Is talking to Alexa enough? Or do we need more?

Siegfried Leung

Amazon's Alexa-enabled Echo started as a smart speaker that served as an in-home hub for the company. There are now many Echo devices with different sizes and form factors, such as those with a screen. Alexa has also come a long way, as she has acquired many new "skills" over the past few years and she can now be the virtual butler of your smart home. Other than playing music, reading news, checking weather, or selling stuff from Amazon, Alexa can now help the user around the house if you have the supported smart home products and services.

This study is going to focus on Echo devices that have a screen – a logical and sensible next-step for the Echo smart speaker product line. These devices are no longer just smart speakers as they can now deliver both audio and visual information to the users. The screen provides another dimension for users to interact with these devices, and Alexa has also evolved to a more capable virtual assistant, who can talk as well show you things.

The key question of this study, as reflected by the title, is to investigate whether these Alexa-enabled devices is better with a screen. This study will examine the user experience to obtain insights into the following questions:
1) For Echo's users, does the added screen and its related functions provide a better experience relative to other screen-less Echo smart speakers?
2) In consideration of the next generation of similar Echo products, what are features or services to include, improve, or exclude?

Data set:

This study will investigate the user experience based on Amazon reviews submitted on five Echo products, including three devices with screen and two screen-less smart speakers:
Echo Show 1st Generation (with screen, #reviews = 4000)
Echo Show 2nd Generation (with screen, #reviews = 2048)
Echo Spot 1st Generation (with screen, #reviews = 4000)
Echo Plus 1st Generation (screen-less, #reviews = 3208)
Echo Plus 2nd Generation (screen-less, #reviews = 1832)

The reviews were downloaded at the end of February, 2019, using the web scrapper implemented in Chrome. The detailed procedures are outlined as follow:
https://www.scrapehero.com/amazon-review-scraper/

The scrapper extracts review information, including author, title, date, review content, and star rating, from each product page, and the reports, with up to 4000 reviews per product, are saved as csv files.

The review data is relatively clean. This study will focus on the review titles, review contents, star ratings and review submission dates. For text data, including titles and contents, punctuations were removed, and lower case was applied. Both reviews and titles were tokenized using the NLTK package, with stop words removed. Bigram and trigram tokens were also generated by linking unigram tokens with underscore. Character count and word count were also computed for both title and content to measure text length. Star ratings, captured as a string in the format of "X.0 out

of 5 stars", was converted to integers. A handful of invalid reviews have a star rating of 0, which were removed from the dataset. No missing data was otherwise found in these key columns.

Data analysis

The first step to explore the reviews data is to examine the satisfaction from each of these productions. Based on the provided star rating, ranging from 1 to 5, the mean star rating was plotted in Figure 1 below. Overall, all devices appear to have good similar ratings. With the exception of Echo Show 1st Generation (mean star rating = 3.7), the other four devices scored more than 4 stars and the Echo Plus 2nd Generation has the best rating of 4.6.

