Predicting the Launch Success of New Products Using Amazon Reviews: A Method Combing Temporal Topic Models and Bass Model

By

[redacted]

A dissertation submitted in partial fulfilment for the degree of MSc in Business Analytics at Warwick Business School – University of Warwick.

Executive Summary

As a modern digital word of mouth, the rapid rise of UGC has a strong influence on consumer buying behaviour. Compared with the social media marketing of businesses and manufacturers, UGC, such as online reviews, is more credible since it can better reflect purchased consumers' true perception of the quality dimensions of the product. In this dissertation, we apply the extension of the probabilistic topic models named Structural Topic Models to extract dimensions of customer satisfaction with the phone product (consumer durables) from online reviews of Amazon UK, so as to obtain valuable information about product quality improvement. At the same time, we utilize the Bass model to calculate the innovation coefficient (coefficient of external influence) and the imitation coefficient (coefficient of internal influence) of each review. In order to analyze the dimensions of product quality changes over time, we take the coefficient difference between the innovator and the imitator of each review, rating score, review time, and brand difference as the covariates of the temporal topic model. After that, we use regression models to explore the importance of the proportion of sentiment related topics in predicting the success of the new product release. Overall, the results of the dissertation reveal consumers' mobile phone preferences and perception of product quality dimensions, and also reflect the market competitiveness of the three mainstream mobile phone brands.

Table of Contents

Index of Tables

Index of Figures

1. Introduction & Background

With advances in technology and online media, users increasingly post their opinions on products on social platforms in various forms of content, such as images, text, and videos. These contents can be referred to as user-generated content (UGC), providing diverse sources of information to extract product quality dimensions based on consumers' experience (Tirunillai and Tellis, 2014). The product quality, reflected in different dimensions of the product, is important in driving a product's success in the marketplace such as customer satisfaction and market share (Tellis et al., 2009). Numerous researches have shown that UGC has a great impact on the success of the product marketplace, including sales and demand (Onishi and Manchanda 2012).

As an important component of UGC, online reviews written by previous consumers provide a vital source of information that affects many potential consumers' decisions on purchasing new products. As indicated by Li et al. (2011), in an online shopping environment consumers' product reviews will be considered more credible by consumers than expert's opinions. This shows that UGC, especially online product reviews, has become one of the main driving forces for initial product sales. For example, studies have shown that product sales and consumer behavior would be impacted by online Word of Mouth (WOM) (Liu, 2006; Godes and Mayzlin, 2004; Hu et al., 2008), and the attributes of online reviews have an influence on product sales.

The hidden semantic structure in UGC text bodies, especially online product reviews, can be captured and aggregated applying the topic models. These topic models, such as Latent Dirichlet Allocation (LDA) and Structural Topic Model (STM), help to detect words and phrase patterns within these give product reviews and then cluster the similar expressions that can best reflect the characteristics within these unstructured textual reviews. These clustered groups of words with different textual features called topics can also reflect the quality of the product. Over time, in most cases, the number of product reviews increases with the accumulation of sales, and the clustered topics of reviews changes with enrichment of review content.

There are many forecasting models for product sales growth, such as trial-repeat models that divide the total sales into the trial purchase and repeated purchase. One of the well-known models that predicts the growth of initial purchase of new products especially durables the Bass diffusion model (Bass, 1969). When it comes to durable goods, it usually refers to a type of product that yields utility over time so that the interval between successive purchases of

such goods is relatively longer. Examples of consumer durables that would be purchased infrequently include electronics, automobiles, and jewellery. According to the Bass model, a model presenting a rationale of the new product interaction of current and potential adopters, overall adopters can be classified as innovators and imitators based on the timing of their adoption. This model also proposed that the products with imitation coefficients (q) much larger than the innovation coefficient (p) are considered to be successfully released. Therefore, by comparing the above coefficients, the product could be divided into success and failure.

1.1 Intended Aim

The intended aim of this dissertation is to explore topic differences of the online reviews generated by innovator and imitator in the success (or failure) product, so as to predict whether the new products can be successfully launched. Specifically, this dissertation will conduct Structural Topic Model (STM), an extension of LDA, to explore topics on a sample of UGC consists online reviews of Amazon UK across 4 firms in new product segmentation of the cell phone market. In order to achieve this aim, we will first compare the latent topics difference obtained by applying Topic models on corresponding Amazon UK product reviews between innovators and imitators, for both success and failure of the new launched products particularly in cell phone segmentation. Then we will explore that under the control of sentiment, whether applying the combination of Topic models to extract the proportion of different emotional topics and the parameters obtained from Bass model on online reviews (Amazon UK reviews) can predict the success or failure of new product launches to a certain extent. The results can help companies mainly positioned in the mobile phone market to effectively capture the target consumers' dynamic perception of product quality, thereby further improving the product design and sales services.

1.2 Research Questions

This dissertation aims to answer the following two questions:

- Q1: Whether the topics obtained from Topic models on online review Amazon, particularly in cell phone segmentation (consumer durables), are different between innovators and imitators of related initial launched products?
- Q2: Under the control of sentiment, whether combining the Topic Models to extract the proportion of different emotional topics and the Bass model on the online review can predict the success (or failure) of new product launches?

Since the review data extracted from the Amazon does not contain features related to real sales, we made the following four assumptions in the dissertation exploration:

- **Assumption 1:** Every x consumers purchase the product, then an equidistant sample of them posts an online review.
- **Assumption 2**: Each review is regarded as a real sales transaction.
- **Assumption 3:** The date of the review shown on the website is the same as the time of purchase.
- **Assumption 4:** The date of the first review of each mobile phone product is the product release date.

1.3 Dissertation Outline

The rest of the dissertation is structured as follows. Section 1 briefly introduces the background knowledge and the aim of the dissertation. Section 2 provides a relatively comprehensive literature review, while Section 3 introduces the research methods including data collection, pre-processing, the Bass model, topic model application, and regression. Section 4 then describes the topic solutions and regression analysis. Section 5 summarizes the findings and limitations of our dissertation and then proposes the possible future research directions.

2. Literature Review

2.1 Online Reviews and Consumer Behaviour

Referred as electronic word of mouth (eWOM), consumer product reviews are considered as one of the unique features of online shopping that people rely on most when making purchase decisions and have attracted widespread attention from public and academia (Huang et al., 2009). Previous researches have shown that online review is of vital implications to product's success since it influences consumer buying behaviour and product sales (Chevalier and Mayzlin, 2006; Hu et al., 2008; Lee and Shin, 2014). They find that different components of online reviews have an effect on consumer satisfaction to the target product, such as the volume of reviews (Cui et al., 2012), the sentiment expressed in context (Liu et al., 2017; Fan et al., 2017), rating of reviews (Chevalier and Mayzlin, 2006), topic models extracted from reviews (Korfiatis et al., 2019).

2.2 Topic Models and Prediction on Online Reviews

Topic models use an unsupervised approach to explore latent semantic structures of textual data, providing insights in understanding a large number of unstructured text data. In recent research, utilizing user-generated content such as online review to obtain critical potential dimensions of consumer satisfaction with quality by applying topic models has become increasingly popular (Tirunillai and Tellis, 2014; Guo et al., 2017; Korfiatis et al., 2019). Among the topic models mentioned above, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is one of the most frequently used models to extract the proportion of multiple topics for specific aggregation of online review. As an extension of tradition probabilistic topic models, such as Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM) (Blei and Lafferty, 2007), Structural Topic Models (STM) tries to improve the estimation the topics by including additional covariates, and the observed covariates affect two components of the model (Roberts et al., 2016), namely the proportion of documents associated with a topic (topic prevalence) and the word rate associated with topics (topic content).

In recent research, text features extracted from online reviews using temporal topic models are gradually used for prediction, such as Financial Latent Dirichlet Allocation (FinLDA) for financial time series prediction (Kanungsukkasem and Leelanupab, 2019) and LDA for clinical trial terminations prediction (Geletta et al., 2019). Applications in marketing research have gradually included forecasting sales. The existing researches are usually based on the

number of reviews and numerical variables representing the valence to incorporate the impact of product reviews on sales. For example, Archak et al., (2011) applied text mining to decompose the review text into segments describing different product features, thereby incorporating the review text into the consumer choice model. They have evaluated the model based on a unique data set from Amazon in the field of digital cameras and camcorders, containing sales data and consumer review data within 15 months. Their results demonstrated the textual information in online reviews can be used to analyse the consumers' relative preferences for different product functions, thereby predicting the future sales changes of the product.

