above 0.5, implying that more than 50% of the contents in these reviews are rated neutral, and those scored below 0.5 are primarily 5-star reviews.

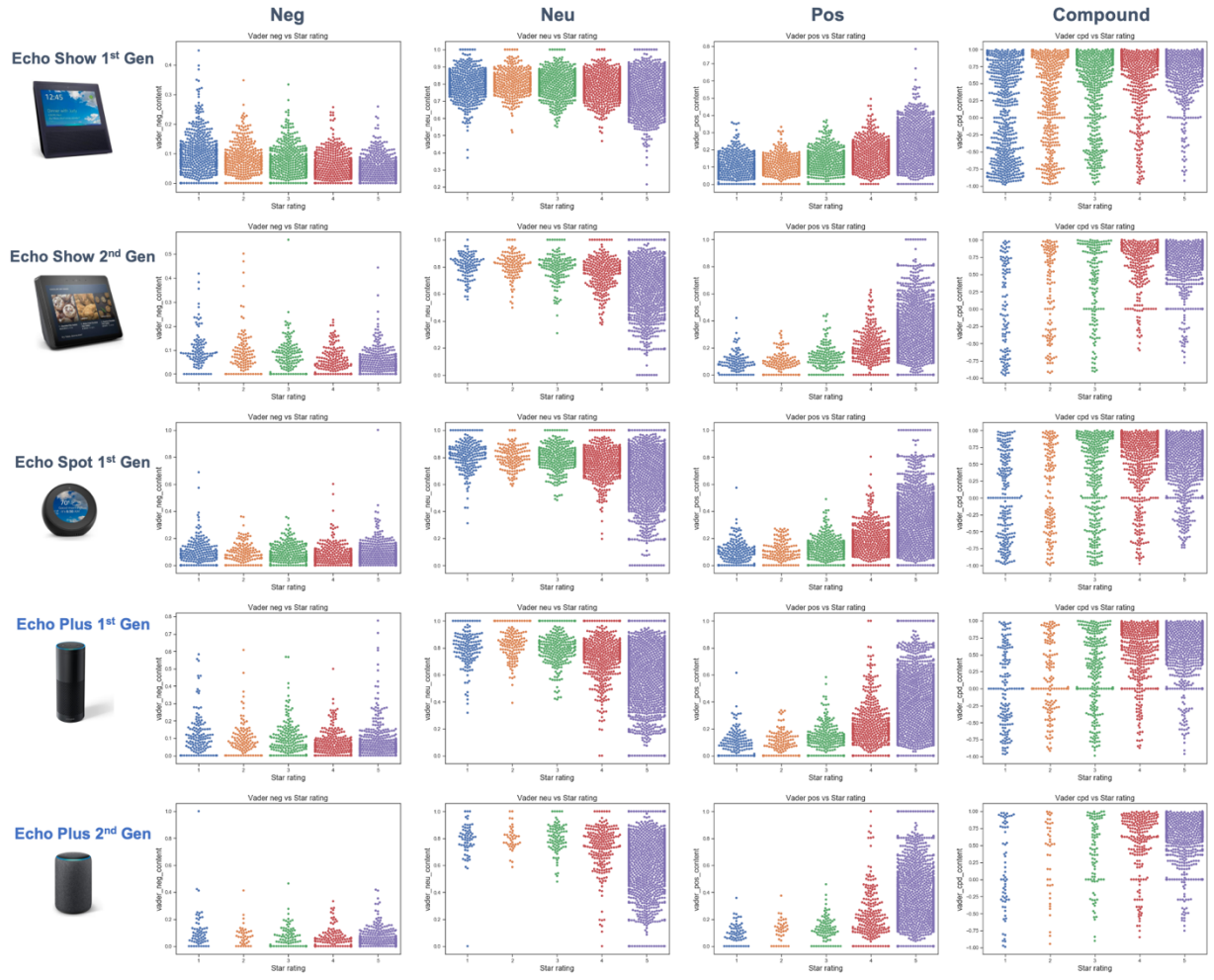

Figure 2. VADER analysis of Echo devices. The VADERS scores (Pos, Neu, Neg, and Compound) for each device are plotted against the review's star ratings. Reviews of 1-, 2-, 3-, 4- and 5-star are colored in blue, orange, green, red, and purple, respectively.

