1-2022

Do sequels outperform or disappoint? Insights from an analysis of Amazon echo consumer reviews

Kyong Jin SHIM
Singapore Management University, kjshim@smu.edu.sg

Siaw Ling LO
Singapore Management University, sllo@smu.edu.sg

Su Yee LIEW
Singapore Management University, suyee.liew.2017@sis.smu.edu.sg

Citation

SHIM, Kyong Jin; LO, Siaw Ling; and LIEW, Su Yee. Do sequels outperform or disappoint? Insights from an analysis of Amazon echo consumer reviews. (2022). *2022 55th Hawaii International Conference on System Sciences HICSS: Virtual, January 4-7: Proceedings*. 2950-2959.
Available at: https://ink.library.smu.edu.sg/sis_research/6847

Do Sequels Outperform or Disappoint? Insights from an Analysis of Amazon Echo Consumer Reviews

Kyong Jin Shim
Singapore Management University
kjshim@smu.edu.sg

Siaw Ling Lo
Singapore Management University
sllo@smu.edu.sg

Su Yee Liew
Singapore Management University
suyee.liew.2017@smu.edu.sg

Abstract

Rapid technological advances in recent years drastically transformed our world. Amidst modern technological inventions such as smart phones, smart watches and smart home devices, consumers of electronic digital devices experience greatly improved automation, productivity, and efficiency in conducting routine daily tasks, information searching, shopping as well as finding entertainment. In the last few years, the global smart speaker market has undergone significant growth. As technology continues to advance and smart speakers are equipped with innovative features, the adoption of smart speakers will increase and so will consumer expectations. This research paper presents an aspect-specific sentiment analysis of consumer reviews of the first three generations of Amazon Echo. Our study demonstrates a novel cross-generation visualization of directional changes in consumer sentiment using the Bollinger Bands and volume charts.

1. Introduction

Over the years, technology has rapidly transformed our world and daily lives. Modern technology has paved the way for multi-functional devices such as smart phones, smart watches, and smart home devices. Computing devices are increasingly becoming more portable, faster, high-powered, and more affordable. Further, increased Internet connectivity, improved communication via instant messaging applications, and established infrastructure systems enable rapid information sharing about new products and ideas at speeds never seen before – resulting in the rising speed of technology adoption.

The last few years have witnessed rising adoption of smart home devices [1]. As of 2021, the global penetration rate of smart home devices stands at 12.2% with 349 million smart home device shipments [2]. Market experts forecast that the global penetration rate of smart home devices will reach over 21% by 2025 [3] and continue to rise. Among others, 'smart speakers' such as Amazon Echo products are powered by powerful Artificial Intelligence software programs. From reporting today's top news to finding food recipes and calories, these virtual voice assistant speakers are drastically transforming our home environment by bringing in the automation of routine daily tasks and improving productivity.

Globally, Amazon is dominating the smart speaker market with 28.3% market share as of 2020 – closely followed by Google with 22.6% market share [4]. Since its first launch in 2014, four generations of Amazon Echo devices have entered millions of homes [5, 6]. In China, the three Chinese vendors – Xiaomi, Baidu, and Alibaba – accounted for over 96% of the domestic smart speaker sales volume by early 2020 [7]. As such, today's smart speaker market is much more dynamic with increasing competition and constantly evolving technologies.

As the global adoption of smart speakers continues to rise and more sophisticated AI-assisted features enter smart speakers, consumers' expectations are raised [8]. Sound quality and device portability are important factors for selecting smart speakers [9]. Many smart speaker users use the voice search to find information such as the latest news headlines, weather, traffic, recipes, and many more [10]. For such voice-enabled speakers to engage users in 'natural' communication, it is important for the devices' underlying algorithms to be able to accurately understand the conversation context. In 2019, Google rolled out BERT [11], a significant change to Google's search algorithm for improved understanding of the context behind users' search queries. With improved 'context-aware' natural language processing capabilities, consumers can expect superior voice-enabled user interfaces on smart speakers.

Increased connectivity to the World Wide Web (WWW) in the last two decades has radically changed the way consumers shop. From clothes to medical services, consumer reviews serve as important indicators of the trustworthiness of businesses [12]. Online product reviews have become a vital source of information for supporting customers' purchase

URI: https://hdl.handle.net/10125/79698
978-0-9981331-5-7

decisions [13, 14, 15]. Modern consumers are well-informed, and they consider ratings and reviews to be essential to their shopping experience [16].

This research study explores consumer reviews on three generations of Amazon Echo device between June-2015 and March-2021. Our goal is to inform changes in consumers' expectations with regard to technical aspects and humanized aspects of the Amazon Echo devices. In particular, our research seeks to answer the question: "Do sequels outperform or disappoint?" To answer this question, we applied text mining methods to an aspect-specific sentiment analysis.

2. Related Work

Since mid-2000, the world has witnessed an exponential growth of social media. Today, social media platforms are a major source of information and news for over 3 billion active users. Social media platforms offer countless possibilities in terms of sharing user-generated content such as photos, videos, business or product reviews, etc.

Modern consumers actively refer to business, product and service ratings and consumer reviews [13, 15, 16] for decision making. Over the years, this has resulted in large volumes of consumer reviews over a wide spectrum of businesses, products, and services. Consumer reviews contain rich information for both businesses and individual consumers.

Based on these reviews, sentiment analysis has gained a lot of attention in recent years. In particular, several prior studies have investigated and proposed methods for aspect-specific sentiment analysis. Prior studies analyzed consumer reviews of camera products [17, 18], electronic gadgets [19, 20], hotels [21], and other businesses [22].