Similar to other topic models, STM defines a process of data generation for each document and find the most possible values of parameters in the model based on these data. According to Roberts et al., (2019), the key innovation of this model is that it allows users to combine arbitrary metadata into the topic model. The technical details derivation of model formulas can be found in Roberts et al., (2016) and the graphical representation of this model can be seen in **Figure 1**.

Within STM, a topic is considered as a multiple of words, where each word has a probability of belonging to a certain topic. Also, a review document is considered as a mixture of topics. Researchers constrain the output to a low number of the highest probable words and topics that are usually less than 20 and assign a topic label. If we consider in this way, adding all the topic proportions within the given review document will get 1, and the total word probability of a given topic is 1. The clustered topic is a combination of dominant keywords that are typical representatives in the written document, which is here online reviews. According to Korfiatis et al., (2019), we assume that a corpus A assembled from R documents $[r_i \in \{r_1, r_2 \dots r_R\}]$, where each review document r_i contains w words. Each word from a certain review is indexed as m, where $m_i \in \{m_1, m_2 \dots m_r\}]$. In the review corpus $c \in \{1,2 \dots C\}$, each word partially belongs to the general vocabulary, while $w_{r,m}$ represents a word in the vocabulary. At the beginning of the estimation process, the number of the estimated topics $k \in \{1,2 \dots K\}$ is considered to be the main input variable in topic models. The topic distribution is influenced by the topic prevalence covariates and topic content. However, in the case of the absence of covariates, the model is simplified to a (fast) implementation of the related topic model (Blei and Lafferty 2007). In the absence of β covariates and point estimates, the model will be simplified to an implementation of the Correlated Topic Model (Blei and Lafferty 2007). A $\boldsymbol{p} * \mathbf{1}$ vector X_r can be used to describe the topic prevalence covariate of STM. Within the p dimensions, the document-level covariate composed of different dimensions affects the importance of the topic k_i in each online review r_i.

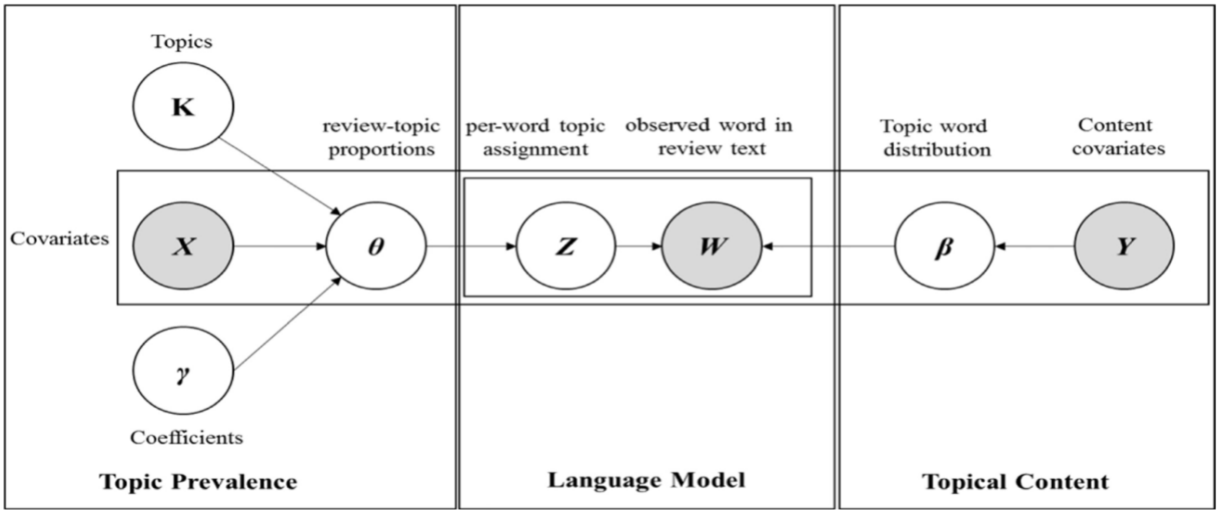


Figure 1: The Process of the Structural Topic Model Denoted by Plate Notation (Roberts, et al., 2016)

2.3 Forecast Product Sales Though the Bass Model

The Bass diffusion model (Bass, 1969) provides one of the competing paradigms in predicting, especially the innovative product diffusion and technology forecasting. It assumes that two mechanisms model diffusion patterns (Massiani and Gohs, 2015): venturesome and daring innovators adopt new products independently, while imitators purchase new products influenced by the decision of existing adopters. Innovation coefficient (p) and imitation coefficient (q) are proposed in the Bass model (Bass, 1969) indicating there are two different categories diffusion: for the successful new product whose imitation coefficient much larger than the innovation coefficient ($q > p$), the periodic sales will reach a maximum value and gradually approach to zero; while others whose imitation coefficient less than innovation coefficient ($q < p$), periodic sales volume will gradually decrease to zero starting from release time. This model is later extended, such as the model concerning successive generations (Norton and Bass, 1987) in technology products and the Generalised Bass model with pricing (Bass et al., 1994).

Specifically, the Bass diffusion model consists of a simple differential equation, that attempts to predict how many consumers will eventually adopt the initial launched product and the approximate time they will adopt. It is assumed that two types of influences help the spread of technological innovation, including the inter-personal method through Word of Mouth and mess media communication (Fitzsimmons et al., 2007). The Bass model formulation is as follows:

$$\frac{f(t)}{1-F(t)} = p + \frac{q}{m}Y(t) = p + qF(t) \qquad (1)$$

Where:

- $F(t)$ represents the fraction of the cumulative number of people who have adopted action at time t. In this case, it refers to the purchase of a new product. In other words that is the cumulative probability that the innovation is adopted by time t by the consumer in the targeted segment. As time changes, t gets larger, $F(t)$ will gradually approach 1 (**Figure 2**). Besides, $F(0) = 0$.

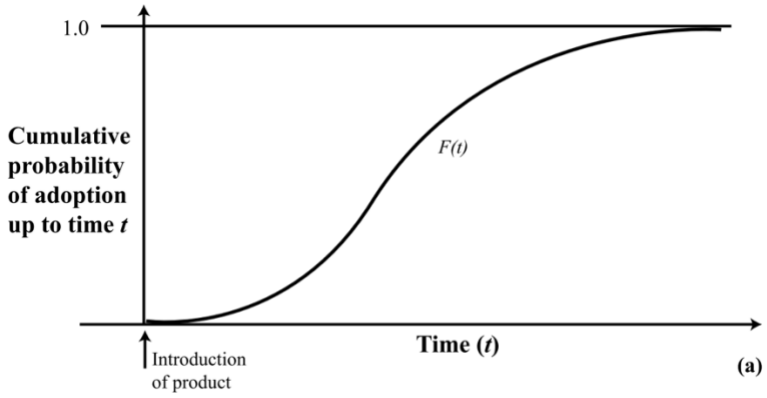


Figure 2: Cumulative Probability of Buying a New Product (Lilien et al., 2007)

- $f(t)$ represents the change of the $F(t)$, which is the rate of the adoption likelihood changing at time t. This is linked with the probability density function (**Figure 3**).

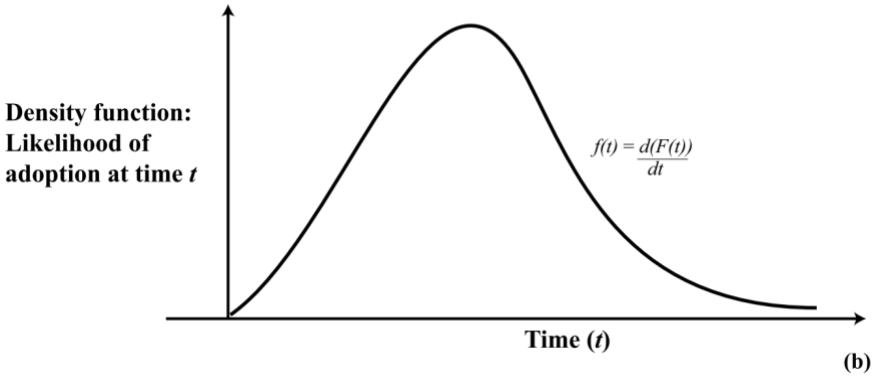


Figure 3: Likelihood of Adopting a New Product (Lilien et al., 2007)

- The left side of the Equation (1) is the conditional likelihood that the innovation is adopted by a consumer at time t, assuming that the initial product release time is 0.
- p is the coefficients of external influence (coefficient of innovation).
- q is the coefficients of internal influence (coefficient of imitation).
- $Y(t)$ is the cumulative number of buyers in the (0, t) time interval.

