Automated Brand & Product Perception Discovery at Scale from Online Reviews with Topical Phrase Mining.

Jason Knaut
University of Illinois at Urbana-Champaign
jknaut2@illinois.edu

ABSTRACT

Online reviews provide a real-time window into consumers' perceptions of the products they purchase at an unprecedented scale. In this paper, I present a novel framework for the unsupervised extraction of clusters of product-related phrases that capture a product category's most important attributes in the eyes of consumers. The framework combines topical phrase mining and opinion mining techniques to synthesize a robust product perception data set that generalizes to the diverse category, brand, and product-level perception analyses of market researchers, product managers, investors, and others.

Keywords

product attribute extraction, product reviews, opinion mining, text analysis, topic modeling

1. INTRODUCTION

Traditionally, market researchers relied on surveys to capture consumer perceptions and sentiment about brands and products. These methods suffer from a few major challenges. First, surveys are very expensive and time-consuming to prepare and to conduct at scale. Secondly, survey results may be easily biased by factors such as unrepresentative consumer samples and improper question wording. Another limitation of the survey approach is in measuring perception trends over time. People are not very good at answering questions like "what did you think of the version of our product two years ago." One solution to this is to conduct ongoing surveys. Again, due to the expense, it is unrealistic that such an approach scales in a cost-efficient manner.

The explosion in publicly available consumer opinions across social media and online reviews presents an opportunity to answer valuable business, marketing, and product questions, almost in real-time, and at a small fraction of the expense incurred with the traditional survey and focus group approaches. However, making use of the volumes of unstructured text reviews requires methods for automating the process of transforming these raw text reviews into a form that can be analyzed similar to quantitative and structured data.

In this paper, I propose a novel framework for transforming online reviews and related metadata into a robust tabular product perception data set. Whereas other research has used rule-based phrase extraction methods (e.g., part-of-speech rules), I aim to incorporate automated topical phrase mining to discover not only the underlying product-related phrases but also to cluster those phrases into product attributes that largely mirror what a human expert would construct based on product category knowledge.

The remainder of the paper is organized as follows. Section 2 describes the mobile phone review data set utilized to demonstrate the practical value of the contributions. Section 3 describes the end-to-end framework for extracting product attributes and mining opinions from reviews. Section 4 explores generating valuable category, brand, and product-level insights by analyzing the product perception data set. Section 5 concludes with a summary of the approach and provides directions for future work.

2. DATA SET

2.1 Overview of the Amazon Reviews Data

As the world's second largest retailer [4], Amazon has amassed an enormous collection of product feedback and opinions from customers in the form of reviews and ratings. Over 223 million of those reviews created between 1996 and 2018 have been made publicly available for research [5] ("Amazon 2018 Data Set")[1].

In this paper, I explore the proposed framework using a subset of the Amazon 2018 Data Set focused on the mobile phone product category. [4] prepared per-category data sets. The data relevant to mobile phone reviews are the `reviews` and `metadata` data sets for the Cell Phones & Accessories category. The Cell Phones & Accessories data set contains 10 million reviews across 590,000 product SKUs.

The `reviews` data includes the raw text review as well as related information such as star rating, date of review, product SKU ("Amazon Standard Identification Number" or "ASIN"), and more. The `metadata` includes brand name, product descriptions, price, categories (sub-categories of the parent Cell Phones & Accessories category), and more. I discuss further preprocessing of the data set in Section 4.

2.2 Real vs. Fake Reviews

One risk of relying on online reviews for seemingly objective analysis is the bias introduced by fake reviews. In early 2015, Amazon acknowledged it had a fake review problem when the company filed its first lawsuit against people charging for product reviews [1]. In October 2015, Amazon filed a lawsuit against more than 1,000 people using the Fiverr marketplace to offer positive product reviews for a nominal fee [8]. While Amazon historically claimed that less than

[1]Amazon 2018 Data Set is available at https://nijianmo.github.io/amazon/index.html

1% of reviews on its site are fake [6], other third-party companies have presented estimates closer to 30% [7].

This is an important sidebar in the context of the proposed framework and following analysis as conclusions should be drawn with caution if potentially biased by the presence of fake reviews. This is exceptionally critical when fake reviews are present disproportionately across brands or time.