A more recent study [23] presented a large-scale sentiment analysis on Amazon Echo consumer reviews. While this prior study compares frequent features across three different smart speaker products, our study investigates aspect-specific sentiment changes across *multiple generations* of Amazon Echo. This study demonstrates that Bollinger Bands and volume charts provide an effective means for visualizing sentiment changes over time with improved interpretability.

3. Dataset

For the collection of consumer reviews, we used the chrome extension 'Web Scraper' [24] to extract publicly available data from Amazon's website. The data collection for this research study focuses on the first three generations of Amazon Echo (referred to as Gen 1, Gen 2, and Gen 3 in the remainder of this paper) between June-2015 and March-2021. All reviews analyzed in our present study are English-written.

The data fields extracted are "username", "location_date", "rating", "title", "review" and "helpfulness". Gen 1 data span from June-2015 to March-2020, Gen 2 data span from October-2017 to March-2021, and Gen 3 data span from October-2019 to March-2021. The main review content comes from the "review" data field, and we further augmented it with the 'title' text which may also include relevant keywords indicative of consumer sentiment. Table 1 shows the total number of reviews and sentences for each Amazon Echo generation broken down by star ratings ranging from 1-star to 5-star.

Table 1. Distribution of the reviews

Rating	Gen 1	Gen 2	Gen 3
5 star	7,779	8,419	10,211
4 star	5,443	5,324	1,384
3 star	3,461	2,808	639
2 star	2,110	1,584	351
1 star	3,279	2,753	729
Total No. Reviews	22,072	20,888	13,314
Total No. Sentences	93,095	61,984	24,496

Table 2. Word count of reviews

Statistic	Gen 1	Gen 2	Gen 3
Mean	84.18	56.57	35.23
Median	46	35	21
Max	2,520	2,927	1,680

Table 2 shows that across all generations, some reviews have very high word counts. Not all product reviews may be fully positive or negative in sentiment. Thus, for topic modeling and sentiment analysis, the review texts are split into one or more sentences.

4. Methodology

This section describes our overall methodology for identifying 'aspects' or product features from a large volume of consumer reviews of Amazon Echo products and performing aspect-specific sentiment analysis. First, we describe how we leverage BERT [25] and Google Natural Language API [26] for sentiment classification. Next, we explain how aspect extraction and prediction are carried out. Lastly, we discuss our aspect-specific sentiment classification method.

4.1. Sentiment analysis

There are many pre-trained embedding approaches that have been released recently. Among others, BERT

[25] is a bi-directional language representation that has achieved state of the art results for many NLP tasks. In our solution, we use a pre-trained BERT model to predict a single sentiment (positive, neutral, or negative) for sentiment analysis. Pre-processing steps performed on the review sentences include expansion of contracted words, normalization (lower-case), removal of special characters, and removal of sentences whose word counts were below five. Stop-words removal and lemmatization are not performed as they might affect the accuracy of the model prediction.

In order to increase the reliability of the sentiment predicted from the pre-trained BERT model (with macro F1-score of 85%), our study leverages additional sentiment prediction using Google Natural Language API [26] on the review sentences. The final sentiment score is taken from the average of the two sentiment scores produced by two different models. The sentiment score ranges between -1 and +1 with -1 being most negative and +1 being most positive. A sentiment score below -0.25 is considered as negative, and a sentiment score above 0.25 is considered as positive.