- m is the ultimate potential market size.

According to Bass (1967), the initial purchase of the product is determined by both innovators and imitators. The important difference between these two types of participants is the influence of purchase: the innovators' first purchased time is not affected by the number of people who have previously purchased the specific product, whereas the imitators are with a sense of learning from the consumers who have already purchased. In terms of product success, innovators will have greater influence at the beginning, however, it will decrease as time goes by. Based on the Bass model theory, where t' is the certain moment between 0 and t.

$$Y(t) = \int_0^t s(t')d(t') = m\int_0^t f(t')dt' = mF(t) \quad (2)$$

Also, sales at any period t can be indicated as $s(t) = mf(t)$, while as previously defined the cumulative sales up to time t is $Y(t) = mF(t)$. Substituting these sales for $f(t)$ and $F(T)$ in the Equation (1), we can get the following equations:

$$\frac{\frac{s(t)}{m}}{1 - \frac{Y(t)}{m}} = p + q\frac{Y(t)}{m} \quad (3)$$

$$s(t) = \left(p + q\frac{Y(t)}{m}\right)\left(m - Y(t)\right) \quad (4)$$

Combined with Equation (2):

$$s(t) = mf(t) = pm + (q - p)Y(t) - \frac{q}{m}(Y(t)^2) \quad (5)$$

Then in order to estimate the p, q, m coefficients, the following analogues are used as (Das, 2017):

$$s(t) = a + bY_{t-1} + c{Y_{t-1}}^2 \quad (6)$$

Where

Y_{t-1} is the cumulative sales from time 0 to time $t - 1$.

Then this means,

$$a = pm \quad (7)$$

$$b = q - p \quad (8)$$

$$c = -\frac{q}{m} \quad (9)$$

Combine the above three formulas, we can get

$$b = q - p = -cm - \frac{a}{m} \quad (10)$$

$$cm^2 + bm + a = 0 \quad (11)$$

By solving this quadratic equation, we can get

$$m1 = \frac{-b + \sqrt{b^2 - 4ac}}{2c} \quad (12)$$

$$m2 = \frac{-b - \sqrt{b^2 - 4ac}}{2c} \quad (13)$$

Then we take the largest and positive value among the two values of m:

$$m = \max(m1, m2)$$

Then according to Equation (7) and (8), the coefficient for p and q can be calculated as:

$$p = \frac{a}{m} \quad (14)$$

$$q = b + \frac{a}{m} \quad (15)$$

The rapid growth in online product review forums results in increased use of the Bass diffusion model for initial product launch prediction. Several researchers have explored the product sales prediction method combining the Bass model with different aspects of online reviews, such as rating from reviews of Yahoo! users (Dellarocas et al., 2007), sentiment from online review (Fan et al., 2017; Lee et al., 2017).

However, few studies have investigated how the distribution of topics obtained from review text influence the product launch success combining the application of the Bass model. Therefore, this dissertation would like to explore the relationship between the distribution of topics obtained with different sentiments using the topic model STM and the difference between innovation and imitation coefficients (p, q) using the Bass model, so as to predict the success of the initial product release.

3. Methodology

3.1 Data Collection

We obtained the data from ***Amazon Multi Language Reviews Scraper***, which is available from Github (Rémy, 2018). The dataset extracts product review related information from Amazon, one of the largest review aggregators. We mainly downloaded the reviews of 15 brands from three mobile phone companies from the official flagship store of Amazon UK corresponding to each brand as samples, including Samsung, Huawei, and Apple. The targeted product information was fetched via Python on key word search method and was stored in the format of JSON, covering the first review after the product is initially launched on Amazon to the most recent reviews on the date of download (17th June 2020). We initially collected 5,885 reviews from 15 brands, with an average length of approximately 36 words. Among these sample reviews, approximately 75.51% of them are written in English (4,444 reviews). The average rating of all the sample review is 4.51, calculated on a full score of 5.

Table 1 shows the overall summary of the sampled data set, including the original number of reviews downloaded from Amazon UK, average comment length (in words), average rating score (over 5), number of English reviews, and the percentage oof English review.

Table 1: Overall Information of Sampled Reviews (Classified by Product and Brand)

Brand	Original number of reviews	Average number of words	Average rating of reviews	Number of reviews in English	Percentage of reviews in English (%)
Samsung S8	1,207	42.07	4.50	947	78.46%
Samsung S9	252	38.04	4.48	194	76.98%
Samsung A10	274	26.05	4.31	200	72.99%
Samsung A40	490	37.69	4.50	394	80.41%
Samsung A70	222	50.28	4.28	165	74.32%
Average Samsung	489	38.83	4.41	380	76.63%
Huawei P Smart Pro	623	31.87	4.58	473	75.92%
Huawei Y5P	524	24.43	4.48	359	68.51%
Huawei P40	431	26.40	4.68	315	73.09%
Huawei P30 pro	199	72.51	4.33	179	89.95%
Huawei P30 lite	432	26.38	4.68	316	73.15%
Average Huawei	441.8	36.32	4.55	328.4	76.12%
iphone 7	405	38.40	4.28	295	72.84%
iphone 8	191	38.38	4.41	141	73.82%
iphone 11	224	32.43	4.83	170	75.89%
iphone X	186	49.08	4.49	140	75.27%
iphone XR	225	31.13	4.61	156	69.33%
Average iphone	246.2	37.88	4.52	180.4	73.43%

3.2 Data Pre-processing

In the pre-processing, textual data was cleaned and standardized for further analysis. First, we changed all the JSON formatted information into the data frame. We were interested in English textual content, and thus filtered reviews written in English. In addition, we eliminated the non-English character words that are not related to product information, including punctuation marks, numbers, pounds signs, URLs, HTML tags, empty space, etc. Besides, the standardised dates were extracted via regular expression on their review date, and all the reviewer ID were extracted from the author URL.

We removed typical stop words (e.g. “then”, “and”, “is”) in tokenization. Then, according to the histogram distribution, those words with character length less than 4 or longer than 15 were removed, accounting for approximately 9.5% of the total number of words. Part-of-speech was applied to tag words, and only nouns, adverbs, and adjectives were reserved for further topic model analysis since they are associated with consumer satisfaction and therefore influence their purchase behaviour. Words from the document term matrix that not meet the TF-IDF threshold, appearance rate of at least 2% of the product reviews (Tirunillai and Tellis, 2014), were removed to prevent outliers. After pre-processing, each individual review was treated as separate files, which contains meaningful and unordered words. The TF-IDF and top 10 most important words appearing in each brand of mobile phones were also proposed.

3.3 Application of the Bass Model

The Bass model (Bass, 1969) was used to estimate the diffusion of the initial release of the cell phones since they are considered as consumer durables that are infrequently purchased. We selectively applied the Bass model to study how adopters interact with potential adopters in the process of adopting new products, and estimated the corresponding innovation coefficient (p) and imitation coefficient (q) on Amazon reviews of mobile phones. At the same time, we made the following assumptions:

Assumption 1 X consumers have purchased the product and an equidistant sample of them posts a review. This is because only a small percentage of buyers are willing to write product reviews or offer their feedback. According to the statistics shown on the Amazon website, only 5%-10% or even fewer buyers are willing to leave feedback. Take Samsung S8 as an example, if we only count sales and reviews in the official flagship store, S8 was first available at Amazon UK on 31 March 2017, and received its first reviews 24 days after its listing. Even until the day when the data was downloaded (14 June 2020), the total number of original

reviews is 1207. This phenomenon requires us to assume that every x person who purchases the initial phone product, one person will write a review.

