# Topic Analysis on Amazon Smartphone Reviews

*Chu Pan*

*12/9/2019*

## 1 Introduction

The modern smartphone has taken a long journey to reach us in 2019, and it has changed drastically along the way. It is an evolution that has taken the market by storm. In the past decade, smartphone display resolution grew by 11 times (Bryant 2019), 3.5mm headphone jack came and went (Swearingen 2018), traditional behemoths like Nokia exit the stage while innovative brands like Apple, Samsung, Huawei established a new business ecosystem in the market. With the boom of new technology, understanding market trends has become an increasingly difficult task. In this circumstance, online product review—a free resource mirroring consumers' thoughts on products—has been brought to attention. Analysis of online reviews may help sellers in understanding consumers' interests and concerns before launching new products.

This article applies Latent Dirichlet Allocation to examine topical tendencies from a corpus of over 20,000 Amazon smartphone reviews between 2013 and 2018. The research questions we seek to address are as follows:

- What are some of the topics in Amazon smartphone reviews?
- Which topic is the most popular among smartphone buyers?
- How did the topical trends change in the past six years? What does the result suggest about consumers' interests or concerns?

## 2 Data

The empirical data of this study comprises over 20,000 Amazon smartphone reviews spanning Jan 2013 – Dec 2018 that Griko Nibras (2019) has compiled and provided on his website.[1] The dataset includes both locked and unlocked carriers, and scoped on 9 brands: Apple, Google, Huawei, Motorola, Nokia, OnePlus, Samsung, Sony, Xiaomi.

Table 1 lists the number of reviews and tokens included in our corpus. The corpus consists of 20 thousand reviews and 1.8 million tokens, punctuations and numbers removed. What is noteworthy here is that our sample size tripled between 2013—2018 due to an increase in the number of reviews during that time period. The unbalanced sample requires that dispersion and frequency statistics be interpreted carefully for comparisons over time.

Table 1: Corpus of Amazon smartphone reviews, 2013-2018

| | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | Total |
|---|---|---|---|---|---|---|---|
| Texts | 1022 | 1613 | 1789 | 3151 | 6084 | 7146 | 20805 |
| Tokens | 168473 | 186574 | 173015 | 260287 | 436859 | 585842 | 1811050 |

## 3 Methods

### 3.1 Preprocessing

*Language Detection.* The original dataset scraped from Amazon contains reviews in all languages including English, Spanish, Chinese, etc. We used `detectlanguage` function in Google Spreadsheet to detect and eliminate non-English reviews.

[1] https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews#20190928-reviews.csv

*Lemmatization* The tokens in the corpus were lemmatized and converted to lower case in order to reduce the inflectional forms from each word to a common base or root.

*Stopword Removal.* Although creating a document-feature matrix is completely automatic, we can control the output by pre-processing the corpus. It is standard practice to remove common syntactical stopwords (such as *the* and *of*) and tokens associated with our search terms (such as *mobile* and *phone*). These words occur so frequently, and with such regularity in all documents, that they overwhelm topical variability. To avoid this situation, we used `quanteda` package in R to remove stopwords and search terms.

*Non-Standard Word Removal.* It was determined that standard stopword removal is not sufficient for this corpus. Smartphone reviews have properties that differ from the scientific journals and news articles typically used in topic modeling. Most obviously different is the use of series and version labels for electronic products. For example, in "Moto Z Droid version XT1635", the phrases "Z Droid" and "XT1635" do not hold any semantic meaning. They are tantamount to proper nouns for a particular Motorola release. These words are not useful in detecting meaningful topics in our data. Therefore, we used POS tagger in `spacyr` to identify word types and eliminate all non-standard dictionary words from the corpus. That helped us reduce runtime while still having good results.