A cursory analysis of the Amazon mobile phone reviews data indicates the 2014 to 2015 period likely contains an anomalous number of fake reviews compared to pre- and post-periods. Both the average ratings and the distribution of five-star reviews jumped in 2014 and 2015—coinciding with increased public visibility around Amazon's fake reviews–before reverting to normalized levels. Five-star reviews accounted for 55% of all reviews created in 2014 and 2015 compared to 43% and 46% for the 2011-2013 and 2016-2018 periods, respectively.

3. FRAMEWORK & METHODOLOGY

This section describes a framework for transforming `reviews` and `metadata` into the robust tabular product perception data set that generalizes to broad category, brand, and product-level analysis. I discuss each of the major components in turn below.

3.1 Preprocessor

The Preprocessor performs four primary functions on the raw input data sets. First, the overall data set is filtered to only include mobile phones. This is accomplished by retaining only the products where the `categories` array in `metadata` contains `'cell phone'`.

Next, relevant `metadata` fields—namely `brand` and `product`—are joined with the `reviews` data. Other `metadata` fields could be valuable depending on the desired analysis. Adding those fields would be a straightforward extension of this process.

Thirdly, the Preprocessor filters the metadata-enriched `reviews` to only include the most popular mobile phone brands (i.e. frequently reviewed). The threshold is set at a minimum of 400 reviews across the 2003 to 2018 time period.

Finally, the Preprocessor stores the metadata-enriched reviews data set for consumption by downstream framework components. Streamlining the input data will dramatically reduce the memory footprint of those downstream components. The Preprocessor also prepares a second data set that only contains the relevant review text, with each line containing the text for a single review. This output will be provided as input to the Product Phrase Extractor described in the next subsection.

Applying the Preprocessor to the Cell Phone & Accessories reviews data set yields 489,857 reviews across 53 brands and 7,514 product SKUs. See Table 1 for statistics on the most frequently reviewed brands.

3.2 Product Phrase Extractor

Perhaps most critical to the framework is the Product Phrase Extractor. The Product Phase Extractor is tasked with automatically detecting not only relevant product attribute phrases but also clusters of phrases that mirror the product dimensions a human expert would conclude holistically represent dimensions of value to consumers.

Brand	Product SKUs	# Reviews	% Total
Samsung	1,624	119,196	24%
Blu	447	57,221	12%
LG	827	54,526	11%
Apple	353	49,579	10%
Motorola	686	41,442	8%
Nokia	577	34,457	7%
HTC	475	21,944	4%
Blackberry	394	21,490	4%
Sony	227	10,850	2%
Huawei	267	9,238	2%
Other	1,637	69,914	14%

Table 1: Most Frequently Reviewed Mobile Phone Brands

This component employs unsupervised topical phrase mining—specifically the ToPMine algorithm [2]. ToPMine is a computationally efficient method that provides a unified framework for both novel phrase extraction as well as phrase clustering—two critical challenges for reliably automating product attribute extraction.

I incorporated the ToPMine implementation created by the author of [2] [2]. Only minor modifications to the `run.sh` script were necessary to produce outstanding results:

- `inputFile =` review-only text from Preprocessor;
- `minsup=10`;
- `minsup=10`;
- `maxPattern=3`;
- `numTopics=30`;
- `thresh=20`;
- `gibbsSamplingIterations=700`.

The ToPMine toolset successfully produced 30 coherent topics. Sixteen of the topics distinctly aligned to either core product attributes (e.g., screen, apps, etc.) or auxiliary product attributes (e.g., customer service, carrier plans) one would associate with the mobile phone product category. A summary of the product attribute clusters is shown in Table 2. Table 3 shows the top phrases extracted for a subset of these product attributes. The other fourteen topics captured more general statements (e.g., device names, "good/great phone", "waste of money/return").