Assumption 2 Each review is treated as a sale transaction. To estimate the coefficient in a more convenient way, we assume each review as a single transaction under the premise of ***Assumption 1***.

Assumption 3 Review written time is the same as purchase time. According to Amazon's third-party seller feedback policy, consumers can leave their comments and rating for the products within 90 days after the purchase date. To facilitate modelling and analysis, we assume that the review date and the purchase date of the phone product are the same.

Assumption 4 We assume the date of the first review of each type of phone is the product release date. There is a specific time to market for the mobile phone products of Huawei and Samsung on Amazon, but there is no specific available date displayed for Apple. We arranged the reviews in chronological order for each newly released cell phone and estimated the Bass model based on the arrival of review, which is the arrival of each transaction. Specifically, we ignored the first three reviews of each product and estimated the coefficient of innovation(p) and imitation (q) when each review (each transaction) appears. These coefficients of p and q allow us to distinguish adopters between innovators and imitators. In the case that some reviews are with the same date, we assume they are of the same p and q coefficients.

Based on the above assumptions, we treated every review as a transaction and calculated the daily transaction volume. According to ***Assumption 1***, we mark the number of purchases as 0 on the days when the number of reviews is not available. We have also summarized the monthly sales and overall cumulative sales corresponding to each model.

According to the Bass model, a product whose imitation coefficient (q) is much larger than the innovation coefficient (p) is considered to have been successfully released. We calculated the values of innovation(p) and imitation(q) corresponding to each review. Then, we compared the value of p and q and divided the reviews into the reviews in the success phase and the reviews in the failure phase. After this, we combined these values with the topic obtained from the STM method.

3.4 Application of Topic Models (STM)

In our Structural Topic Model (STM), the specific time of review was considered as one of the main variables that affected the topic prevalence to obtain temporal topics. In order to better perform the analysis of the topic model, we divided the process into three steps. In the first step, we used STM extract the corpus for further analysis by applying the established pre-processing packages and techniques. The second step was about to determine the number of topics that best described the corpus based on the importance and representativeness of the corpus. In the third step, we combined the topics and the coefficients of innovators (p) and imitators (q) of different reviews and analyzed their connections with product launch success.

3.4.1 Text Preparation and Corpus Extraction

Based on the results of data pre-processing in ***Section* 3.2**, words that are with the same root form were derived into different groups in the process of stem and lemmatization.

3.4.2 STM for Topic Selection

We used the ***stm*** package in R for text analysis developed by Roberts et al., (2014) and Roberts et al., (2016), which allows us to estimate topic models with document-level covariates. We used the function ***textProcessor*** to convert and process the data for analysis, including removing the custom stop words (e.g. add ‘Samsung’ into stop words for Samsung related reviews). Then we utilitzed the function ***prepDocuments*** to process the loaded data into the right format. This includs setting the lower threshold to remove infrequent terms that appear less than 1%. As mentioned in the previous part, the innovation of STM lies in incorporating metadata via topical prevalence and topical content into the topic modelling framework. We chose review rating and the difference between innovator coefficient (p) and imitator coefficient (q) as one of the factors that affect covariates in the topic prevalence.

We decided ***K*** topics for each review ***r*** with Vocabulary volume of ***V***. The diagnostic properties are considered as critical criteria in deciding different user-specified topic numbers, including held-out likelihood, residual dispersion, exclusivity, semantic coherence and bound. Among these criteria, introduced by Mimno et al., (2011), semantic coherence refers to the measurement related to pointwise mutual information. The core idea behind is that words that are most likely to appear under the topic should appear in the same document simultaneously when considering a semantically consistent model. In addition, they also proposed that this

measurement standard has a good correlation with human judgments on the quality of topics. On the basis of semantic coherence, Roberts et al. (2014) found that exclusivity measure should be considered in topic number selection since by forming a few topics dominated by common words can easily achieve semantic coherence.

In order to form the initial number of topics for further analysis, we mainly conducted an evaluation of topic quality based on the semantic relevance and exclusivity of the generated topics. The semantic coherence of topic ***K*** that contains a list of ***M*** most probably words in topic) is proposed the related expressions (Equation 16), where $D(v_i, v_j)$ refers to that co-occurrence times of two words in a given document.

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log\left(\frac{D(v_i, v_j) + 1}{D(v_j)}\right) \quad (16)$$

Our measurement of exclusivity was based on the FREX labelling metric (Bischof and Airoldi, 2012; Bischof and Airoldi, 2016) that attempts to find words that distinguish topics. These words are with the balance of both frequent and exclusive, with calculation related to harmonic average (Equation 17). According to their definition, ECDF refers to the empirical CDF and ω refers to a weight with a default value as 0.7. Also, $ECDF\left(\frac{\beta_{k,v}}{\sum_{j=1}^{K}\beta_{j,v}}\right)$ is considered as the frequency score calculated by empirical ECDF of word v in the topic distribution of kth (Rdrr.io, 2019). Following Bischof and Airoldi (2012), words with high FREX scores must be high in both dimensions since it does not allow the compensation between dimensions.

$$FREX_{k,v} = \left(\frac{\omega}{ECDF\left(\frac{\beta_{k,v}}{\sum_{j=1}^{K}\beta_{j,v}}\right)} + \frac{1-\omega}{ECDF(\beta_{k,v})}\right)^{-1} \quad (17)$$

We evaluated the probable number of topics in the area by applying the user-specified initialization function in the R package called ***searchK***. We set the optimal topic numbers between the minimum value of 3 and the maximum value of 30, with the increment size of 3 topics in each step. In addition, the initialization type is set as 'Spectral' that applies a non-negative matrix decomposition of the word co-occurrence matrix (Roberts et al., 2016), and the topical prevalence is set to "~factor(rating) + p_q_difference + factor(general_brands) +s(day_count)". Then we computed and plotted four diagnostic values for Models with a different number of topics (***K***). Based on the balance of diagnostic semantic coherence, exclusivity, and held-out likelihood, we narrowed the range of the number of possible topics

given the set of 15 independent reviews from 15 sub-products under the 3 mobile phone brands. After that, we applied this method again to determine the number of topics ***K*** corresponding to each independent data set. Then, we selected and labelled the topic solutions for each sub-product, by using the FREX criterion and combing it with the topic quality and the topic correlation.

3.5 Combine STM with the Bass model

In previous steps, we obtained coefficients of innovation (p) and imitation (q) of the initial launched product from phones and the labelled topics of each review data set respectively. Then we plotted the differences of the labelled topics corresponding to the difference between p and q over time. This allows us to check whether features about emotional topics extracted from online review overtime utilising STM help to enhance the predictability of the success or failure of an initial launched products from the baseline model.

Usually, an online review contains one or more quality dimensions of the product, that is, each review is a mixture of one or more topics. Based on the regression model, the marginal effect on estimating the topic distribution from the textual solution from product reviews can be evaluated and compared. In this dissertation, the regression model is defined as follows:

$$Innovation\_Imitation\_Difference_i = \alpha ProductRating_i + \sum_{i=1}^{I}\sum_{k=1}^{K} \beta_i TopicProportion_{i,k} + \varepsilon_i$$

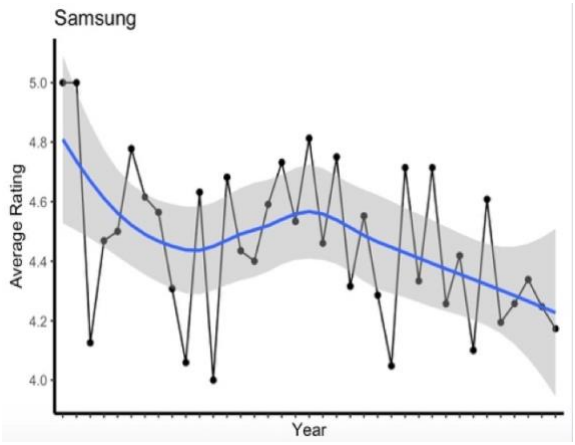
In the above regression, i refers to each specific review from the overall reviews I that $i \in \{1,2 \dots I\}$, and k is about the specific topic per corpus under overall topics K that $k \in \{1,2 \dots K\}$. The $Innovation\_Imitation\_Difference$ refers to the difference between the innovation coefficient and imitation coefficient, calculated as $\boldsymbol{p} - \boldsymbol{q}$. The positive value of $Innovation\_Imitation\_Difference$ is considered to be relatively successful initial product launch, while the negative value indicates that the release of the new product is a failure. The $ProductRating$ refers to factor variable about specific rating scores from each review from 1 to 5, and the $TopicProportion$ is the theta calculated from the STM model of the topic k under the review i. In order to illustrate the prediction effect, we added one topic variable at a time incrementally on the baseline model and compared the marginal changes. The explanatory power of the four topics with the most significant marginal influence is examined by this analysis.