### 3.2 Modeling

The main objective of this project is to analyze topical tendencies of Amazon smartphone reviews. The simplest approach to statistical analysis of online reviews is to count tokens. But if we use word counts to draw conclusions about the topical trending across different reviewers and time frames, we risk making mistakes because words are variable and ambiguous (Jockers and Mimno 2012). Variability arises because reviewers often have a choice of several synonyms. In order to make claims about topical tendencies, we would have to summarize the results of hundreds of word associations. Even if we exhausted all possible synonyms and antonyms of each token, they may not suffice to represent a topic. For instance, *return*, *shipping*, *refund* may not be associated by semantics, but under the context of e-commerce, they are all under same topic of "seller's reliability". Such a relationship cannot be identified solely by word association. Ambiguity adds further complications: if we count the occurrence of a single word, we may unwittingly conflate multiple meanings of that word (i.e. "pixel" as a picture element and "pixel" as a Google smartphone).

Statistical topic models, such as Latent Dirichlet Allocation (LDA), use contextual clues to group related words and distinguish between uses of ambiguous words (Jockers and Mimno 2012). In this study, we implemented an LDA model with the R language and environment (R Core Team 2017) for topic mining. Our model employs the "bag of words" approach to text analysis. That is, it assumes a document is generated by picking a set of k topics and then for each topic picking a set of words. There is an important parameter that must be specified upfront: k, the number of topics that the algorithm should use to classify documents. Small k tends to result in topics of a broad and general nature; larger k is usually associated with more focused topics and slower computation. There is a tradeoff between accuracy and efficiency. After much trial and error, it was determined that k=8 yielded the most semantically meaningful results in a reasonble runtime.

Our topic model will contribute to answering our research questions in two ways. Firstly, the model reduces the dimensionality of the corpus by assigning each word to one (or more than one) of the eight clusters or topics. This level of complexity is rich enough to express much of the variability of the corpus, but small enough to be interpreted by humans (Jockers and Mimno 2012). Secondly, the model is able to identify the primary topic in each document. When linked with time series data, the model offers a way of exploring macro scale topic trend over time, which gives a direct answer to our research question #3.

## 4 Results

### 4.1 Eight Topics in Amazon Smartphone Reviews

We begin by asking what are some of the topics in Amazon smartphone reviews. Using Latent Dirichlet Allocation, we were able to identify 8 topics[2] from the corpus:

[2] A more comprehensive list of the topics and associated terms can be found in Appendix A.2.

Table 2: Topic labels and terms for the Amazon smartphone review corpus

| topic | label | terms |
|---|---|---|
| 1 | battery | battery, charge, life, day, ... |
| 2 | camera & display | camera, quality, display, light, ... |
| 3 | brand | note, love, iphone, samsung, ... |
| 4 | basic features | app, button, text, call, ... |
| 5 | seller's reliability | new, product, seller, refurbish, ... |
| 6 | carrier | unlock, card, sim, tmobile, ... |
| 7 | general concern | get, month, time, call, ... |
| 8 | operational system | update, window(s), android, support, ... |

Among the eight topics, Topic 7 describes rather general topical coherences. Topic 3 seems to focus on brand, but a deeper look at the original texts reveals that the topic refers specifically to Samsung's long-running patent war against Apple. Other topics correspond more to seller's reliability or specific features of the products. These topics reflect customer preference in a way that is less vulnerable to special events.

### 4.2 Topic Popularity

Popularity of the topics can be measured by the proportion of each topic in the entire corpus. A difference among those proportions is evidence that some topics are more likely to occur than others. Figure 1 offers a corpus wide view of the proportion of topics.