Continuing the study of the relationship between review length and star rating, the VADER Compound scores were plotted against the star ratings in Figure 3 for two of the selected devices, Echo Show 1st Generation and Echo Plus 1st Generation. The graphs look similar for both devices and the review lengths do not seem to correlate with the Compound scores. The long

reviews are not exclusively those with low star ratings or negative Compound scores, including those with either high star ratings or positive scores.

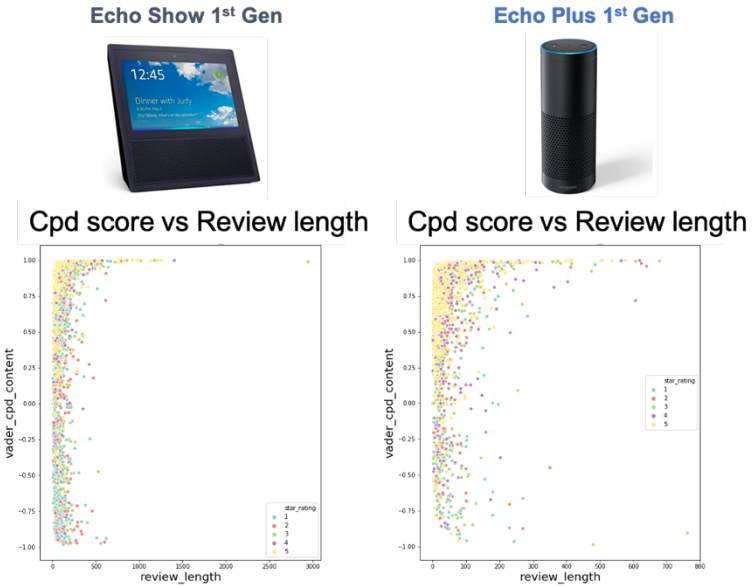

Figure 3. VADER Compound scores vs review length.

In addition, the Compound scores are further used to explore whether the sentiments change over time, which could be, for example, due to software and feature update. In Figure 4, the Compound scores computed from reviews of Echo Show 1st Generation and Echo Plus 1st Generation are plotted against review submission date. The graphs show no relationship between the Compound scores and time. Instead, the graphs do show peaks of review submission after sales events, including after Christmas and after Prime Day.

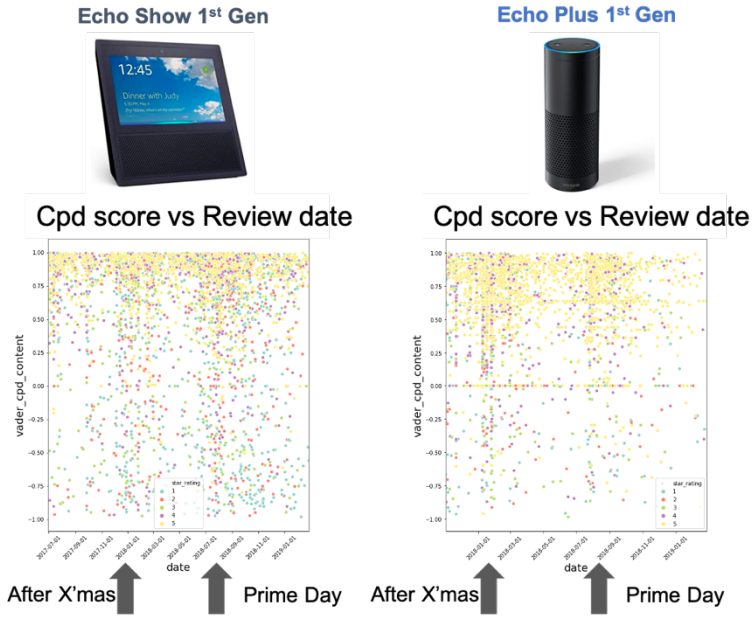

Figure 4. VADER Compound scores vs review date.

Sentiment polarity ratio: scoring sentiments based on Pos and Neg

As shown in Figure 2, while the VADER score distributions appear to make sense, many reviews with low star rating were scored as positive, i.e. Cpd score < 0. Based on the relative proportions of positive, negative, and neutral texts, the amount of positive text appears to be the most significant component that contributes to the sentiment evaluation. Notably, many reviews do score with a low Pos or Neg scores, i.e. close to 0, while many have a large portion of neutral text. As shown in Figure 3, the review lengths could further confound the analysis.