4. Analysis and Results

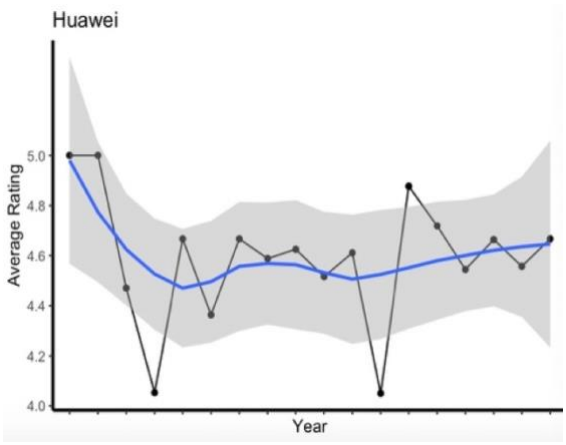
4.1 Overall Analysis

The original data is about 5,885 reviews of 15 products from the three major brands (Apple, Samsung, and Huawei) in the mobile phone category of Amazon UK. After pre-processing the reviews, 4,444 reviews are ready to be analyzed in STM and then they are combined with the Bass model for prediction. This includes 1900 reviews of Samsung (s8, s9, a10, a40, a70), 1642 reviews of Huawei (P Smart pro, Y5P, p40, p30, p30 pro, p30 lite), 902 reviews of Apple (iphone 7, iphone 8, iphone 11, iphone X, iphone XR).

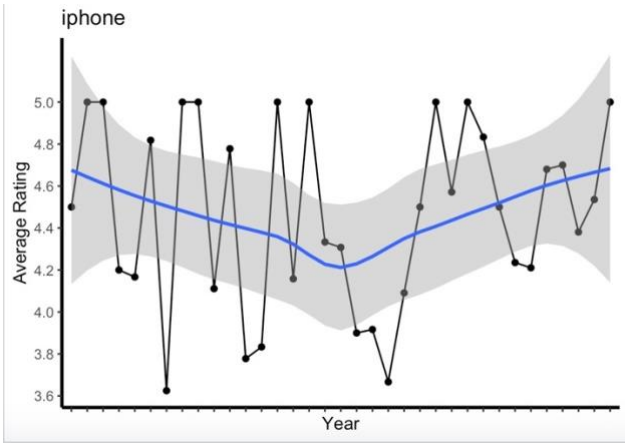
In the above reviews, the average ratings shown in reviews of different brands of mobile phones and the trend of rating changes over time are as shown in **Figure 4**. Since the data set of each mobile phone product is composed of 5 sub-products launched in different years, the average rating represents the consumer's overall recognition of the brand over a period of time, which is shown as the blue line in the figure with the maximum score. From the brand perspective, Samsung's overall average score has shown a downward trend in recent years, specifically from the initial 4.8 scores (the leftmost starting point is April 2017) to the recent 4.2 scores (the right side is June 2020). Besides, Apple's overall average score shows a downward trend and then an upward trend, with the overall score fluctuated greatly. As can be seen from the figure, Apple's average score is around 4.5 scores from the beginning (October 2016), and then it experienced a continuous drop to the lowest average score of 4.1 (July 2018). Recently, its average score has relatively improved to around 4.6 points (June 2020). Compared with the average score of Apple, the constant fluctuation of Huawei's average score level is relatively gentle, with the overall average score level at approximately 4.6 scores (from January 2019 to June 2020).



(a) Samsung



(b) Huawei



(c) Apple

Figure 4: Changes in Ratings of Different Mobile Phone Brands Over Time

4.2 Define the Optimal Number of Topics

As described in the methodology, we used the ***searchK*** method to find the optimal number of topics for the overall data set including Samsung, Huawei, and Apple, and the results are shown in the figures below. This search method was applied twice with the search range set differently. At the same time, we set the prevalence to "~factor(rating) + p_q_difference + factor(general_brands) +s(day_count)".

In the first search, we assumed that the optimal topic number is within the range from 2 to 40 and the increment step of 2 topics (**Figure 5**). Through observation, we narrowed the range of the optimal number of topics to 6 to 26 in the second attempt and set the increment step of 1 topic (**Figure 6**). We combined the three-dimensional value of each topic number, including semantic coherence, residual value, and held-out likelihood. We found that semantic coherence has achieved the local minimum when the number of topics is equal to 7, 11, 13, 16, 20 and the residuals have reached the local minimum value when the topic number is 8,13,15,18, 20. After comparing and checking the groups of FREX words under each optimal number of topics, we finally decided 13 topics as the most likely and suitable number of topics for STM model analysis.

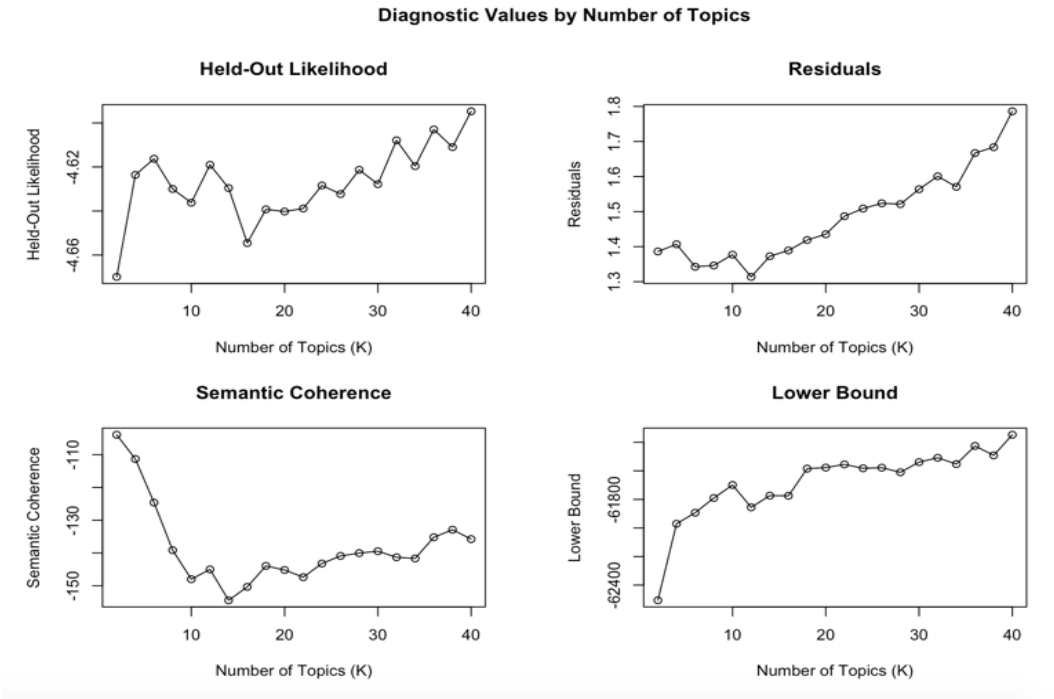


Figure 5: Determine the Optimal Number of Topics (Range From 2 to 40, Step = 2)