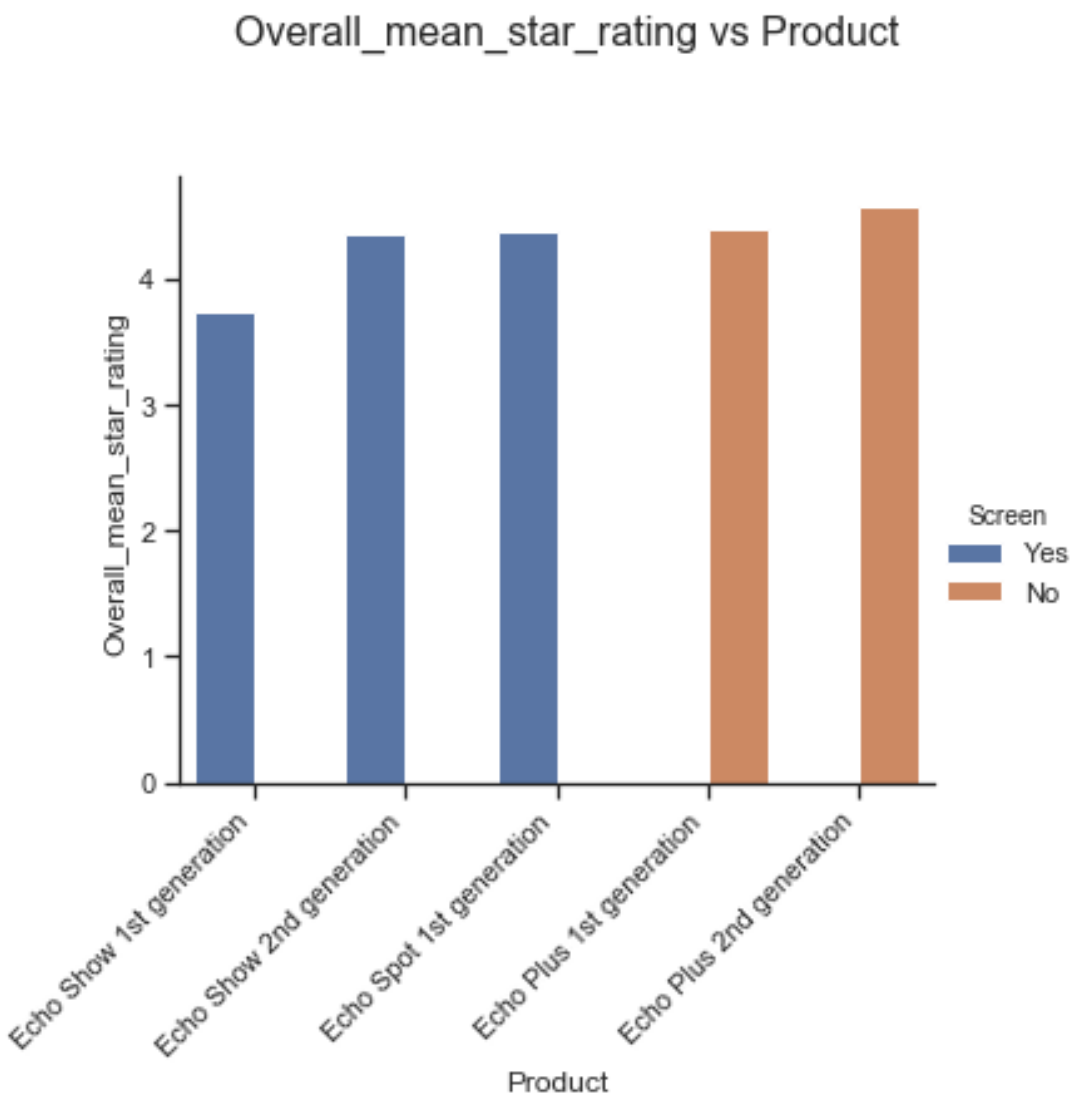

Figure 1. Mean star ratings of selected Echo devices.

The review length was further examined. The mean word counts were calculated from title and contents of reviews for each device. As shown in Figure 2, the titles and review title and contents

are shortest for 5-star reviews, and the word counts increase as the star rating decreases. Interestingly, the longest reviews are either 2-star or 3-star reviews, not the 1-star reviews. In addition, reviews for Echo Show and Echo Spot are generally longer than those for Echo Plus. The Echo Show 1st Generation has the longest reviews across all star ratings, while Echo Plus 1st Generation has the shortest reviews. In some cases, the difference is quite significant. For example, the 5-star reviews of Echo Show 1st Generation have 92 words in average, which are 3 times as long as those 5-star reviews from Echo Plus 1st Generation.