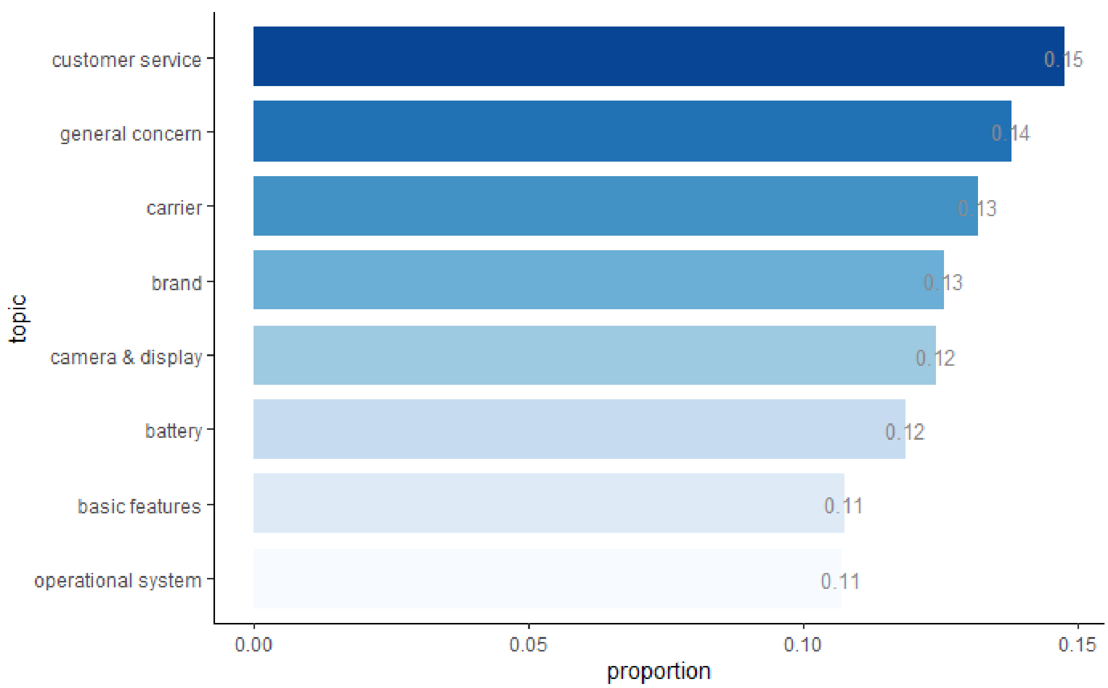


Figure 1: Topic popularity: proportion of occurrence in the entire corpus

Figure 1 shows that "seller's reliability" is the most-mentioned topic in this corpus, followed by "general concern" and "carrier". A word cloud representation of the topic's key terms in Figure 2 provides a good initial understanding of its scope. Regarding "seller's reliability", reviewers are most concerned about:

i) fast delivery: product is shipped and delivered on time.


Figure 2: Word cloud of topic labeled “sellers reliability"

ii) product condition: decent packaging; product looks brand new with no damage.
iii) accessories: the product comes with accessories (authentic charger, headphone, etc.) as promised.
iv) return & refund: free return is guaranteed; refund request is processed in a reasonable time frame.

We noticed that contents related to “seller’s reliability” are often placed at the beginning of the review paragraph. And the topic often determines the tone of the entire comment. Extract 1 shows one example of how “seller’s reliability” is mentioned in Amazon reviews. Sentences related to “seller’s reliability” are highlighted in bold. The paragraph starts with an evaluation of the product’s condition and accessories, leading the tone of this review. Despite the complaint about the product’s compatibility, this customer closed the paragraph with a positive ending. The effect of “seller’s reliability” can be seen more clearly in Extract 2. The initial version of this review was submitted within 5 days of purchase and included only the first two sentences, which focused exclusively on product condition. We can see that the upscale packaging and attractive visuals have made a favorable impression. It was until two months later that the buyer wrote an detailed update on the product’s functionality. Despite all the flaws and problems with Galaxy phones, the customer claimed that she “really do like this phone” in the end.

> Extract 1.
>
> **The phone looked almost new, barely to no scratches. Only bad part is that the phone didn’t come with authentic charger** and Verizon setting on phone is super imprinted. Had to disable a lot of Verizon apps and even defaulted messaging app,since I’m a cricket service user. Besides that great phone in excellent condition. (product rating: 4 stars)
>
> Extract 2.
>
> **Came as advertised, nicely packaged with the original seal. Practically mint condition with absolutely minimal scratching, maybe one or two tiny ones on the back if you look really hard.** Otherwise, working well, **practically new**. UPDATE: Had the phone just under a month and the rear camera started malfunctioning but would still work, then just under 2 month mark it consistently fails to start up. Unfortunately this seems to be a common problem with Galaxy phones, however, I tried practically every solution available but no luck. Was told its probably a hardware issue if a factory reset failed to fix the issue. Maybe the unit I got was a bit defective or had underlying water or fall damage? Anyway, will be replacing for another and hope I don’t have the same issue as I really do like this phone. (product rating: 4 stars)