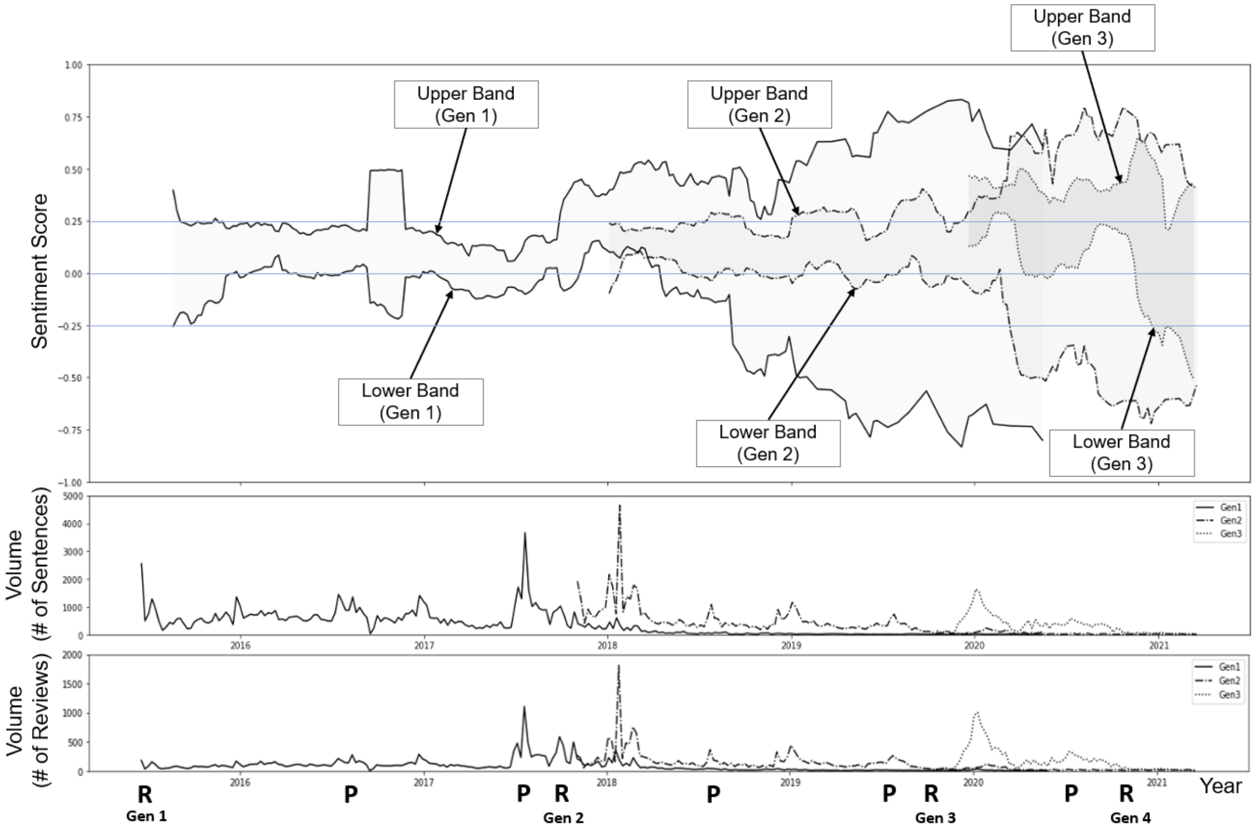

Figure 1. Sentiment and volume change over time (Bollinger Bands & Volume Charts)

Table 3. Percentage distribution of sentiments

Sentiment	Gen 1	Gen 2	Gen 3
Positive	41.7%	43.3%	54.4%
Neutral	28.5%	25.7%	22.9%
Negative	29.9%	31.0%	22.7%

Table 3 shows the distribution of the sentiments for Gen 1, Gen 2, and Gen 3. We observe an increasing trend of positive sentiment and a generally decreasing trend of negative sentiment with each subsequent release of the Amazon Echo device.

Next, we observe how the sentiment changes over time with the release of each new generation. We use Bollinger Bands [27] to visualize the sentiment scores over time along with the changes in 'volume' or the number of review sentences. In finance, Bollinger Bands are technical analysis tools defined by a set of trend lines that are plotted two standard deviations (in a positive direction and in a negative direction) away from a simple moving average of the price of a security.

In our study, we first calculate the average sentiment score over all review sentences on a weekly basis. When we plot these weekly average sentiment scores over time, the chart shows volatility – making it challenging to visualize and observe the underlying 'trend' of the sentiment score changes. Using a Bollinger Bands chart which plots weekly moving averages by connecting the average sentiment scores over a specified period (e.g. 10 weeks), the chart helps us comprehend the underlying movement of the sentiment score (as sentiment does not move only in one direction).

Once we have the weekly average sentiment scores, we use '10 weeks' for period and the default standard deviation of 2 to plot the Bollinger Bands (Figure 1). Figure 1 shows three charts:

1) Bollinger Bands (top) shows the sentiment trend over time. When the upper band and the lower band are far apart, the sentiment *volatility* is high. When the two bands are close together, the *volatility* is low.
2) Volume (middle) shows the total number of review sentences (aggregated weekly) over time. Recall from Section 3 that we perform aspect extraction and sentiment scoring on each review sentence instead of performing it on the entire review content.
3) Volume (bottom) shows the total number of reviews (aggregated weekly) over time

At the bottom of Figure 1, we denote major events concerning Amazon Echo product launches and sales along the X-axis (Year). "R" denotes a product release, and "P" denotes the 'Amazon Prime Day' sales which occurs annually.

Gen 1 was first released and made available to Amazon Prime members in March-2014. Afterwards, it was released to the mass public in June-2015. Figure 1 denotes this latter date with "R" (Gen 1). As for **Gen 1**, the average sentiment score generally stayed well within the 'neutral' range (between -0.25 and +0.25) from June-2015 to around the time of **Gen 2** release in October-2017. Upon **Gen 2** release, the average sentiment score for **Gen 1** started diverging in both directions – positive and negative. In other words, the **Gen 1** sentiment volatility started increasing when its sequel – **Gen 2** – was released. At the same time, the number of **Gen 1** reviews drastically reduced while the number of **Gen 2** started spiking. Our investigation reveals that the **Gen 1** reviews upon **Gen 2** release exhibited high polarity in either positive or negative direction, and this resulted in increasing distance between the two bands downstream.

We observe a similar pattern for **Gen 2**. Upon product release in October-2017, the average sentiment score generally stayed well within the 'neutral' range until around the time of **Gen 3** release in October-2019. The average sentiment score for **Gen 2** started diverging in both directions upon the release of its sequel – **Gen 3** – in October-2019. The number of **Gen 2** reviews drastically reduced, and the number of **Gen 3** reviews started spiking.

The simple moving average for **Gen 3** was generally higher than those of its predecessors. This indicates that **Gen 3** was perceived more positively than its predecessors. However, the overall sentiment score took a negative turn around October-2020 when its sequel **Gen 4** was released. Additionally, the average sentiment score for **Gen 3** started diverging in both directions – positive and negative. This is the same pattern that we observed for two of its predecessors.

Other than the sentiment changes upon each product release, we observe that the volume (number of reviews and number of sentences) tends to *increase* around the Amazon Prime Day sales period. For **Gen 2** and **Gen 3**, it is noticeable that within about two months upon the product release, the volume tends to increase significantly. For example, **Gen 2** was released in October-2017 – and we note that the volume significantly increased in January 2018. A similar observation is made for **Gen 3** (e.g. the volume peaks suddenly towards the end of December-2019 into January-2020).