#	Product Attribute	#	Product Attribute
1	buttons	9	chargers
2	sim/sd card	10	form factor/case
3	call/text	11	screen
4	apps	12	user friendly
5	camera	13	carrier
6	call quality	14	processor speed
7	entertainment	15	battery
8	customer service	16	web/email/contacts

Table 2: Product Attributes Generated by ToPMine

[2]The ToPMine implementation is available at http://elkishk2.web.engr.illinois.edu/code/ToPMine.zip

buttons	call/text	apps	camera	entertainment
power button	text messages	download apps	front camera	play games
qwerty keyboard	data plan	app store	low light	mp3 player
home button	calls and texts	installed apps	camera takes	music player
touch screen	triple minutes	apps running	taking pictures	watching videos
volume buttons	talk and text	apps installed	picture quality	fm radio
home screen	make calls	google play	camera quality	play music
physical keyboard	send receive	open apps	front facing camera	watch movies
press button	send text	apps downloaded	camera good	youtube videos
back button	texting calling	running apps	camera flash	social media
fingerprint scanner	picture messages	play games	camera great	media player
customer service	form factor	screen	battery	web/email/contacts
customer service	back cover	touch screen	battery life	web browsing
tech support	gorilla glass	screen protector	removable battery	web browser
customer support	fits pocket	screen size	battery life good	surf web
service provider	protective case	big screen	hold charge	address book
verizon store	water damage	screen resolution	replace battery	alarm clock
apple store	water resistant	screen brightness	battery charge	connect wifi
send back	feels solid	large screen	battery lasts	web site
call customer service	feels cheap	screen cracked	days without charging	easy set
serial number	fits hand	home screen	good battery life	contact list
answer questions	water proof	cracked screen	battery dies	email accounts

Table 3: Top Phrases for a Subset of Learned Product Attributes.

3.3 Product Attribute Selection

The Product Attribute Selection stage of the framework is an optional step should an expert want to manually curate either the auto-generated product attributes output from the Product Phrase Extractor or specific phrases within the learned product attributes.

In this step, I chose to discard the fourteen general topics extracted in the previous step. An alternative approach would be to include all clusters of extracted review phrases but exclude the general phrase clusters when conducting the desired analysis. No further modification to the auto-generated product attributes was performed.

3.4 Product Attribute Vocabulary

The Product Attribute Vocabulary data structure serves two primary purposes. First, the vocab data structure processes the raw phrase attributes and underlying phrases learned by the Product Phrase Extractor. A few important cleaning steps include:

1. eliminating infrequent phrases: phrase frequency $<$ 50);
2. constraining overall topic size: max phrases per product attribute = 500); and
3. ensuring phrases map to only one product attribute: if multiple product attributes contain a phrase, assign the phrase to one product attribute with the highest ToPMine frequency).

Secondly, the vocab data structure provides an encoding and decoding mechanism to convert phrase strings to integer ids and vice-versa for more efficient memory usage and computation. This vocab data structure also provides constant time lookup for all phrases belonging to a given topic.

The vocabulary of product attribute phrases learned from the mobile phone reviews contains 880 distinct phrases across the fourteen product attributes.

3.5 Phrase Candidate Generator & Tagger

The Phrase Candidate Generator serves to create phrase candidates within the sentences of individual reviews to match against the Product Attribute Vocabulary. This component utilizes Python's NLTK library to process the mobile phone review text provided by the Preprocessor. Each review is split into an array of sentences. Then each sentence is cleaned (e.g., stopword removal, lowercased, number and punctuation remove) and the resulting tokens are converted into unigrams, bigrams, and trigrams: the phrase candidates. Finally, the candidates are compared to the Product Attribute Vocabulary which eliminates candidates not found in the dictionary while simultaneously encoding those candidates that do match.

The Phrase Candidate Tagger then tags each sentence of each review with the phrase(s) and topic(s) that match the Product Attribute Vocabulary. For example, consider the following sentence "The screen resolution and camera quality are amazing!" Assume phrase `screen resolution` has vocab id = 311 and topic id = 10 and `camera quality` has vocab id = 102 and topic id = 5. Therefore this sentence would be tagged with `phrases = [311,102]` and `topics = [10,5]`. This compact representation will be key in retaining granular information about phrase and topic sentiment at an individual review level.

3.6 Sentiment Scorer

The Sentiment Scorer computes a sentence-level, lexicon and rule-based sentiment score for each sentence across all reviews using the VADER algorithm [3]. The VADER algorithm is an appropriate choice given it was designed for application to social media. Social media user-generated content serves as a better match for the colloquial voice and short-form writing in online reviews. Sentiment scores in the VADER algorithm range from -1 (negative) to 0 (neutral) to +1 (positive). While negative, neutral, positive, and compound scores are generated, the Sentiment Scorer only

retains the compound score.