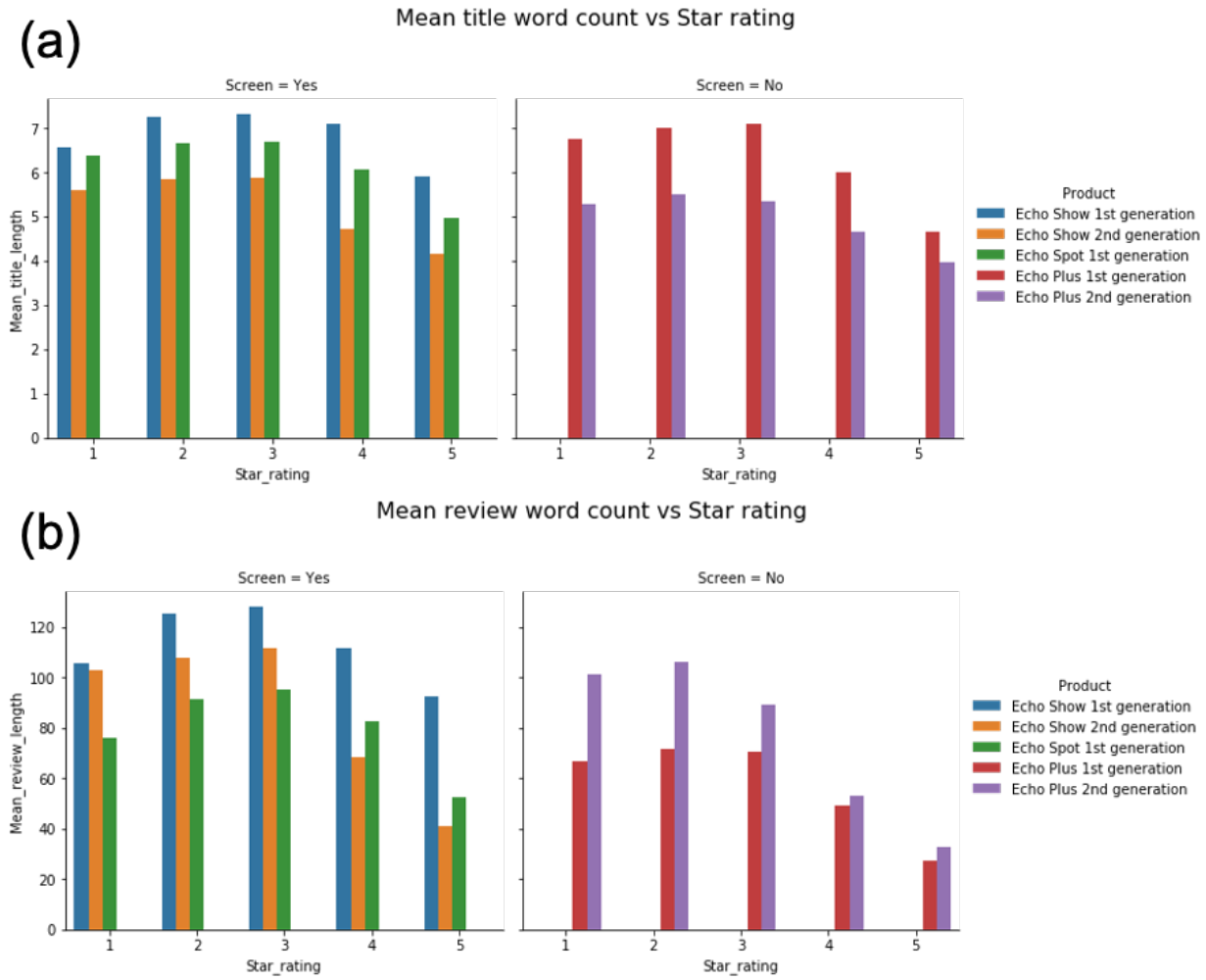

Figure 2. (a) Mean title lengths (word count) of reviews grouped by star rating. (b) Mean review content lengths (word count) of reviews grouped by star rating.

In Figure 3, the mean review length was plotted against the mean star rating, showing that reviewers, when they are not as satisfied with the device as reflected by the lower star rating, would write more. Interestingly, the two screen-less devices, with their relatively higher mean star ratings, have shorter reviews than the three devices with screen.

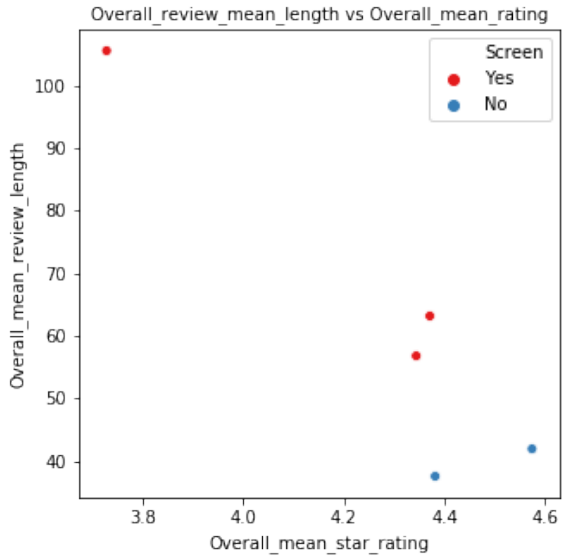

Figure 3. Mean review length vs. mean star rating of selected devices.

Further examination of the data shows that the reviews for all the five selected devices are highly skewed with mostly 5-star reviews. The number of reviews and their relative proportion for each star-rating group are plotted in Figures 4 and 5, respectively. Figure 4 shows that 5-star reviews made up for about 70% of the reviews for all devices but Echo Show 1st Generation, from which ~50% are 5-star reviews.