In summary, the Bollinger Bands chart allows for improved detection of *underlying trends* in terms of sentiment changes over the course of three Amazon Echo product releases. Volume charts are useful in observing whether a large move in either direction of sentiment polarity is substantiated by a large number of consumer reviews. If a large move in a negative direction is substantiated by a large volume (as opposed to a small number of consumer reviews), this perhaps suggests to the business (device manufacturer) to intervene before the situation becomes a critical mass. Section 4.2 details a machine learning method for extracting 'aspects' or product features from a large volume of consumer reviews. With this, the business can efficiently target specific product features that resulted in positive or negative public sentiment.

4.2. Aspect extraction and prediction

Many consumer review sites adopt 'star ratings' that represent the summarized rating for the product. However, not all reviews may be fully positive or negative in sentiment. For example, consider the following review with a 2-star rating for **Gen 2**:

> *"Good speaker. I don't like that it doesn't work with Spotify. Alexa is very useful as alarm and for entertainment. Overall, I'm quite satisfied. However, cloud service could be better."*

From the review text, it can be inferred that the reviewer rates "speaker", "alarm", and "entertainment" *aspects* of **Gen 2** positively. However, he indicates negative sentiment towards 'cloud service' and 'lack of Spotify compatibility' *aspects* of the product. Despite the fact that the number of positively rated aspects outweighs that of negatively rated aspects, the star rating stands at only 2 (out of 5 stars). As such, the star rating can be *misleading* and *omitting* of aspect-specific information to the readers.

While the review text provides rich context around the reviewer's feelings and attitude towards a product, it can be challenging and time-consuming to peruse all reviews manually. Thus, we leverage topic modeling to automate 'aspect' (or product feature) discovery from a large volume of reviews.

Table 4. Review sentences with predicted aspect(s)

Aspect(s)	Review sentence example
Q&A	"i tried to start up several conversations with alexa but she would misunderstand me or tell me she was not sure about that"
Connectivity	"i have been trying to get bluetooth streaming from my samsung galaxy s6 to work for more than 2 months"
Sale support	"a piece of junk and amazon support is even worse"
Voice	"it just does not hear me have done the voice training over and over"
Comparison, Q&A	"google home is better at answering the questions"
Sound, price	"too expensive can get better sound quality with bose jbl speaker"
Companionship	"talking to her all day and love being around her"
Smart home	"used it to control lights, door and thermostat"

Overall, the 'aspect' extraction from the consumer reviews involves two steps: 1) inference from topic modeling results and 2) manual evaluation by human judges fluent in English and with good knowledge of the domain (e.g. smart speakers). Topic modeling is capable of scanning and summarizing a large volume of texts. Manual evaluation by human judges is performed on a subset of the review sentences to add a layer of validation. This subset consists of sentences retrieved from those consumer reviews that have garnered at least 100 'helpfulness' scores. Human judges manually extracted aspects from these review sentences.

Similar text pre-processing steps, as described in Section 4.1, are performed on the review sentences for topic modeling – with additional steps such as stop words removal and lemmatization. Furthermore, only verbs, nouns and adjectives are considered in topic modeling.

A total of 12 aspects are identified: *sound, smart home, connectivity, comparison, voice, price, app, shelf life, question answering (Q&A), companionship, sale support, and none*. We perform multi-class aspect classification to predict the aspect(s) for each review sentence. RoBERTa (robustly optimized BERT approach) model [28] is pre-trained with nearly 10 times more data, and it achieves better results than the original BERT. In this study, RoBERTa model is used to train on 2,000 review sentences with manually labelled (by human judges) aspects. The AUC (Area Under the Receiver Operating Characteristic curve) score on the validation data is 88.6%. The trained model is then used to predict the rest of the unlabeled review sentences. Table 4 shows examples of the review sentences with predicted aspects.

4.3. Aspect-specific sentiment analysis

Table 5 shows the percentage distribution of the aspect-specific sentiments (positive and negative) across Gen 1, Gen 2, and Gen 3. The percentage value indicates the proportion of consumer review sentences where a certain aspect (e.g. sound, connectivity, etc.) was detected by our algorithm. We calculate this percentage for each Amazon Echo generation. For instance, in Table 5, 22% of the 'positive' consumer review sentences for Gen 1 are found to be containing 'sound' aspect in the texts. And, 12.6% of the 'negative' consumer review sentences for Gen 2 are found to be containing 'voice' aspect in the texts.

The upwards pointing arrow (↑) indicates a generally increasing trend in the percentage (of positive or negative review sentences) from the predecessor to the sequel. The downwards pointing arrow (↓) indicates a generally decreasing trend in the percentage (of positive or negative review sentences).

Generally, the aspects can be grouped into four categories. The first category (↑↑) consists of aspects with an increasing trend for both positive and negative sentiments, and we call this category "trending". The second category (↓↑) has decreasing positive trend and increasing negative trend, and we categorize this as "disappointing". Third category (↑↓) consists of aspects with increasing positive trend and decreasing negative trend, and we call this "outperforming" category. The last category (↓↓) consists of aspects with decreasing trend for both sentiments, and we call this "sunsetting" category. Lastly, the aspect "sale support" does not have significant changes in the trend and "None" refers to sentences with no aspects detected by our algorithm.

As shown in Table 5, the aspects that are 'trending' are mainly the technical features of the Amazon Echo devices such as sound, smart home, and connectivity. The aspects that are 'sunsetting' are humanized functionalities such as Q&A and companionship. Detailed analyses and insights are covered in the next section.