3.7 Product Attribute Scorer

The Product Attribute Scorer fuses (i.e. zips) the sentiment scores learned by the Sentiment Scorer to the respective tagged phrases and topics generated by the Product Candidate Generator & Tagger.

3.8 Robust Data Preparer

With all of the processing and data augmentation complete, the Robust Data Preparer organizes the resulting data into a tabular, denormalized data set. The rationale for this final stage is to demonstrate the versatility of product perception analysis that can be conducted across different dimensions (e.g., brand, product, review year, product attribute, etc.). The final Product Perception Data Set has the columns/fields outlined in Table 4 with each row representing a specific product attribute for a distinct review.

Field	Data Type
id	int
brand	string
product	string
review_year	int
product_attribute	int
product_phrase_ids	list[ints]
sentiment	float

Table 4: Schema of the Product Perception Data Set

4. ANALYSIS & DISCUSSION

In this section, I demonstrate how the Product Perception Data Set produced by the framework can uncover valuable consumer insights at different levels of granularity.

4.1 Category-Level Perceptions

Category-level perception analysis focuses on answering questions about the mobile phone product category more broadly.

4.1.1 *How has product attribute satisfaction evolved over the last decade?*

To answer this question, I calculated the average sentiment for each product attribute and compared the change between the base year (2009) and end year (2018).

Overall sentiment has declined across 15 of the 16 product attributes, with only `processing speed` remaining unchanged. Table 5 summarizes the smallest and largest sentiment changes by product attribute.

One hypothesis for the observation is that the novelty of mobile phones has worn off; as consumers we come to expect a lot more from our devices and the pace of innovation is slowing down. For example, screens are far more durable and phones are often water resistant, yet `form factor` experienced the second largest drop in sentiment.

4.1.2 *Has the relative importance of different product attributes evolved?*

Changes in the relative frequency at which product attributes appear in reviews over time could indicate changing relative importance to consumers. Such an insight could help product designers inform how they invest time and resources in future products.

Smallest Sentiment Changes			
Product Attribute	2009	2018	Δ
processor speed	0.33	0.35	+0.02
customer service	0.07	0.05	-0.02
camera	0.34	0.27	-0.07
call/text	0.14	0.08	-0.07
User friendly	0.20	0.12	-0.08
Largest Sentiment Changes			
Product Attribute	2009	2018	Δ
apps	0.29	0.09	-0.20
form factor	0.33	0.15	-0.18
battery	0.16	0.010	-0.15
buttons	0.21	0.07	-0.14
entertainment	0.32	0.18	-0.14

Table 5: Evolution of Mobile Phone Product Attribute Sentiment

For this analysis, I selected the product attribute in the base year that had approximately average mention frequency. This was the `web/email/contacts` attribute.

Table 6 summarizes the analysis and informs the following insights. First, it is clear that `web/email/contacts` have evolved from the hero features of mobile phones in 2009 to a table stakes afterthought. The `entertainment` attribute experienced a similar, though slightly smaller, shift in importance over the same period. Secondly, the `battery` attributes have significantly risen in relative importance. Consumers mentioned `battery` only about 1.2 times the frequency of `screen` in 2009. That relative frequency more than doubled to 2.5 times in 2018.

2009		2018	
Attribute	Rel. Freq.	Attribute	Rel. Freq.
user friendly	2.4	battery	15.8
camera	2.0	call quality	10.5
call quality	1.8	user friendly	10.5
battery	1.8	camera	8.4
screen	1.5	sim/sd card	7.4
web/email/cont.	1.0	screen	6.3
sim/sd card	1.0	customer service	5.6
calls/text	0.9	form factor	2.8
entertainment	0.8	calls/text	2.5
form factor	0.8	carrier	2.5
buttons	0.7	chargers	2.3
chargers	0.5	buttons	2.2
customer service	0.4	processor speed	2.1
apps	0.3	apps	1.8
carrier	0.2	entertainment	1.1
processor speed	0.1	web/email/cont.	1.0

Table 6: Evolution of Relative Frequency of Reviews Mentioning Product Attributes (base category = `web/email/contacts`)

4.2 Brand-Level Perceptions

Brand-level perception analysis tackles questions about how brands are generally perceived against their competitors based on their collective product offerings.