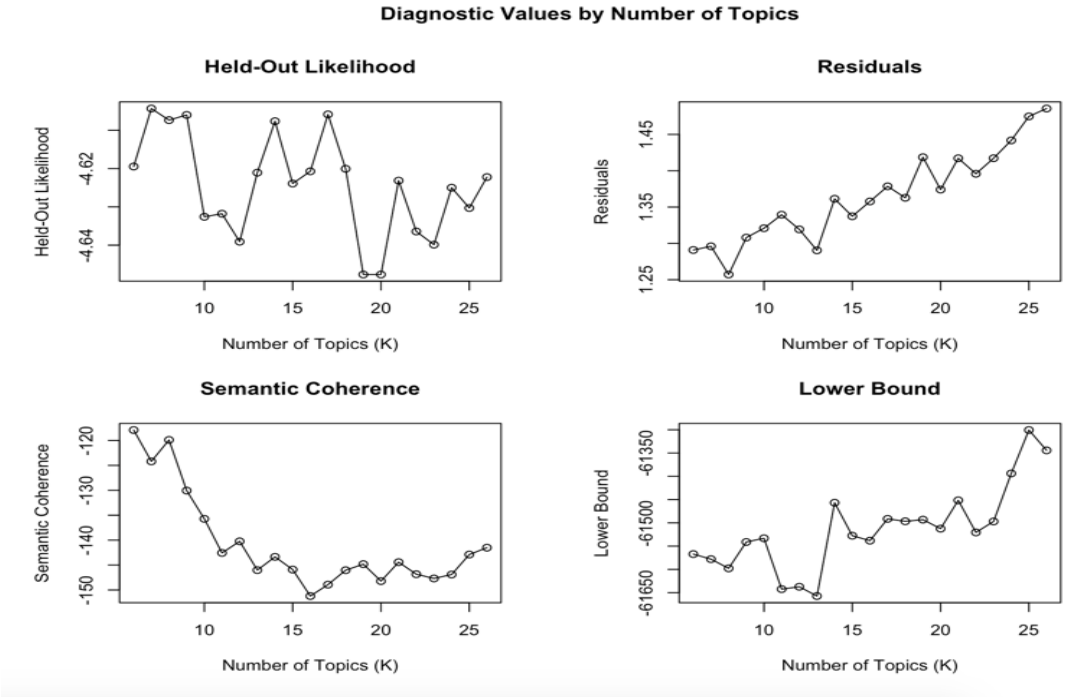


Figure 6: Determine the Optimal Number of Topics (Range From 6 to 26, Step = 1)

4.3 Application of STM

4.3.1 Semantic Coherence and Exclusivity

We utilized ***selectModel*** under the STM package to find the model of desirable properties. According to Roberts et al. (2016), ***selectModel*** first runs two EM steps in order to discard the model with low probability and then converges the maximum EM iteration on the default value of the 20% of the most possible model. With the prevalence function of "~factor(rating) + p_q_difference + factor(general_brands) +s(day_count)" and initialization type of "Spectral", which is the same value as the previous step of finding the optimal number of topics, we obtained the possible models. Then we visualized the results via ***plotModels*** in the STM package, which is designed to calculate the average value for topics within each model and draw these values with numbers.

As can be observed from **Figure 7**, combining the numerical distribution of semantic coherence (x-axis) and exclusivity (y-axis) at the same time, Model #3 (marked in light green) is considered relatively optimal since it is relatively located in the upper left corner of the chart.

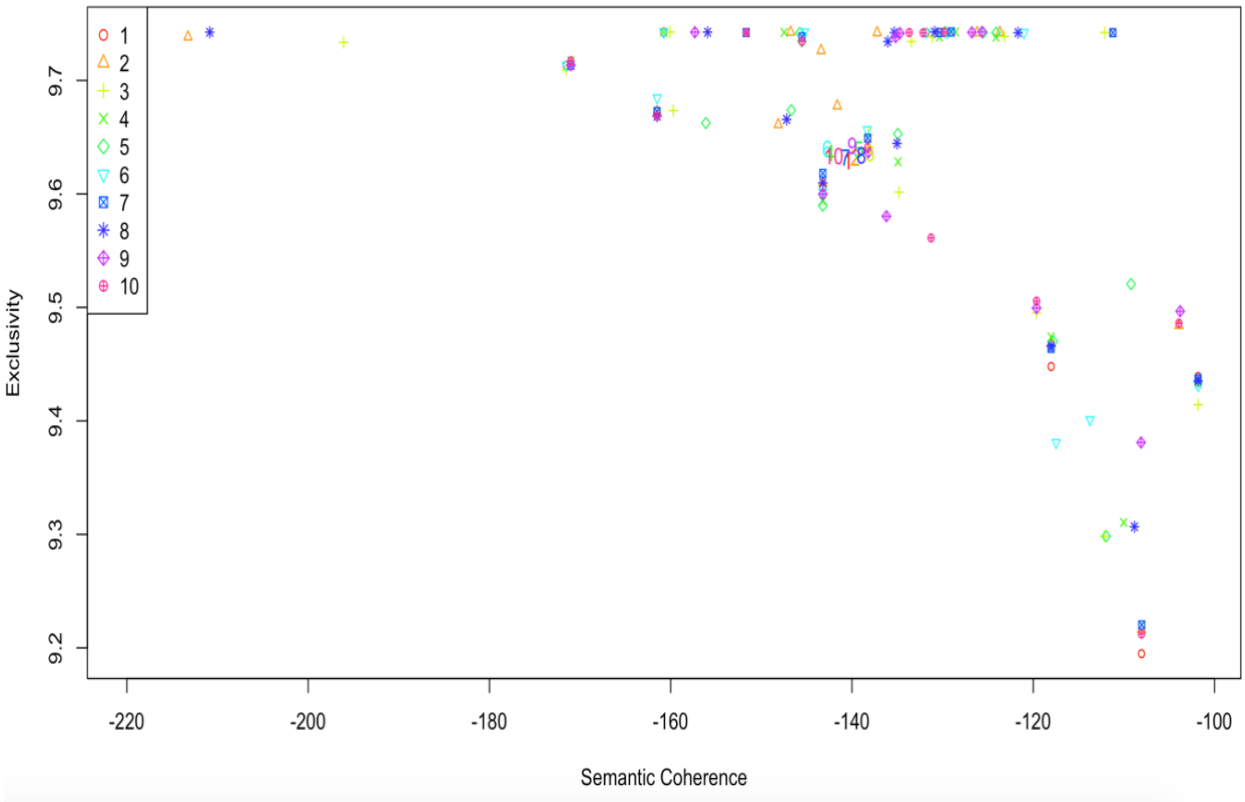


Figure 7: Optimal Model Selection with Both Semantic Coherence and Exclusivity

4.3.2 Topic Labels & FREX

Besides, the detailed results of this Model #3 can be viewed in **Appendix A**, which contains the four types of words under each topic, specifically Highest Prob, FREX, Lift, and Score. Among these four types of words, we mainly determined the topic label of each topic based on FREX (Bischof and Airoldi 2012; Airoldi and Bischof 2016), because it balances the frequency and exclusivity of words under the selected topic.

The label of topics in STM is determined based on the top 10 FREX words under the given topic. In addition, we tried to estimate the emotional expressions of different topics based on life experience. The sentiments related topics can be divided into 3 categories, including positive, neutral, and negative. For example, considering the principle of FREX, we found that word 'complaint' and 'fault' appears at a relatively high position, indicating the possibility of dissatisfaction of the product purchase. Thus, we labelled this topic with negative emotion.
In the same way, we also found other positive topics, such as Topic #2 related to the speed of operation and Topic #3 related to phone quality and customer service.

At the same time, the proportion of each topic in the textual content is also calculated and ranked from high to low according to the proportion in **Table *2***. As we observed from this table, in the reviews of the mobile phone market including Samsung, Huawei, and Apple, the topic with the largest proportion is about secondary features, accounting for about 13.05%. The second and third largest topics with larger proportions are phone call function related and price & complaint related, accounting for respectively 10.64% and 9.37%.