5. Insights

5.1 Trending

Our analysis results reveal four aspects of the Amazon Echo devices that appear to be ‘trending’. They are *sound, smart home, connectivity, and comparison*. These aspects are termed ‘trending’ because the percentage of review sentences mentioning these aspects continues to increase from one generation to the next. This coincides with the observation that the percentage of review sentences mentioning aspects from other categories reduces from Gen 1 to Gen 3.

Table 5. Percentage distribution of aspect-specific sentiments across Gen 1, Gen 2, and Gen 3

	Category	% Positive Sentences			% Negative Sentences		
		Gen 1	Gen 2	Gen 3	Gen 1	Gen 2	Gen 3
(↑↑) sound	*Trending*	22.0%	33.7%	41.1%	11.8%	19.2%	22.6%
(↑↑) smart home		9.0%	9.1%	9.8%	3.9%	3.9%	4.8%
(↑↑) connectivity		6.9%	7.9%	9.9%	12.8%	16.1%	17.3%
(↑↑) comparison		4.4%	9.0%	10.2%	3.7%	6.5%	6.5%
(↓↑) voice	*Disappointing*	8.7%	7.1%	7.0%	12.7%	12.6%	14.8%
(↑↓) price	*Outperforming*	3.7%	3.6%	4.3%	6.3%	4.2%	4.6%
(↓↓) app	*Sunsetting*	12.4%	8.3%	5.2%	10.4%	8.9%	7.7%
(↓↓) Q&A		9.9%	8.0%	4.9%	14.1%	11.1%	8.6%
(↓↓) companionship		5.2%	4.2%	3.7%	0.7%	0.5%	0.3%
(↓↓) shelf life		2.1%	1.8%	1.7%	9.2%	8.4%	8.6%
sale support		1.4%	1.5%	0.8%	7.5%	8.0%	7.8%
None		33.3%	28.0%	25.9%	23.9%	20.7%	19.9%

For smart speakers, ‘sound’ is an important feature significantly impacting consumers’ music and news listening experience. As shown in Table 5, across all three generations of Amazon Echo, ‘sound’ is the most talked about aspect. Consumers’ experience with regard to the ‘sound’ aspect is rather mixed. Positive reviews are observed:

> *“can honestly say in many areas the sound quality is now actually better than gen 1 echo”* **(Gen 2)**

> *“went thru a range of music streamed over bluetooth switching between gen 1 and gen 2 devices and found gen 2 to be an improvement over gen 1 for the vast majority of the music i tried”* **(Gen 2)**

> *“the 3rd gen ... the sound quality is much better than the previous gen”* **(Gen 3)**

In contrast, we observe negative reviews:

> *“gen 2 sounds like a loud phone in a shoebox”* **(Gen 2)**

> *“after comparing the sounds between gen 1 and gen 2 it was obvious to me that gen 1 was superior”* **(Gen 2)**

> *“audio quality for the alexa voice is still pretty muddy compared to the echo gen 1”* **(Gen 2)**

As evident in the above consumer reviews, from the consumer reviews on Gen 2 and Gen 3, we observe that those consumers having prior experience with one or more predecessors tend to compare the sequel with its predecessor(s). We assign ‘comparison’ aspect to such reviews.

Another trending aspect is ‘smart home’ which concerns smart speakers’ ability to control light, television, speaker, and other smart home devices. This aspect is highly linked to ‘connectivity’ aspect. Positive consumer reviews abound:

> *“a nice choice to start your connected home life”*

> *“.. projector plugged in.. and it works perfectly”*

> *“intercom function has been helpful twice already 10 days of use and the sound is impressive”*

> *“works fantastic on my enclosed porch and syncs well with my other alexa devices”*

> *“i love it to communicate with other rooms”*

> *“nice stereo sound with 2 echo speakers paired together sound is 360 degrees and fills entire room”*

> *“i now have a six speaker setup and it is like going to a rock concert every time i put music on”*

> *“the 3rd gen ... is a great value especially if you pair it with fire 4k”*

However, there are reviews reflecting negative experience with regard to ‘smart home’ aspect of Amazon Echo devices:

> *“i know last year’s models had some connectivity issues with philips hue bulbs”*
>
> *“works good for light automation but does not connect to pandora or amazons”*
>
> *“its great for apartments but if you want to add it to a big house its not worth it”*
>
> *“bought two led bulbs for the front room they were a pain to install from that company but they got us through it and echo turns onoff and dims the lamps on command”*
>
> *“voice sync issues with samsung tv”*
>
> *“we had a lot of trouble getting her to connect to spotify and apple music instead of amazon music"*

In summary, with regard to ‘smart home’ and ‘connectivity’ aspects, consumer reviews reflect mixed sentiment. For a home to be ‘smart’, connectivity amongst multiple appliances and devices must be easy to set up and configure. While Amazon Echo devices interface smoothly with certain appliances and devices, consumers experience the opposite when they attempt to connect Amazon Echo devices with other appliances and devices. As shown above, the reviews often mention specific brands, products, and features that work well or poorly with Amazon Echo devices. Such insights will be useful to Amazon in understanding consumers’ pain points.