Extract 3 gives one example of how "seller's reliability" can be used in a negative tone. Our LDA model identifies "seller's reliability" as the primary topic in this paragraph with a topic probability of 49% (Appendix A.3), which is thrice as high as the mean probability of this topic in the entire corpus.[3] The customer's attitude can be measured by the product rating. This customer rated one star for the purchase, the lowest possible rating on Amazon. For this case, we can say for sure that product quality itself was not the main factor of the poor rating. The same product sold by other sellers on Amazon received an average product rating of 4.2. This customer also admitted it was a "great phone". Seller reliability and customer service were the main issues here.

> Extract 3.
>
> Great phone, **useless and hopeless support. XXX[4] does not stand behind their product. I still could not claim the 12 month manufacturer warranty. The support team have only one remedy - contact the seller, Amazon in my case. After 2 month Amazon is not responsible for any product support, but the manufacturer should. The support team does not get this. The manufacturer should provide the option (through the support team) how to claim the warranty.** In my case the camera flash does not work and the phone is 3 month old. (product rating: 1 star)

This example also points out an interesting direction for future studies: whether lengthy description of the topic is associated with negative customer experience. Indeed, Extract 3 has an unusually high topic probability of "seller's reliability" in this corpus. The author used an entire paragraph describing his unpleasant experience with the seller. When the shopping process goes smoothly, customers tend to take it for granted. Therefore, "seller's reliability" appears frequently in this corpus, but only briefly[5] in most reviews. The only case in which customers may want to elaborate on "seller's reliability" is when something goes wrong and they have to complain. A more rigorous examination of the association between topic probability and customer satisfaction would require hypothesis testing and regression analysis. Due to time limitation, we did not implement any of those. The finding remains an inspiration for future works.

It is noteworthy that topic popularity in this study should not be taken directly as the importance of topics. One could definitely argue that "seller's reliability" outweighs the other topics for other reasons. Although there is no specific time frame a buyer has to leave a review on Amazon, most customer reviews are submitted within two weeks of purchase. Given such a limited time period, buyers can hardly explore all features of their smartphones. In contrast, a judgment on accessories and product appearance can be completed at first glimpse. The dominance of "seller's reliability" can be partly explained by the time frame of Amazon customer reviews. Also, it may be the case that commenting on camera and display requires too much engagement and professionality, while complaining about slow shipping is way more straightforward. Yet still, it is interesting to see that first impression shaped by delivery and product condition could influence customer satisfaction even before they use the product.

In short, "seller's reliability" is the most popular topic in this corpus of Amazon smartphone reviews. This topic tends to appear at the beginning of the review paragraphs, as delivery and product condition are crucial to buyers' first impression on the purchase. Though the finding may not translate directly into topical importance, we could safely make the following inference on ecommerce: customers not only care about the products they receive but also the service that comes along with the purchase. When shopping online, the first physical contact customers have with a specific seller is often unwrapping the package that has been delivered to them. This is an unique opportunity for sellers to make a favorable impression. Fast shipping, premium packaging and strong customer support are tied to customer satisfaction and should be considered as crucial parts of ecommerce.

### 4.3 Topic Trend Over Time

At this point, we have had a general understanding of topic popularity in this corpus. However, as previously mentioned, our sample size tripled between 2013 and 2018. Since we used more data from 2018 to train the

[3]The mean probability of "seller's reliability" over the entire corpus is 15%.
[4]The retailer name is hidden for data security reasons.
[5]two or three sentences as in Extract 1 and 2

LDA model, the results may be more representative for 2018 than 2013. A natural question to ask is: how does 2013 reviews differ from 2018? Is there any change in topic trend and tendencies during the past six years? To figure out those questions, we linked the LDA output with time series data and aggregated topic proportions by year. The result is displayed in Figure 3 and Figure 4 below.