In order to better understand and evaluate both the positive and negative sentiments expressed, a new scoring model was developed to focus on assessing the positive and negative contents. The key concept here is to evaluate the polarity of the sentiments expressed based on the relative proportions of the positive and negative contents. The sentiment polarity, calculated as the ratio of the Pos and Neg scores (Pos/Neg), provides a measure to characterized how polarized the input text is by taking into account of both positive and negative sentiments. In other words, one could evaluate how positive a review is not only by the amount of positive text but also how much negative text is present. Few examples are shown in Figure 5. If a review (example 1) has 50% positive text (Pos=0.50) and 10% negative text (Neg=0.10), the Pos/Neg ratio is 10, indicating that the positive text in the review is 10-fold more than the negative text. On the other hand, if a review (example 2) has 10-fold more negative text than the positive text, then the Pos/Neg ratio is 0.1. If the positive text and the negative text is in the same relative proportion as shown in example 3, then the Pos/Neg ratio will equal 1.

Example	Pos	Neg	Pos/Neg	Log(Pos/Neg)
1	0.50	0.05	10	1
2	0.05	0.50	0.1	-1
3	0.25	0.25	1	0
4	0.20	0.02	10	1
5	0.09	0.90	0.1	-1
6	0.01	0.01	1	0

Figure 5. Sentiment polarity ratio examples.

As both Pos and Neg are relative proportions, the Pos/Neg ratio is independent of the review length or the actual Pos and Neg scores. For example, if a review has 20% of positive text and 2% of negative text as shown in example 4, the resulting Pos/Neg ratio will be 10 just like for example 1 which has higher Pos and Neg scores but shorter. Similarly, examples 5 and 6 produce the same Pos/Neg ratios as those of examples 2 and 3, respectively, while they each has a different set of Pos and Neg scores. Lastly, a minimal score of 0.001 is applied for both Pos and Neg scores to ensure numerical stability.

The sentiment polarity ratios calculated for the different Echo devices, expressed in Log scale, are shown in Figure 6. For 5-star reviews, the majority of reviews scored Log(Pos/Neg) above 1, i.e. Pos > Neg, indicating that these reviews have more positive text than negative text. On the other hand, the 1-star reviews mostly have Log(Pos/Neg) between -1 and 1, indicating that these

review comments are in general “neutral” with about a 10-fold difference between Pos and Neg. As discussed above, some of the reviews have VADER Cpd score inconsistent with the corresponding star rating, such as those 5-star reviews with Cpd score < 0.05 or 1-star reviews with Cpd score > 0.05. Taking the Pos/Neg ratio as a sentiment measure also helps to reduce the number of such cases, i.e. those with at least 100-fold difference between Pos and Neg, including 5-star reviews Log(Pos/Neg) < -2 and 1-star reviews with Log(Pos/Neg) > 2. Most importantly, the Pos/Neg ratio is interpretable. For example, for a review with Log(Pos/Neg) = 2, or Pos/Neg = 100, the result indicates that the review has 100 positive words for every negative word, which really reflects a high level of satisfaction and high star rating.