5.2 Disappointing

According to an earlier research study [23], voice recognition and understanding are some of the most often complained aspects mentioned in Amazon Echo consumer reviews. We observe the same sentiment in our analysis. ‘voice’ is a ‘disappointing’ aspect with positive sentiment exhibiting a downward trend and an upward trend for negative sentiment from Gen 1 to Gen 3. Based on the top keywords extracted from topic modeling, there is a decreasing satisfaction with regard to the voice recognition and command feature. However, consumers are generally positive with the recognition of commands issued. It is the understanding and hearing of the commands that upset the consumers. This is likely because consumers treat Alexa (the AI service that powers Amazon Echo devices) as a *personal assistant* and expect Alexa to understand and deliver simple command just as human assistants would do. Unfortunately, the ability of Alexa has not met the expectation according to the consumer reviews:

> “*i do not like that it is not retaining the recognition of my voice*”
>
> “*it has been unable to understand the word.. and i cannot let a poorly design voice recognition software into my high tech world”*

Since Amazon Echo is an evolving product with a new sequel expected to be released bi-annually, it is possible for consumers to set higher expectation for the sequel and compare it to its predecessor [29]:

> “*not as good at voice recognition as my first generation echo*”

Consumers of smart speakers equipped with conversational capabilities expect to interact with the devices to perform simple tasks such as playing music or switching on a light via voice commands. When smart speakers fail to perform these tasks well, it frustrates consumers – and it is reflected in this review:

> “*then alexa kept saying i do not understand try again later*”

It is important to assess if there are alternate ways to receive inputs from the device users when Alexa is unable to understand human voice commands. Consumer expectations will further increase as the human machine interaction has evolved to the level where users expect human-like responses and empathy. If conversational agents such as Alexa are able to signal an expression of understanding with cognitive empathy and back-channel responses including “um-hmms” to induce perceptions of listening [30], it may improve the rapport and experience between the device users and the agent.

5.3 Outperforming

Our analysis results reveal that price is an ‘outperforming’ aspect - with an upward positive trend and a downward negative trend from Gen 1 to Gen 3. This phenomenon can be explained by the fact that sequels are priced lower with improved features and functionalities. It is reflected in consumer reviews:

> “*it is accurate excellent sound and lower in price than its previous model*”
>
> “*not only the sound quality which is as good or better than anything out there for twice the price but for the convenience of being able to call up any song by speaking your request*”

Some of the consumer reviews discuss the worth of purchase during sales such as Amazon Prime Day and Black Friday. However, these does not mean customer is satisfied with the pricing since the volume of the negative sentiment is slightly more than the positive sentiment. For Amazon Echo devices, the complementary services of the Amazon Prime membership may play a role in the satisfaction of consumers. A consumer review reflects such satisfaction:

“*we also signed up for the amazon family music unlimited to share with family members a great deal with access to 60 million songs for just a few dollars a month per device*”

It is important to note that the sequel must live up to the new and improved features (as advertised) since some consumers may have paid more expecting better quality and features than other devices:

“*the sound quality in voice recognition are not good enough to justify its price tag being more than double that of the echo dot*”

“*not worth the extra expense for what i used it for*”

5.4 Sunsetting

We call this category ‘sunsetting’ because with each sequel launch, the amount of consumer reviews (both positive and negative) mentioning this category’s aspects decreases.

As for ‘app’, consumers are glad to use the phone app for shopping, timer, alarm, and listening to Amazon Prime music as reflected in these reviews:

“*we use them for our 6 wifi outlets as well and they work great through the alexa app or voice command*”

“*it works well with my amazon music and i am sure it will work well with all the other apps that it offers*”

However, Amazon may consider inter-operability with other products such as Spotify and Apple Music. A consumer review reflects this sentiment:

“*it is not compatible with spotify which is very upsetting*”

“Companionship” is an increasingly important feature for conversational agents. A recent review on factors influencing users’ adoption and use of conversational agents stated that the ability to employ human-like cues and communication modalities was found to significantly influence the intention to use a conversational agent especially among the elderly users [31]. In addition, another analysis on Amazon Echo customer reviews shows that some 30% of consumers would treat Amazon Echo as a human character because of its personified name and its ability to converse. Some even address Echo as best friend, girlfriend, wife, or family [23]. Using the keywords from a prior study [23], we further inspected the consumer reviews in our dataset. Our analysis results reveal that consumers did not really develop a closer relationship with Amazon Echo devices as observed in [23]. Instead, it is mainly about meeting the needs of loved ones - for example:

“*upgraded our original because my wife likes to listen to music*”

“*i purchased a bose 500 with alexa because that wife wanted better audio than the echo dot*”

However, it is common to find consumers relate to the Amazon Echo as a person and address it as ‘she’:

“*she definitely can hear you even at loud volumes better than the 2nd generation*”

“*i love my alexa more than my google however sometimes she does not listen very well*”

“*we have had to get used to each other because she is much more computer responsive than siri which i was accustomed to*”

It is interesting to observe that the keywords such as “joke”, “fun”, “kid” and “life” are seen in the consumer reviews:

“*our granddaughter enjoys playing games with it and hearing jokes*”

“*i love her jokes and tell her good morning for a pleasant response to start the day she is amazing*”

Despite the presence of positive reviews concerning “companionship” aspect, it is believed that consumers’ expectations have elevated over the years due to various options available such as Siri, Google Home. The *new* and *cool* features from a year or so ago are no longer of novelty and not worthy of writing another positive review about.

The next two aspects, under this category, are “Q&A” and “shelf life”. Even though the negative sentiment is having a downward trend, there are more negative sentences than positive sentences. Thus, we focus on analyzing the negative sentences for both aspects.