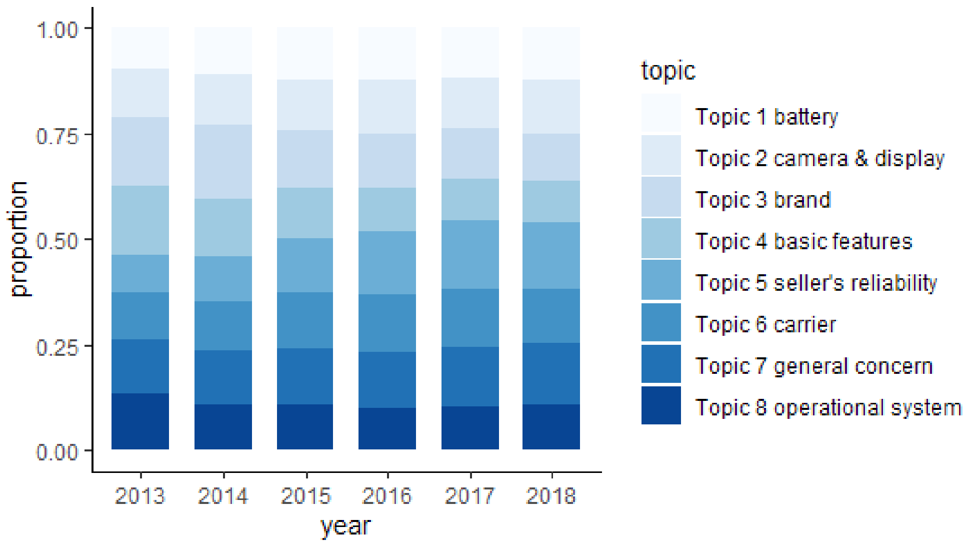


Figure 3: Topic proportions over time

Figure 3 suggests that "seller's reliability" has earned an increasing popularity among reviewers in recent years. "Seller's reliability" started off with a porportion of 0.09, ranked the lowest in 2013. Since then, the topic has gained more attention at a 14% pace of growth and finally became the most-mentioned topic in 2018. The proportion of the topic labeled "carrier" also increased slightly between 2013 and 2016. It is worth noting that the most significant change in "carrier" was not its proportion, but the most frequently used terms inside this topic.[6] 2013 was the year for contract phones, where network providers retained customers by selling smartphones at a competitive price along with their data plans. Therefore the term "contract" stood out on top of 2013 keywords. As 4G network became prevalent, unlocked phones became the mainstream in recent years. Customers prefer smartphones that are compatible to all networks, allowing them to switch between carriers as they wish. Consequently, "unlocked" emerged as a keyword in 2015 and has remained on the list since then. The transition reflects an alternation in customer interest in the past six years.

On the other hand, some topics have shown decreasing tendencies in the past six years. The proportion of the topic labeled "basic features" dropped from 0.16 to 0.09 between 2013 and 2018. With Infinity Display, Leika three-lens camera and all those fancy features striking our nerves every single day, basic functionality of cell phone—such as sending text messages and checking calendar—has gradually faded out of our sight. The topic labeled "brand" has also diminished in the past six years. As mentioned in section 4.1, this topic refers more to the battle between Apple and Samsung than general discussion on brands. As more cell phone manufacturers entered the market, it makes sense that people have been distracted from the topic.

The proportions of other topics were quite stable over time. Battery, camera and display gained noticeable attention in all six years. "Operational system", a relatively unpopular topic in this corpus, maintained a 10% share from 2014–2018. The only exception arised in 2013, when the topic had a proportion as high as 13%. Examining the raw data, we found that Nokia and Samsung were still making Windows phones in 2013. Buyers either loved or hated the system so the debate was reflected in Amazon reviews. After the cutoff, Nokia exit the stage and Samsung turned completely to Android. Since then, IOS and Android have dominated the world. There was not much room left for discussion.

[6] Keyness analysis done in previous coffee break experiment. A word cloud representation can be found in Appendix A.5.