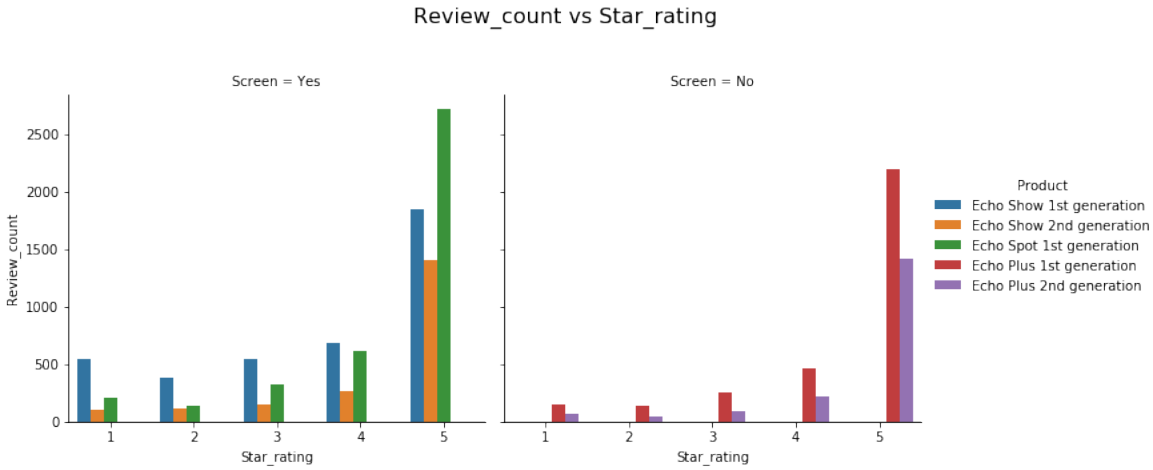

Figure 4. Number of reviews from each star rating group.

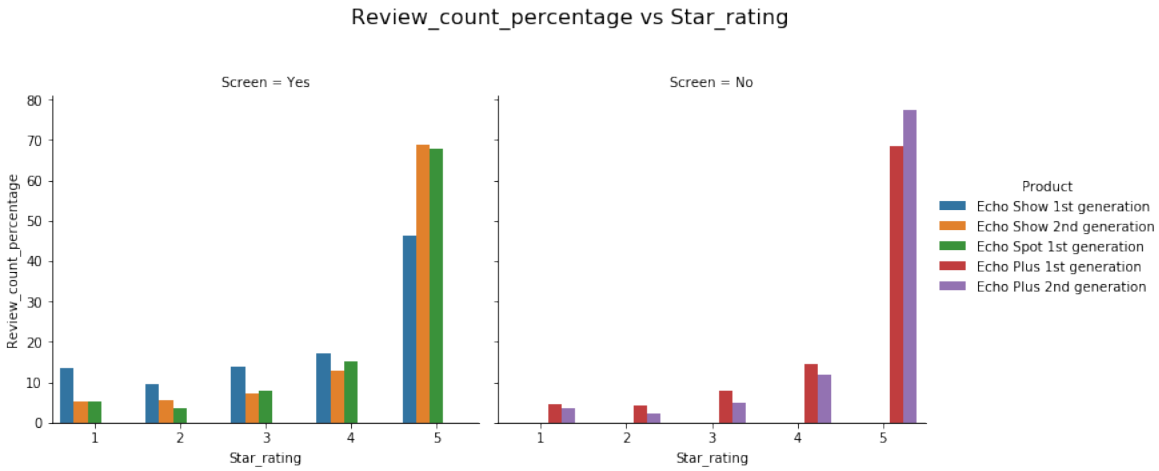

Figure 5. Relative proportion of review count from each star rating group.

The overall high mean star ratings can probably be explained by the skewed reviews. The above analysis also suggests that reviewer's satisfaction might influence how much they write. Understanding the relationship between the reviewer's sentiment and the review content would be of interest. However, being able to extract information regardless of the review's length would be important. The next steps will include examining the reviews further by word cloud analysis, which would help identifying features that are liked and disliked by the reviewers, and by performing a sentiment analysis, which would allow us to measure and analyze the sentiments expressed quantitatively.

Word cloud analysis

Monogram, bigram and trigram word cloud analyses were performed on both review titles and review contents of each star-rating group from all selected Echo devices. Aiming to identify the most liked and disliked features, the following analysis will focus on the 5-star and 1-star reviews. Word clouds were generated based on monogram, bigram and trigram frequencies, which are shown respectively in Figures 6 and 7 from 5-star and 1-star reviews. In each figure, the 20 most frequent words/phrase are shown, and the word size based on the computed frequency.