Table 2: Topic Solution for Online Reviews

#	Topic Label	Prop (%)	FREX Words (Top 10)
1	Purchase Experience	6.86	happy, product, purchase, mobile, reasonable, smartphone, replacement, time, handset, deal
2	Run & Responsiveness (Positive)	8.16	love, transfer, money, smart, responsive, switch, nice, photo, load, range
3	Customer Service & Quality (Positive)	7.39	quality, perfect, service, amazing, customer, quickly, item, sound, camera, time
4	Call Function Related	10.64	issue, finger, review, people, speaker, headphone, update, video, device, software
5	Phone Condition & Brand	5.19	please, condition, wife, earlier, extremely, scratch, brand, gift, standard, decent
6	Charging Function Related	8.64	day, week, couple, charge, month, hour, charger, half, return, seller

7	Model & System	6.96	model, highly, lite, storage, expensive, upgrade, system, cheaper, fine, user
8	Product Delivery	6.05	delivery, absolutely, language, instructions, impress, perfectly, package, original, change, upgrade
9	Touch Experience & Size	6.77	hand, button, edge, size, protector, light, touch, display, power, mode
10	Life & Battery	6.32	battery, life, speed, run, design, wireless, comment, camera, feature, difference
11	Price & Complaint Related (Negative)	9.37	fast, price, complaint, quick, market, fault, space, colour, time
12	Feature & Memory	13.05	easy, excellent, fantastic, camera, worth, brilliant, feature, memory, space, photo
13	Entertainment Features	4.62	cheap, performance, previous, miss, screen, lot, game, photo, feature, music

4.3.3 Plot Expected Topic Proportions

In addition, the visualization of ***plot.STM*** also gave us an intuitive display of the relative expected proportion of each topic by showing the three highest FREX of each topic and topic numbers (**Figure 8**). As indicated by this figure, the top 5 dimensions that consumers are mainly concerned about the phone product of Samsung, Huawei, and Apple are (1) feature and memory, (2) call function, (3) prices and consumer complaints, (4) charging performance, and (5) operation and response sensitivity. The above five aspects correspond to Topics #12, #4, #11, #6, and #2 respectively. At the same time, as indicated by the proportion of Topic #13, consumers will focus less on the performance of the entertainment features of the phone such as games and music. Considering the proportions of different topics presented above, consumers are more inclined to pay attention to the storage features, call quality, price tags, negative complaints of previous buyers, charging performance, and operating speed when purchasing mobile phones under the categories of Samsung, Huawei, and Apple.

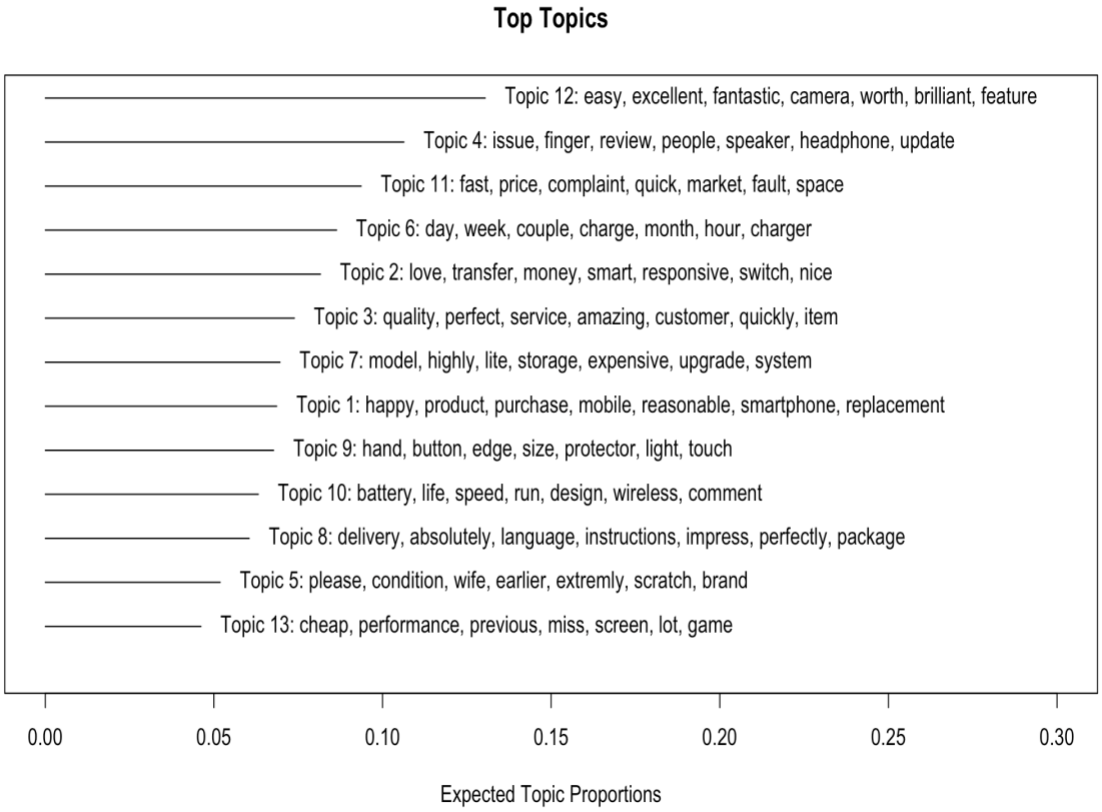


Figure 8: Expected Topic Proportions for Top Topics

4.3.4 Topic Correlations

While getting the proportion of each topic, the relationship between the topics can also be examined through the STM utilizing the function named ***plot.topicCorr***. As mentioned by Roberts et al. (2014), two topics are regarded as correlated if the link connected with them is above a given threshold (can also be self-specified). This can be achieved by an estimated marginal topic proportion correlation matrix. They also pointed out that topics with positive correlations suggest the likelihood of the co-occurrence of both topics discussed in the document. From **Figure 9**, we observed that each topic is related to at least three topics except itself. Among these topics, Topic #11'Price & Complaint Related' (located in the middle) has the largest number of pairwise positive contacts with other topics, specifically 7 other topics including topic related to 'Purchase Experience', 'Model & System', 'Product Delivery ', 'Phone Condition & Brand', 'Customer Service & Quality', 'Run & Responsiveness' and 'Feature & Memory'. In contrast, Topic #4 'Call Function Related 'and Topic #9 'Touch Experience & Size' (both located in the lower-left corner) have relatively few numbers of correlations to other topics, and they are both linked to three other topics. Neither of these two topics is directly related to Topic #11 'Price & Complaint Related'. This might suggest that Topic #4 and Topic

#9 are with relatively high exclusivity, and their semantic consistency between both of them and other topics is relatively low.