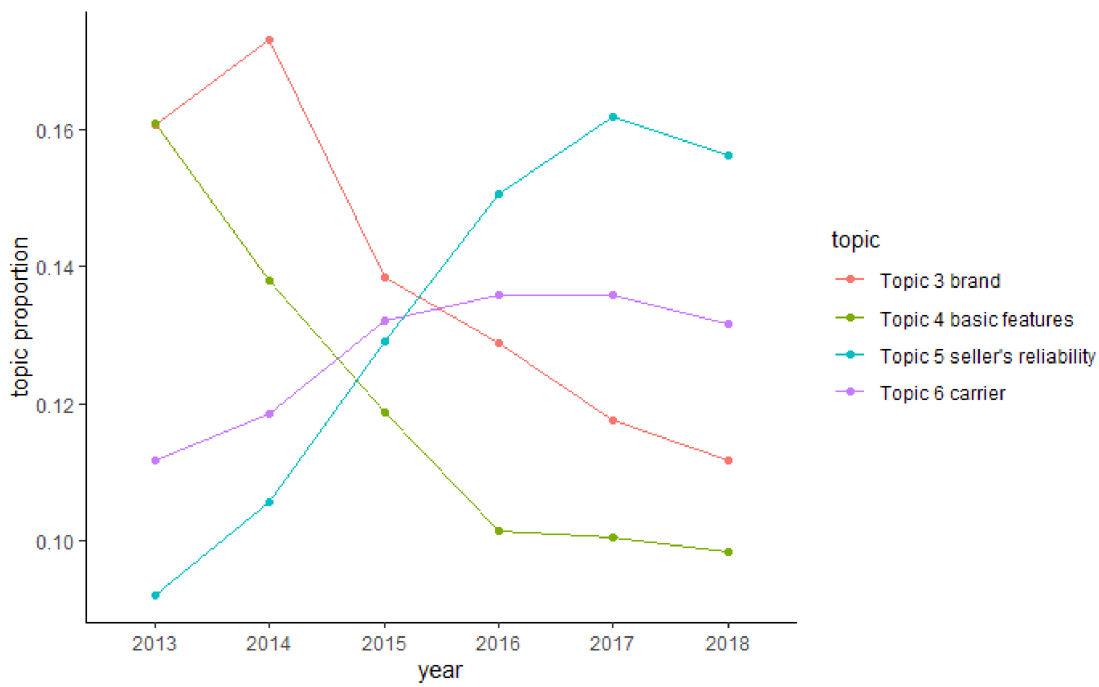


Figure 4: Trendlines for several topics

## 5 Discussion

In this article, our aim has been to understand topical tendencies in Amazon smartphone reviews. We have established a topic model with Latent Dirichlet Allocation and incorporated time series data to examine topic proportions over time. We have analyzed topical effects by empirically examining the commentary on Amazon, and linked the results to customer insights and market trends.

We have argued that the Amazon smartphone reviews can be grouped into eight topics: battery, camera & display, operational system, basic features, carrier, brand, seller's reliability, and general concern. Among them, seller's reliability was the most-mentioned topic with an increasing trend in recent years. We have shown that delivery, packaging, product appearance, and customer support are crucial factors of buyers' satisfaction. Based on the findings, we further hypothesize that these factors have a priming effect on product ratings and should be taken seriously by sellers.

Furthermore, combining time series with keyness analysis, we have noticed a transition in customer preference on contract versus unlocked phones in the past six years. Back in 2013, contract phone was the mainstream. Whereas in 2018, unlocked smartphones have become more favorable among buyers. Battery, camera and display maintained noticable attention from customers between 2013 and 2018, while basic functionality gradually faded out of sight.