The word cloud analyses from the 5-star reviews reveal some of the most popular features (Figure 6). Reviewers, who gave 5-star reviews to Echo Show 1st and 2nd Gen, did mention the screen. When Echo Show was released, Amazon promoted the device's capability of making video calls via the Drop-in feature. Amazon also advertised Echo Show as a companion device in the kitchen. Both of these were present in the word clouds. Interestingly, reviewers of Echo Show 2nd Generation have mentioned "sound" a lot, which is similar to the reviews for the screen-less speakers. Reviews of Echo Spot, which is advertised as a smart alarm clock, show buyers do enjoy using the device as alarm clock. Unsurprisingly, Amazon's smart voice assistant, Alexa, appears to be a prominent feature for all devices. Non-Amazon services that are compatible with Echo, such as Spotify, however, are not present, suggesting that these third-party services are overshadowed by the Amazon's services.

Figure 6. Word clouds from 5-star review contents.

Figure 7. Word clouds from 1-star review contents.

Word clouds from 1-star reviews were generated to examine reviews at the other end of the spectrum. Interestingly, some features appear in both unigram word clouds from 5-star and 1-star reviews, such as Alexa, Amazon, music, screen (for Echo Shows and Echo Spot) and speaker (for Echo Pluses). This may suggest that 1-star reviewers also like these features or the 1-star reviewers actually dislike these features liked by many 5-star reviewers. The bigram and trigram word clouds do provide a bit more insights into reviews. For Echo Show devices, the reviews mentioned about YouTube videos and flash player, both of which is specific to screen-enabled devices. This likely relates to the rollout of Echo Show's own YouTube app at its release. The Echo Show 1st generation was first released with a YouTube app, however, the app was then blocked by Google for some time. The "Alexa app" and "phone" also appear here for multiple devices, which could reflect on interactions with the Echo devices via the Alexa app in the phone. This analysis also helps identifying couple device-specific issues. Some Echo Plus 2nd generation devices might have a problem with WiFi signal (from bigram word cloud). Some 1-star reviewers also reported screen issues, such as screen flickering and lines on the screen, for Echo Show 1st generation. These issues, while do not appear in the word clouds of the review contents, was present the word clouds generated from the review titles as shown in Figure 8.

Figure 8. Word clouds from 1-star review titles of Echo Show 1st Generation. Words and phrases related to screen flickering is highlighted in pink.

The next step is to carry out sentiment analysis to quantify the sentiments expressed in these reviews. The goal would be to connect the results from sentiment analysis and the features/issues identified from the word cloud analysis, so we could profile and compare user experience of these Echo devices.

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 9. 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.

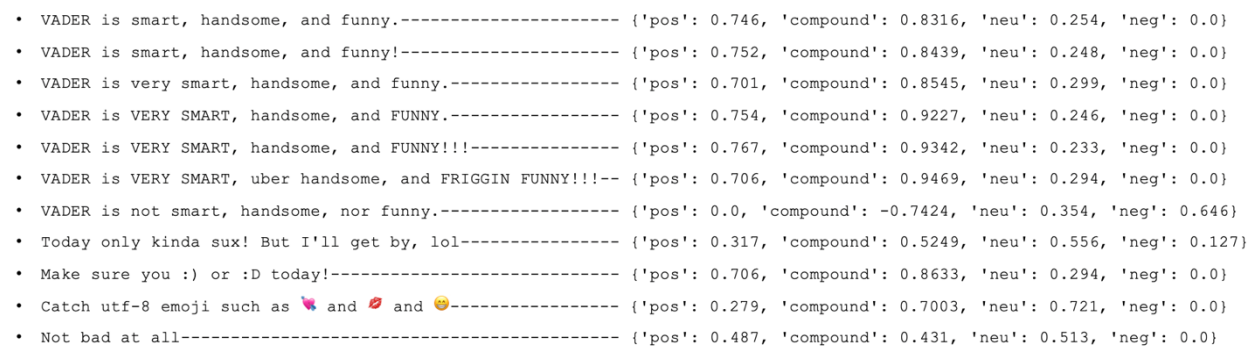

Figure 9. Examples of VADER sentiment analysis.

The VADER analysis of the review contents is summarized in Figure 10. 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 for Echo Show 1st Generation, which has 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 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.