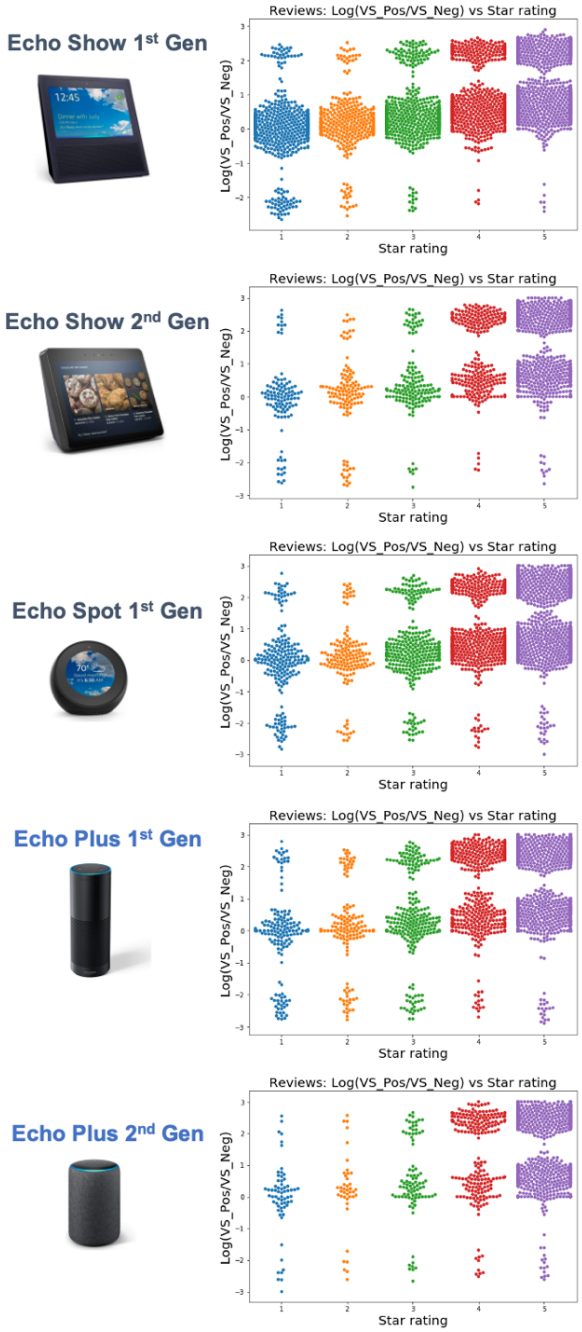

Figure 6. Sentiment polarity ratios for Echo devices.

To provide a similar unidimensional measure of the sentiments like the VADER Cpd score, the geomean of the Pos/Neg ratios is calculated for each star rating group. By plotting the geomean of the sentiment polarity ratios against the star rating, a sentiment profile of the reviews can be constructed for each product as shown in Figure 7. Interestingly, the graph shows two distinct review profiles. With effectively identical sentiment polarity ratios for reviews from 1- to 3-star, the two profiles differ at 4- and 5-star reviews. The higher rating group, which includes Echo Plus 1st Gen, Echo Plus 2nd Gen, and Echo Show 2nd Gen, share polarity ratios of around 20 and 110 for 4- and 5-star reviews, respectively. The lower rating group, including Echo Show 1st Gen and Echo Spot 1st Gen, show polarity ratios of around 10 and 30, for 4- and 5-star reviews, respectively. This result shows that the Echo Plus speakers received more positive reviews than the screen-enabled counterparts, with the exception of Echo Show 2nd Gen, suggesting that the screen-less speakers achieved a higher level of satisfaction without a screen. Relative to the 1st Gen device, the Echo Show 2nd Gen was able to raise the satisfaction level significantly to the same level as that of the Echo Plus speakers. Based on the word cloud analyses as shown in Figure 6, many of the top features of Echo Show 2nd Gen, like the Echo Plus speakers, are related to sound quality and related audio features, which are less prominent in reviews of the 1st Gen device. This appears to suggest that Amazon made Echo Show 2nd Gen a better, more satisfied product by making the device a better smart speaker. In other words, this may reflect that, in early 2019, Echo Show users enjoyed using their screen-enabled devices more as a smart speaker and less so as a multimedia or video call device.

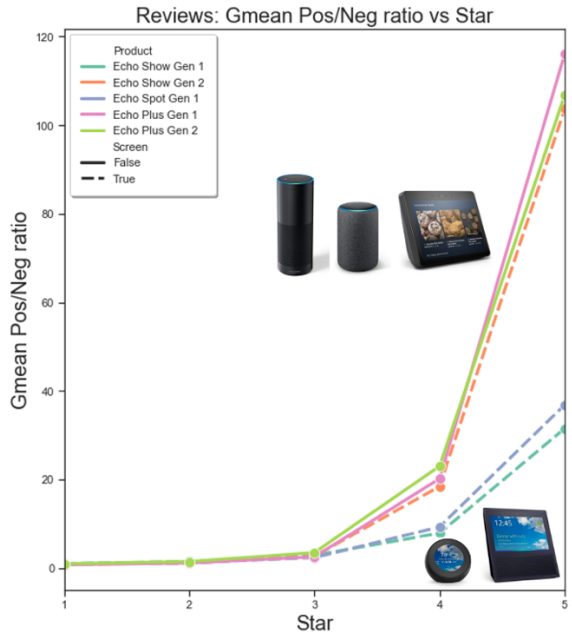

Figure 7. Sentiment profiles based on sentiment polarity ratios and star ratings.

As the sentiment polarity ratio has provides positive results for this Echo dataset, the study could be expanded to examine newer Echo devices as well as other similar smart devices. The concept of sentiment polarity ratio could also be applied to build a sentiment modeling tool. The bigram and trigram word clouds are shown to be informative but a bit messy. The results could likely be improved by a better data cleaning procedure.