The findings of this study have to be seen in light of some limitations. First, the measurement of topic proportion may not fully represent topical importance. In this corpus, the proportion of topic labeled "camera & display" did not stand out in any of the six years, which is quite counter-intuitive. Back in 2017-2018, "camera" and "display" were the most highlighted features in many smartphone advertisements. Huawei P10 was known for its high-resolution Leica camera; Samsung first introduced its Infinity Display—a virtually bezel-less screen—on the Galaxy S8. Both were advertised as the brand's killer features in the market. However, those features did not seem to provoke a huge storm in Amazon reviews. At this point, we cannot draw any conclusion on ineffective marketing yet. It may be the case that commenting on camera and display required too much engagement and professionality, while complaining about slow shipping was way more straightforward. It is also possible that the appeal of camera and display matters in the sense that it determines how much someone spends and whether or not they make the purchase, but not after-sale experience, which is the main focus of Amazon review. After all, there is no single best way to market research. Future works should combine topic mining with results from market trend analyses and consumer interviews to gain a more comprehensive understanding of customer preference.

Moreover, this study is limited by the reliability of Amazon reviews. Amazon does not require users to actually buy the product before they submit a review. In this dataset, we noticed that some reviews looked exactly the same: they were posted by the same user on the same day, but written for different products. During data cleaning, we identified those entries as fake reviews and removed all the suspecious duplicates. But still, there is no guarantee that the dataset provides an unbiased representation of customer experience.

## References


Bryant, Laurence. 2019. "Ten Years of Mobile Evolution: a look at the Samsung Galaxy S series." https://community.arm.com/developer/ip-products/processors/b/processors-ip-blog/posts/samsung-galaxy-s-ten-years-of-mobile-evolution.

Jockers, Matthew L., and David Mimno. 2012. "Significant Themes in 19th-Century Literature." *Faculty Publications – Department of English* 105. http://digitalcommons.unl.edu/englishfacpubs/105.

Nibras, Griko. 2019. "Amazon Cell Phones Reviews." https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews/version/1{\#}20190928-reviews.csv.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Swearingen, Jake. 2018. "We're No Longer in Smartphone Plateau. We're in the Smartphone Decline." http://nymag.com/intelligencer/2018/12/global-u-s-growth-in-smartphone-growth-starts-to-decline.html.

## Appendix

### A.1 Code for topic modeling

Step 1 - Preprocessing

```
library(dplyr)
library(tm)
library(lda)
library(topicmodels)
library(textstem)

# read data from local csv file
review_df <- read.csv("./mydata/amazon-reviews.csv")
# format variable "year" for time series analysis
review_df <- review_df %>%
  mutate(year = format(as.Date(review_df$date, "%m/%d/%Y"),"%Y"))

# create corpus
review_corpus <- Corpus(VectorSource(review_df$body))

# attach metadata
meta(review_corpus, tag="year") <- review_df$year
meta(review_corpus, tag="title") <- review_df$title
meta(review_corpus, tag="rating") <- review_df$rating
meta(review_corpus, tag="brand") <- review_df$brand
meta(review_corpus, tag="product") <- review_df$product

# pre-processing
review_corpus <- tm_map(review_corpus, content_transformer(tolower))  # lower case
review_corpus <- tm_map(review_corpus, removeNumbers)  # remove numbers
review_corpus <- tm_map(review_corpus, removePunctuation)  # remove punctuation
review_corpus <- tm_map(review_corpus, lemmatize_strings) # lemmatization
review_corpus <- tm_map(review_corpus, removeWords,
                        c("the", "and", stopwords("english"))) # remove stopwords

# remove words associated with our search terms
review_corpus <- tm_map(review_corpus, removeWords,
                        c("phone", "cell", "mobile","smart"))

# load pre-compiled wordlist
load("./mydata/mywordlist.RData")
# remove redundant words (described in method section)
review_corpus <- tm_map(review_corpus, removeWords, mywordlist)
# remove extra whitespace
review_corpus <- tm_map(review_corpus, stripWhitespace)
```

Step 2 - Create document term matrix

```
# create dtm
doc.lengths <- rowSums(as.matrix(DocumentTermMatrix(review_corpus)))
review_dtm <- DocumentTermMatrix(review_corpus[doc.lengths > 0])
```

Step 3 - Implement Latent Dirichlet Allocation

```
# topic modeling
ldaOut <- LDA(review_dtm, 8)
```