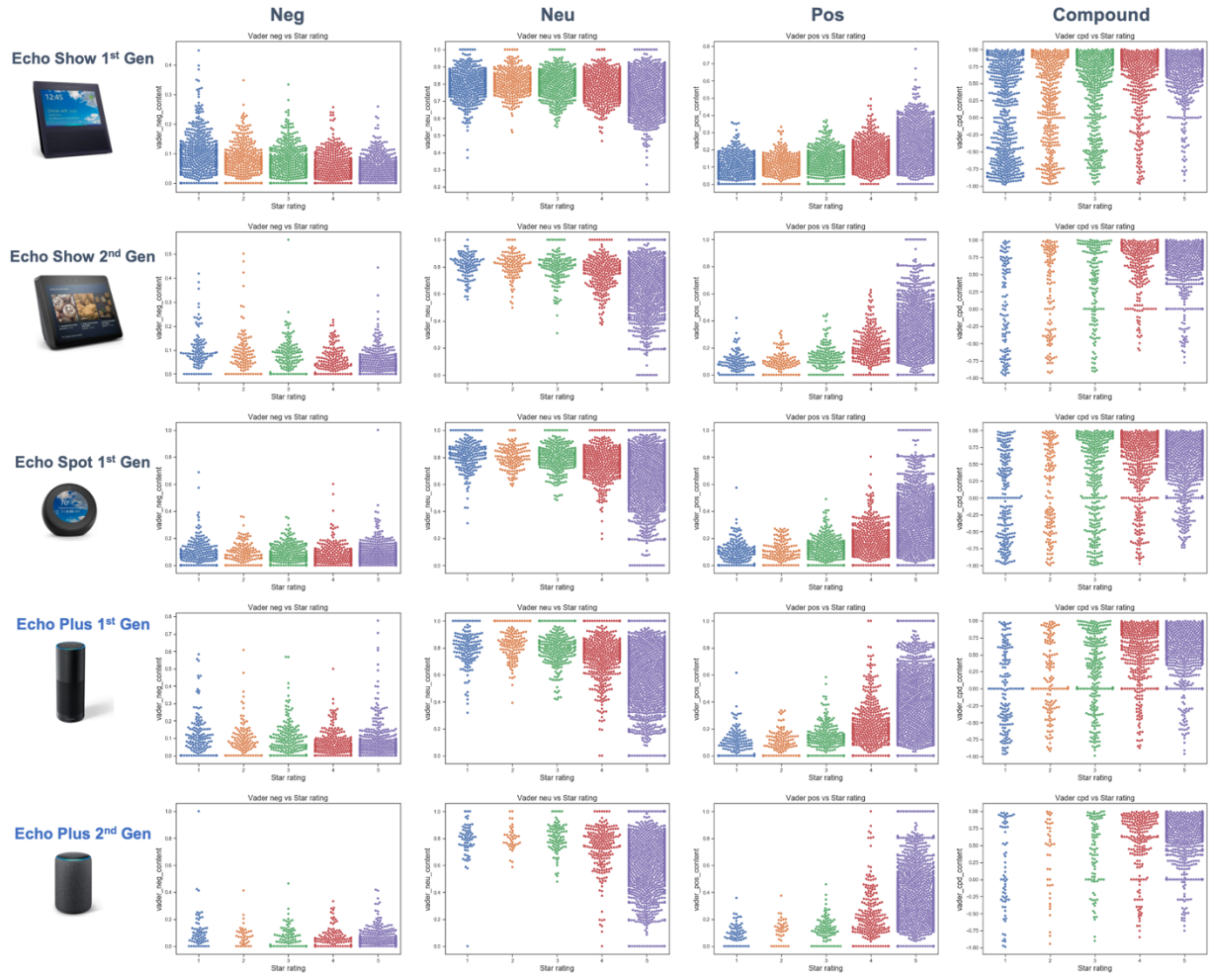

Figure 10. VADER analysis of Echo devices. The VADERS scores (Pos, Neu, Neg, and Compound) for each device are plotted against the review's star ratings. Reviews of 1-, 2-, 3-, 4- and 5-star are colored in blue, orange, green, red, and purple, respectively.

Continuing the study of the relationship between the relationship between review length and star rating, the VADER Compound scores were plotted against the star ratings in Figure 11 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.