4.2.1 How do consumer perceptions of a brand's products compare to the category as a whole?

Brands need to understand how their products are perceived vis-a-vis the competition. One way to understand that positioning is to compare the brand to the average perception across leading brands in the category.

For this analysis, I calculated the average sentiment for each brand across each product attribute. These brand averages were then averaged again to create a category average that assigned equal weight to each brand's position. Without using a simple average of brand averages, the skew in number of reviews would have caused frequently reviewed brands (i.e. Samsung) to dominate the category average.

Figure 1 depicts the relative product attribute perceptions of Samsung versus the mobile phone category (as defined to include the brands shown in Table 1).

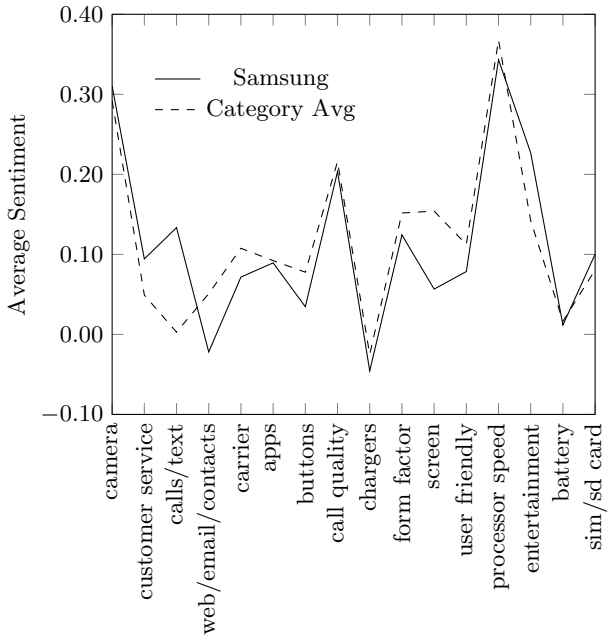

Figure 1: Product Attribute Sentiment: Samsung vs. Category Average

4.2.2 How has a brand's product perception evolved over time?

Similar to the analysis above, a brand can plot the consumer sentiment along the various product attributes at multiple points in time to understand sentiment evolution.

As shown in Figure 2, Samsung has experienced a drop in consumer sentiment along several important product attributes including `battery`, `screen` and `form factor`.

4.3 Product-Level Perceptions

Product-level perception analysis tackles questions about how individual products are perceived relative to other products or how perceptions evolve over time.

4.3.1 How do consumers perceive two different mobile phone models?

Figure 3 visualizes the differing consumer perceptions of Samsung's Galaxy S4 and Galaxy S5 models. This analysis

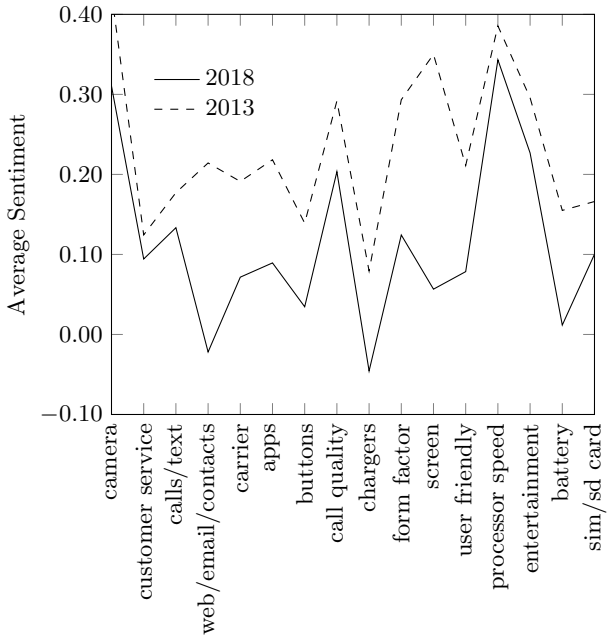

Figure 2: Product Attribute Sentiment: Samsung Comparison Between 2013 and 2018

is conducted for the same point-in-time—specifically from reviews created in 2017.

The analysis reveals product attributes where the next-gen model is outperforming and underperforming the late-model. The next-gen model seems to not meet consumers expectations in terms of `processor speed` and `camera`.