Most reviews with “Q&A” aspect are about the inability of Alexa to understand simple requests or questions:

“*this thing gets hung up on simple request all the time*”

“*stupid alexa misunderstands commands and questions all the time does lots of wrong things* ”

The negative sentiment of “Q&A” aspect may have direct relationship with ‘voice’ aspect since some consumers share that they need to repeat commands to get Alexa to understand it:

“*i do not enjoy repeating a command*”

“*it does not know simple answers sometimes it has a hard time hearing me sometimes and i have to yell for it to hear me*”

As for “shelf life” aspect, consumers are generally positive about the battery life. However, there are quite a few negative reviews on defective products:

“*it worked for a month and then quit*”

“*i had a bad experience with this device it came defective i spent a lot of work to install it*”

Besides that, there are negative reviews about the devices due to issues experienced after the one year warranty period is over:

“*it died three months after the warranty expired*”

“*be warned this device will quit working as soon as warranty expires*”

It is worthwhile to note that some consumers make use of the replacement service and are glad that such an option is available. However, some replacements had the same issues – resulting in frustrated consumers:

“*ordered one and the pins were not lined up so it would not charge sent for a replacement same issue*”

While every company strives to produce defect-free products, defects can arise sometimes due to unforeseen circumstances. Given this, the key concern for the business is to manage consumers' expectations and ensure that sufficient support is allocated to address the issue in an efficient and timely manner.

6. Discussions

Our study analyzed sentiment changes with each 'sequel' product release. To attain deeper insights on which 'aspects' are perceived positively or negatively, we used topic modeling to extract aspects from the review sentences. Next, sentiment scoring was performed on each review sentence. We tabulated aspect-specific sentiment analysis results – positive and negative – for Gen 1, Gen 2, and Gen 3 (Table 5). Lastly, we analyzed sentiment changes and grouped aspects into four categories: 1) trending, 2) disappointing, 3) outperforming, and 4) sunsetting.

Even though there are two types of aspect-specific sentiment analysis, implicit and explicit [32], our study did not differentiate between the two types. Instead, we rely on the RoBERTa model [28] to assign a suitable label or aspect to the review sentence. Even with the advancements of the recent approaches in NLP such as the pre-trained models [25, 28], the linguistic complexity of user-generated content remains a challenge. It is often characterized by highly expressive tokens such as emoticons, localized lingual, misspellings, abbreviations, and the use of sarcasm and metaphor. Thus, aspect extraction and sentiment analysis can be quite challenging, and it usually does not work well with tools trained using standard texts [33]. Even though we made use of pre-trained models in this study, we attempted to fine-tune the models by having human judges perform manual inspection of the review sentences and the aspect labels output by the models.

In summary, to the question "Do sequels outperform or disappoint?", our text mining approach can perform an aspect-specific sentiment analysis and show which 'aspects' or product features outperform or disappoint. Our method can discover which aspects are trending or frequently mentioned in consumer reviews. It can also discover which aspects are sunsetting *(opposite of trending)*. As the concept of voice commerce is continuing to gain immense popularity globally and as smart speakers are increasingly being equipped with new and advanced features and functionalities, smart speaker manufacturers can benefit from being able to automatically mine and derive useful insights into consumer sentiment from reviews.

7. Conclusion, Limitations, and Future Research Directions

This study presents an aspect-specific sentiment analysis of consumer reviews of Amazon Echo and demonstrates a novel cross-generation visualization of directional changes in consumer sentiment using the Bollinger Bands and volume charts. We have analyzed consumer reviews without considering different consumer segments (e.g., elderly users, specific app users, etc.) that may interact with smart home devices differently based on their age, gender, preferences, and lifestyles. A more targeted study which takes into consideration different consumer 'profiles' is expected to produce useful insights about how different generations of a product are perceived. While such information as reviewer's gender and age may not necessarily be available, demographics can be *inferred* from the text [34, 35].

Recently, jointly trained aspect detection and sentiment analysis or multi-task learning have achieved promising results [36]. Multi-task learning offers advantages such as reduced complexity (i.e., single model for two tasks) and reduced overfitting through shared representation, which may help to improve the performance of all the tasks involved [37]. Furthermore, it will be beneficial to extend the study with other disciplines, such as marketing or product design, to further analyze the content of the reviews towards designing personal assistant products that truly can meet the needs of users.

8. References

[1] J. J. Ikoba, "A new survey shows a surge in the adoption of smart home devices", GIZMOCHINA, 7 Jan. 2021, www.gizmochina.com/2021/01/07/new-survey-surge-adoption-smart-home-devices/

[2] A. Holst, "Smart home - Statistics & Facts", statistica, 19 Apr. 2021, www.statista.com/topics/2430/smart-homes/

[3] J. Lasquety-Reyes, "Number of Smart Homes forecast worldwide until 2025", statistica, 11 Nov. 2020, www.statista.com/forecasts/887613/number-of-smart-homes-in-the-smart-home-market-worldwide

[4] L. S. Vailshery, "Global smart speaker market share 2016-2020, by vendor", statistica, 7 Apr. 2021, www.statista.com/statistics/792604/worldwide-smart-speaker-market-share

[5] D. Bohn, "Exclusive: Amazon Says 100 Million Alexa Devices Have Been Sold", The Verge, 4 Jan. 2019, www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp

[6] P. Smyth, "Amazon Echo dominates smart speaker market", LondonLovesBusiness, 11 Feb. 2021, londonlovesbusiness.com/amazon-echo-dominates-smart-speaker-market

[7] Statistica Research Department "Market share of smart speaker brands in China Q1 2020, based on sales volume", statistica, 6 May 2021, www.statista.com/statistics/1197847/china-sales-volume-share-of-smart-speaker-manufacturers

[8] C. Delligatti, "Consumer Expectations And The Smart Speaker Assistant", CupertinoTimes, 5 Mar. 2018, cupertinotimes.com/consumer-expectations-smart-speaker

[9] "What do Consumers Expect From Smart Speakers?", Duoyinfo's Blog, 25 Dec. 2019, www.duoyinfo.com/blog/what-do-consumers-expect-from-smart-speakers_b0019.html

[10] "5 Ways Consumers Interact With Smart Speakers", Apr. 2018, mindstreammediagroup.com/introduction-smart-speakers-voice-search-brand-advertisers

[11] P. Nayak, "Understanding searches better than ever before", Google, 25 Oct. 2019, blog.google/products/search/search-language-understanding-bert

[12] S. Utz, P. Kerkhof, and J. Bos, "Consumers rule: How consumer reviews influence perceived trustworthiness of online stores", Electronic Commerce Research and Applications, 2011. 10.1016/j.elerap.2011.07.010.