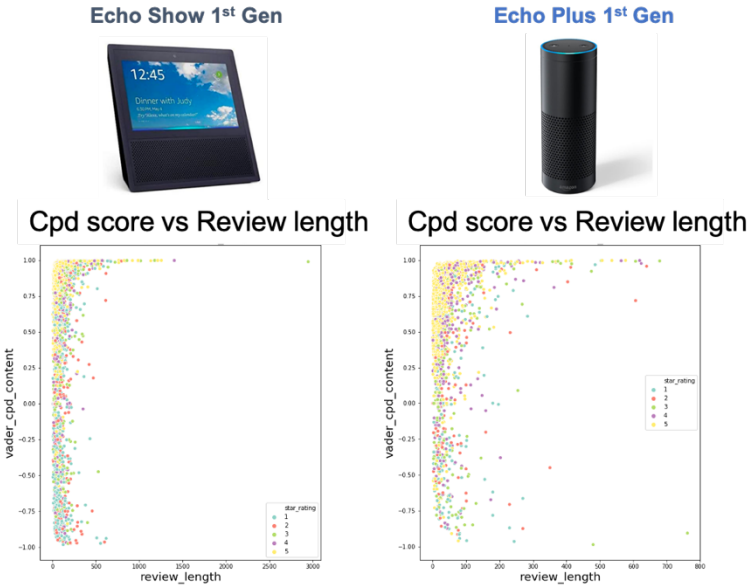

Figure 11. VADER Compound scores vs review length.

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 12, 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.

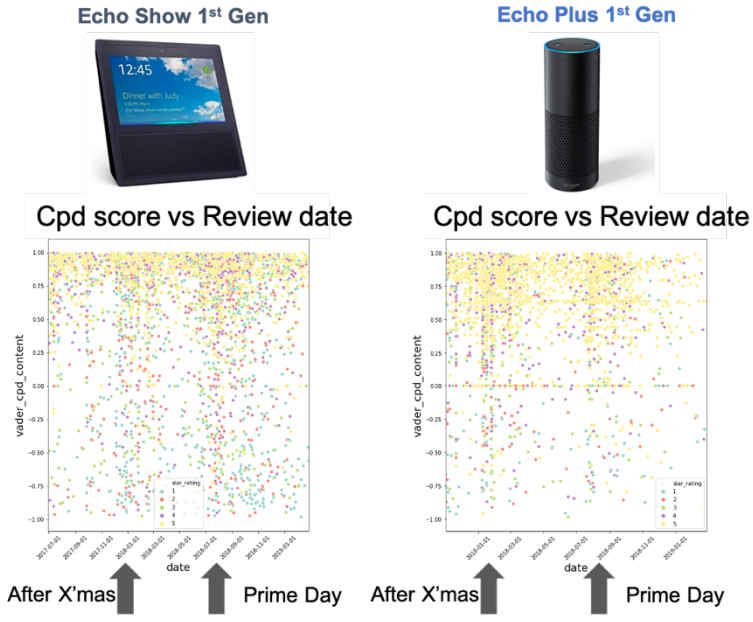

Figure 12. VADER Compound scores vs review date.

Sentiment polarity ratio: scoring sentiments based on Pos and Neg

As shown in Figures 10, 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 11, 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 13. 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 13. 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 14. 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.