Step 4 - Visualize results (code only)

```
# top 10 terms in each topic (re-ordered by score)
# (the scoring favors less general, more specific terms to describe a topic)
# (output in Appendix A.2)
ldaOut.terms <- lda::top.topic.words(posterior(ldaOut)$terms, 10, by.score = T)
# top 3 topics assigned to each document
ldaOut.topics <- as.matrix(topics(ldaOut,3))
# topic probabilities by document
ldaOut.prob <- as.data.frame(ldaOut@gamma),10)
# topic rank
ldaOut.rank <- colSums(posterior(ldaOut)$topics) / nrow(review_dtm)
```

### A.2 Top 10 terms in each topic

Table 3: Top 10 terms in each topic

| Topic 1 battery | Topic 2 camera & display | Topic 3 brand | Topic 4 basic features |
|---|---|---|---|
| battery | camera | note | app |
| charge | quality | love | button |
| life | sony | iphone | text |
| price | display | samsung | motorola |
| case | light | galaxy | call |
| screen | feel | apple | moto |
| day | video | big | message |
| hour | screen | upgrade | volume |
| break | fingerprint | feature | voice |
| drop | android | size | hear |

| Topic 5 seller's reliability | Topic 6 carrier | Topic 7 general concern | Topic 8 operational system |
|---|---|---|---|
| new | unlock | get | update |
| product | card | month | issue |
| seller | sim | time | window |
| refurbish | work | call | nokia |
| receive | buy | take | support |
| brand | att | turn | app |
| purchase | tmobile | start | android |
| condition | network | bad | problem |
| box | carrier | try | connect |
| order | money | screen | software |

### A.3 Topic probability for Extract 3

Here is the R code and outout for the topic probability for Extract 3. `V5` ("seller's reliability") received the highest probability and was thus identified as the defining topic in the paragraph. The number highlighted in red says that Extract 3 have 49% probability of Topic 5 - seller's reliability.

```
> as.data.frame(ldaOut@gamma)[18125,]
              V1          V2         V3       V4        V5        V6        V7        V8
18125 0.07356889 0.007849753 0.04063648 0.042648 0.4917136 0.0144469 0.2286908 0.1004455
```

### A.4 Code for Topic Proportions Over Time

```r
# aggregate mean topic proportions by year
topic.trend <- aggregate(posterior(ldaOut)$topics, by = list(year=review_df$year), mean)
# assign topic names
colnames(topic.trend)[2:9] <- c("Topic 1 battery", "Topic 2 camera & display",
                               "Topic 3 brand", "Topic 4 basic features",
                            "Topic 5 seller's reliability", "Topic 6 carrier",
                            "Topic 7 general concern", "Topic 8 operational system")
# reshape data frame
library(reshape2)
topic.trend <- melt(topic.trend, id.vars = "year")
colnames(topic.trend) <- c("year", "topic", "proportion")
```

### A.5 Word clouds for 2013 and 2018 keywords

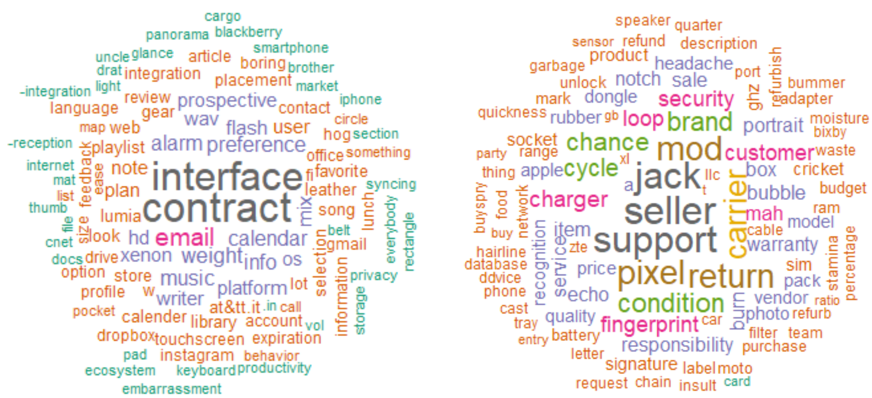


Figure 5: Word cloud for 2013 keywords (left) and 2018 keywords (right)