[13] "Local Consumer Review Survey 2020", BrightLocal, 9 Dec. 2020, www.brightlocal.com/research/local-consumer-review-survey

[14] G. Cui, H.-K. Lui, and X. Guo, "The Effect of Online Consumer Reviews on New Product Sales", International Journal of Electronic

[15] D. Kaemingk, "Online reviews statistics to know in 2021", Qualtrics, 30 Oct. 2020, www.qualtrics.com/blog/online-review-stats

Commerce. 17. 39-58, 2012. 10.2307/41739503.

[16] "The Deloitte Consumer Review. The growing power of consumers.", Deloitte, 2014, www2.deloitte.com/content/dam/Deloitte/uk/Documents/consumer-business/consumer-review-8-the-growing-power-of-consumers.pdf

[17] M. Hu and B. Liu, "Mining opinion features in customer reviews," AAAI, vol. 4, no. 4, 2004, pp. 755–760.

[18] M. A. Shahkhali, F. Ahmadi-Abkenari, "Sentiment Mining on Products Features based on Part of Speech Tagging Approach", International Journal of Computer Science & Network Solutions, vol. 3, no. 12, pp. 1-12, 2015.

[19] S. Mukherjee, and P. Bhattacharyya, "Feature Specific Sentiment Analysis for Product Reviews", 2013. 7181. 10.1007/978-3-642-28604-9_39.

[20] K. Srividya, A. M. Sowjanya, "Aspect Based Sentiment Analysis using POS Tagging and TFIDF ", International Journal of Engineering and Advanced Technology, vol. 8, no. 6, pp. 1960-1963, 2019.

[21] M. S. Akhtar, D. Gupta, A. Ekbal, P. Bhattacharyya, "Feature selection and ensemble construction: a two-step method for aspect based sentiment analysis", Knowledge-Based System, vol. 125, pp. 116-135, 2017.

[22] K. Bhattacharjee, and L. Petzold, "What Drives Consumer Choices? Mining Aspects and Opinions on Large Scale Review Data Using Distributed Representation of Words", 2016, pp. 908-915. 10.1109/ICDMW.2016.0133.

[23] Y. Gao, Z. Pan, H. Wang, and G. Chen, "Alexa, My Love: Analyzing Reviews of Amazon Echo," in IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SC, 2018, pp. 372–380.

[24] "Making web data extraction easy and accessible for everyone", WebScraper, webscraper.io. Accessed 9 June 2021.

[25] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL-HLT, 2019.

[26] "Sentiment Analysis Tutorial", Google Cloud, 8 Jun. 2021, cloud.google.com/natural-language/docs/sentiment-tutorial

[27] J. Bollinger, "Bollinger bands", Bollinger Bands, www.bollingerbands.com/bollinger-bands

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", 2019, ArXiv, abs/1907.11692.

[29] M. Toor, "Customer expectations: 7 Types all exceptional researchers must understand", Qualtrics, 3 Dec. 2020, www.qualtrics.com/blog/customer-expectations

[30] M. M. E. Van Pinxteren, M. Pluymaekers, and J. G. A. M. Lemmink, "Human-like communication in conversational agents: a literature review and research agenda," Journal of Service Management, vol. 31, no. 2. 2020, doi: 10.1108/JOSM-06-2019-0175.

[31] E. C. Ling, I. Tussyadiah, A. Tuomi, J. Stienmetz, and A. Ioannou, "Factors influencing users' adoption and use of conversational agents: A systematic review," Psychol. Mark., pp. 1–21, 2021, doi: 10.1002/mar.21491.

[32] M. Hu and B. Liu, "Mining and summarizing customer reviews," in KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177, doi: 10.1145/1014052.1014073.

[33] O. De Clercq, "The many aspects of fine-grained sentiment analysis: An overview of the task and its main challenges," in Proceedings of HUSO 2016, The Second International Conference on Human and Social Analytics, 2016, pp. 23–28.

[34] K. Santosh, R. Bansal, M. Shekhar, and V. Varma, "Author Profiling: Predicting Age and Gender from Blogs", Notebook for PAN at CLEF, 2013.

[35] J. Marquardt, G. Farnadi, G., Vasudevan, M. Moens, S. Davalos, A. Teredesai, and M. D. Cock, "Age and Gender Identification in Social Media", CLEF, 2014.

[36] M. S. Akhtar, T. Garg, and A. Ekbal, "Multi-task learning for aspect term extraction and aspect sentiment classification," Neurocomputing, vol. 398, pp. 247–256, 2020, doi: 10.1016/j.neucom.2020.02.093.

[37] Y. Zhang and Q. Yang, "A Survey on Multi-Task Learning," IEEE Trans. Knowl. Data Eng., 2021, doi: 10.1109/TKDE.2021.3070203.