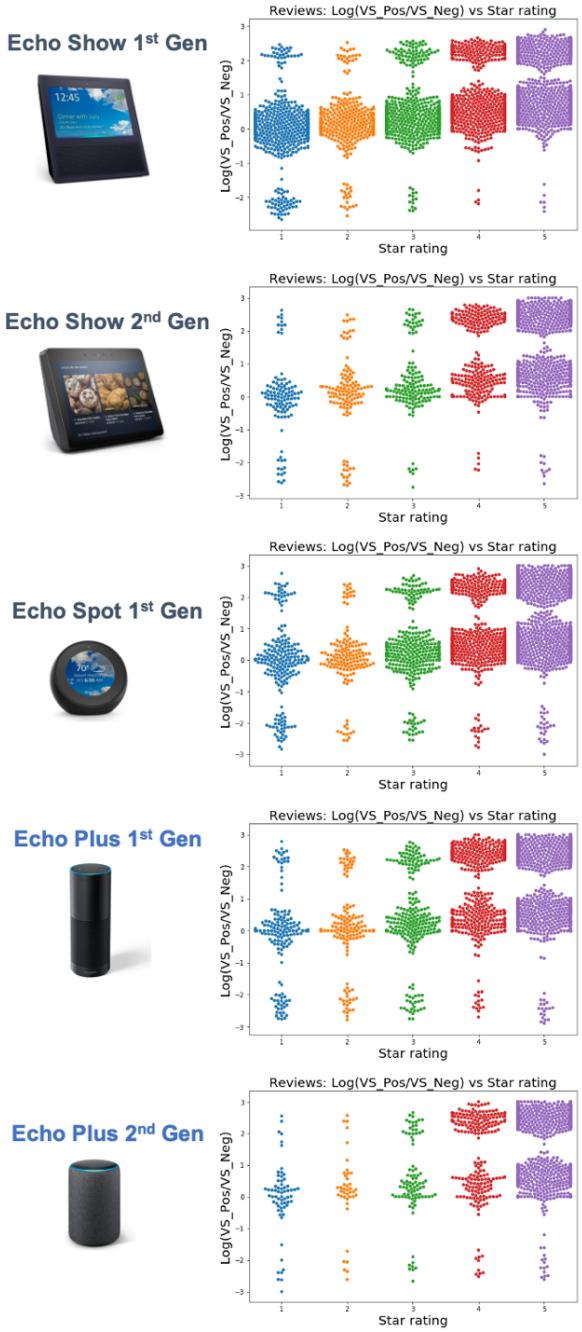

Figure 14. Sentiment polarity ratios for Echo devices.

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 15. 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.

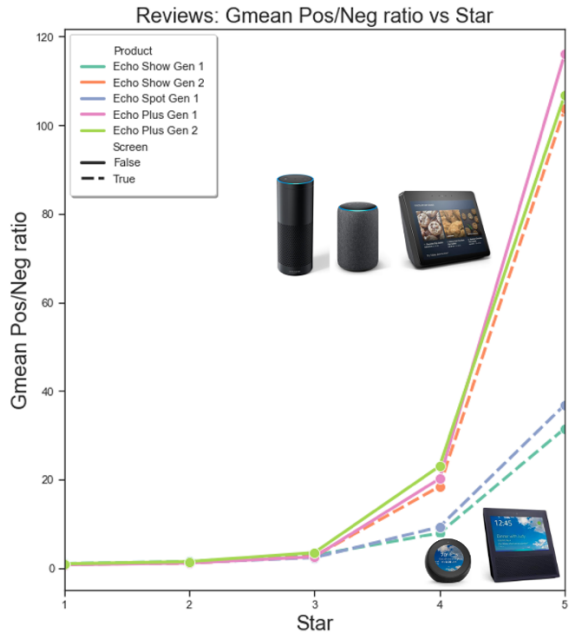

Figure 7. Sentiment profiles based on sentiment polarity ratios and star ratings.

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.

Conclusions

Aiming to study whether Echo device is "better" with a screen, this study examined user experience with a selection of 5 Echo devices with and without a screen based on reviews downloaded from Amazon in early 2019. The initial exploratory data analysis shows that the mean star ratings are similar among the 5 selected devices with the Echo Show 1st Gen has the worst rating at 3.7 star in average. However, the data set is shown to be heavily skewed with a disproportionally high number of good reviews. The review length also appears to correlate with the mean rating. Focusing on analyses that would not be influenced by the skewed distribution and review length, word cloud analyses were performed to identify popular and good features from 5-star reviews and issues from 1-star reviews. VADER analysis was carried out to further study the sentiment expressed in the reviews. While the outputs of the sentiment analysis basically agree with the reported star ratings, some, especially those with lower star ratings, are questionable, such as numerous 1-star reviews being scored very positively. The overall results are also difficult to interpret. Alternatively, focusing on the goods and bads in the reviews, sentiment polarity ratio is devised to model how polarized the sentiment by computing the ratio of the relative amounts of the positive and negative texts. The sentiment polarity ratios, calculated from the VADER outputs, provide a "cleaner" and interpretable sentiment assessment of the current data. By using the geomeans of the polarity ratios, the reviews could be categorized into two groups: a lowering rating group, which includes Echo Show 1st Gen and Echo Spot 1st Gen, 2 of the 3 screen-enabled devices, and a higher rating group, which includes both screenless Echo Plus speakers and the Echo Show 2nd Gen. Interestingly, in conjunction with the word cloud analysis, Echo Show 2nd Gen achieves a better user satisfaction by having better audio-related features, very similar to that of the fellow screenless Echo smart speakers in the group. This finding appears to suggest that users enjoy using Echo more as a smart speaker, and having an additional screen does not automatically make the device "better".