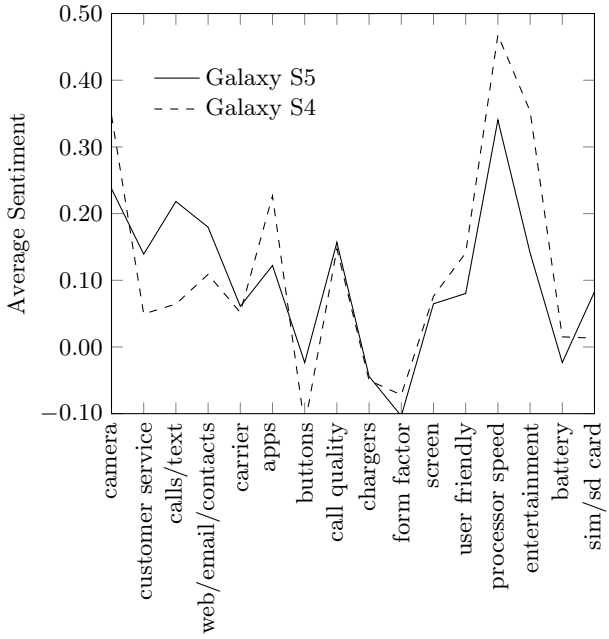

Figure 3: Product Attribute Sentiment: Contrasting Samsung S4 & S5 Devices

5. CONCLUSION & FUTURE WORK

In this paper, I described a novel framework to extract and categorize consumer perceptions from a large corpus of mobile phone reviews and related metadata. The proposed text transformation pipeline successfully discovered meaningful product attributes by applying an automated topical phrase mining approach. Topical phrase mining proved quite powerful for the task-at-hand by clustering the diverse vocabulary of extracted product-related phrases (critical to opinion mining and sentiment analysis) while simultaneously clustering those phrases into product category-specific attributes. Opinion mining the reviews surfaced product attribute perceptions which I then demonstrated could be utilized to objectively answer valuable market research inquiries at different levels of granularity—namely product category, brand, and individual product levels—that would not be possible in the absence of product attribute discovery and opinion mining.

The findings in this paper would benefit from three areas of future work. First, the presence of fake reviews in the data set inject bias into my approach of scoring customer perceptions. Industry and academia have increasingly proposed solutions to fake review identification, many of which combine elements of natural language processing and classification. I suggest inserting a Fake Review Filter component prior to the Preprocessing component of the framework to minimize positive skew in sentiment and perceptions.

Secondly, the quality of the opinion mining would benefit from exploring more advanced phrase-level versus sentence-level sentiment algorithms. Current state-of-the-art sentiment analysis incorporates context-aware embedding approaches as opposed to the more traditional lexicon-based approach applied in my framework. That said, I suspect that phrase-level sentiment algorithms would have a smaller impact on the quality of results given the overwhelming majority of sentences in the mobile phone reviews utilized in these experiments contained at most one meaningful product attribute phrase.

Finally, pairing the proposed perception discovery framework with an interactive, information retrieval application would dramatically enhance the value of this overall approach to market researchers. Traditional survey-based approaches to consumer perception discovery enable a researcher to drill into overall findings down to individual responses. A similar rich experience could be delivered by allowing a researcher to drill down into summaries of reviews along a given product dimension or even down to individual reviews containing insights to better understand the reasoning for the higher-level product attribute perceptions.

6. REFERENCES

[1] Bishop Todd. Amazon files first-ever suit over fake product reviews, alleging sites sold fraudulent praise, 4 2015. Last accessed 2 August 2020.

[2] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *Proc. VLDB Endow.*, 8(3):305–316, Nov. 2014.

[3] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[4] Kantar. Kantar's 2019 top 50 global retailers, 2019. Last accessed 2 August 2020.

[5] J. Ni, J. Li, and J. McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[6] NPR. Some amazon reviews are too good to be believed. they're paid for., 7 2018. Last accessed 2 August 2020.

[7] Picchi, Aimee. Buyer beware: Scourge of fake reviews hitting amazon, walmart and other major retailers, 2 2019. Last accessed 2 August 2020.

[8] Weise, Elizabeth. Amazon cracks down on fake reviews, 10 2015. Last accessed 2 August 2020.