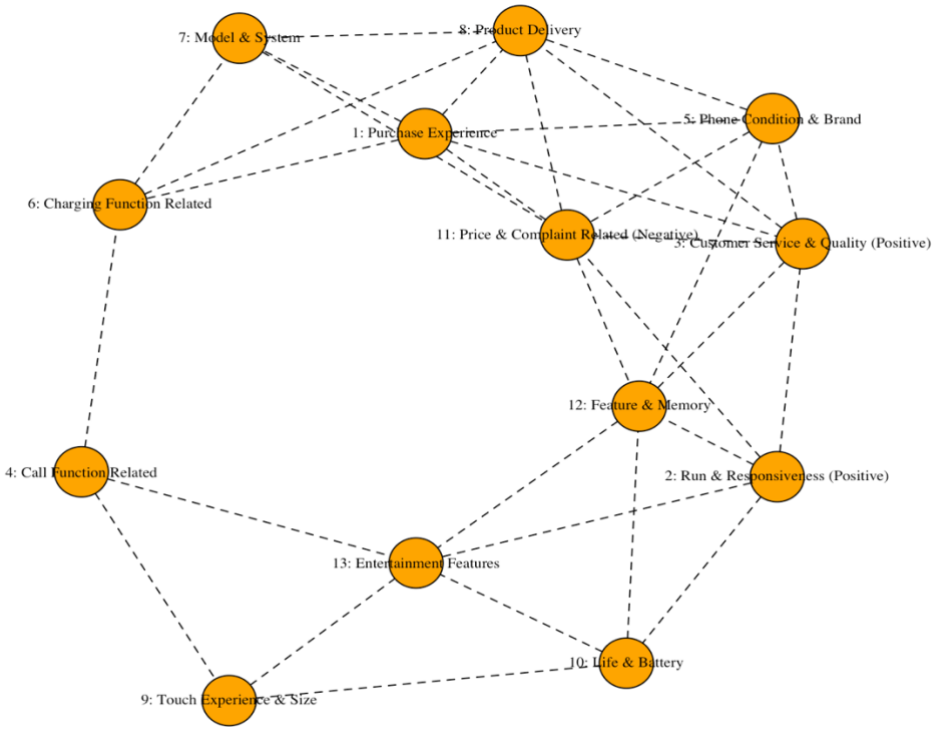


Figure 9: Correlation Between Different Topics

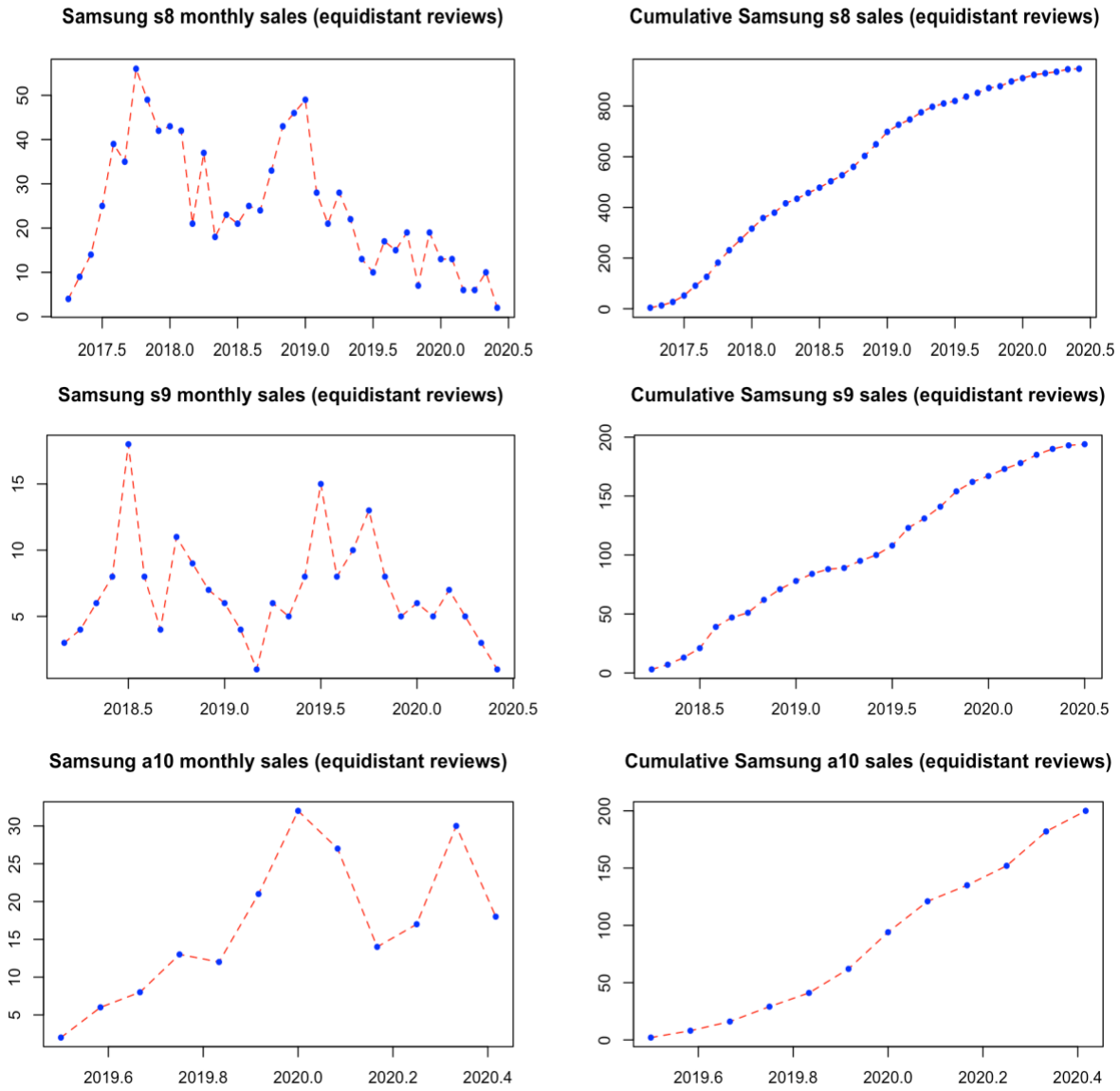
4.4 Obtain Innovation Coefficient (p) & Imitation Coefficient (q) From the Bass Model

We applied the Bass diffusion model to calculate the specific imitator coefficients (p) and innovator coefficients (q) for each online review from each of the sub-products under the Apple, Samsung, and Huawei brands (e.g. iphone11). Then we merged datasets of different sub-products together and estimated the relationship between topics and metadata on the basis of the combination of reviews and Bass model coefficients (see ***Section 4.5*** for details).

4.4.1 Estimate Monthly Sales and Cumulative Sales

By utilizing the Bass model, we created sales charts of their product reviews corresponding to the isometric sales for different sub-products. **Figure 10, Figure 11, Figure 12** shows the monthly sales and cumulative sales chart of 5 products of Samsung, Huawei and Apple respectively. The x-axis represents the time of reviews (in months), and the y-axis represents the number of reviews. According to the assumptions in the methodology, we assumed that the actual monthly sales and cumulative sales are multiples of the review numbers within a

given time. The blue dots in the figure represent the number of sales (reviews) per month. The larger the number of blue dots, the longer it has been since the product was released (the time range is from the first review of each product to 14 June 2020). Among all the products, iphone7 was the first to be released, and Samsung s8 was with the largest number of cumulative reviews. The monthly sales of all products fluctuate up and down over time, and such changes in sales can reflect they are in the different stages of their product life cycle.



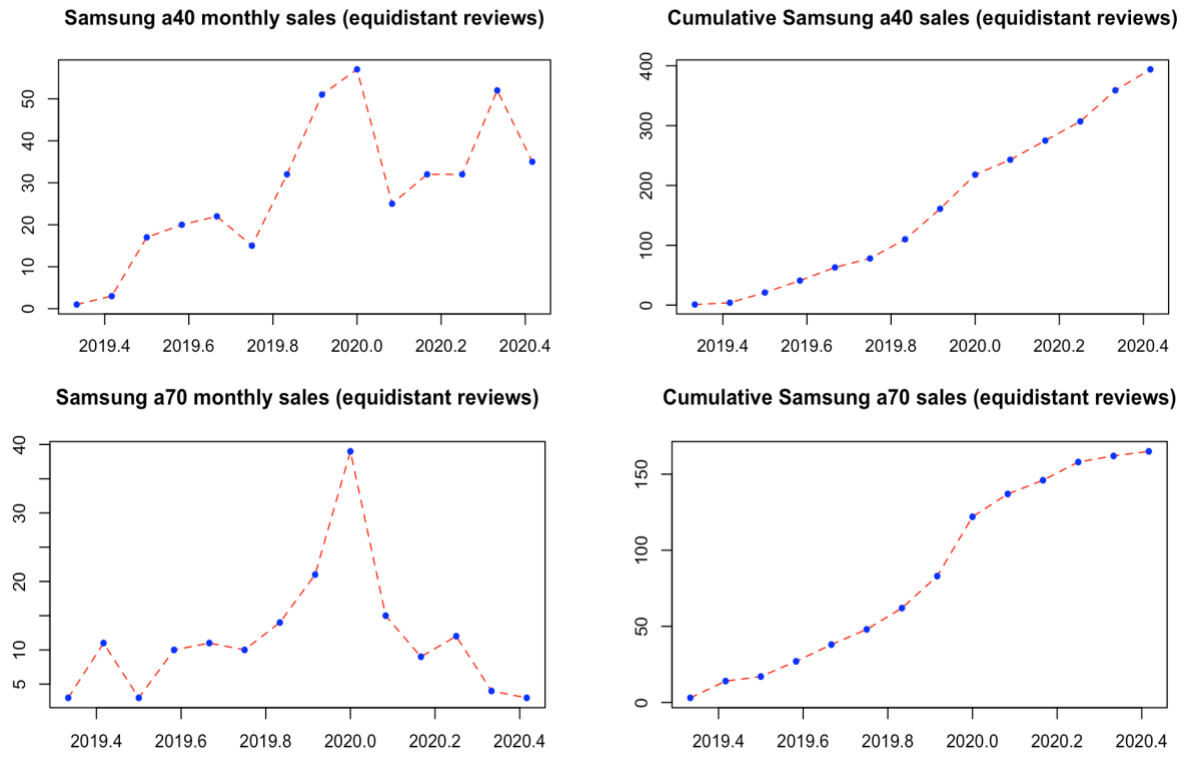
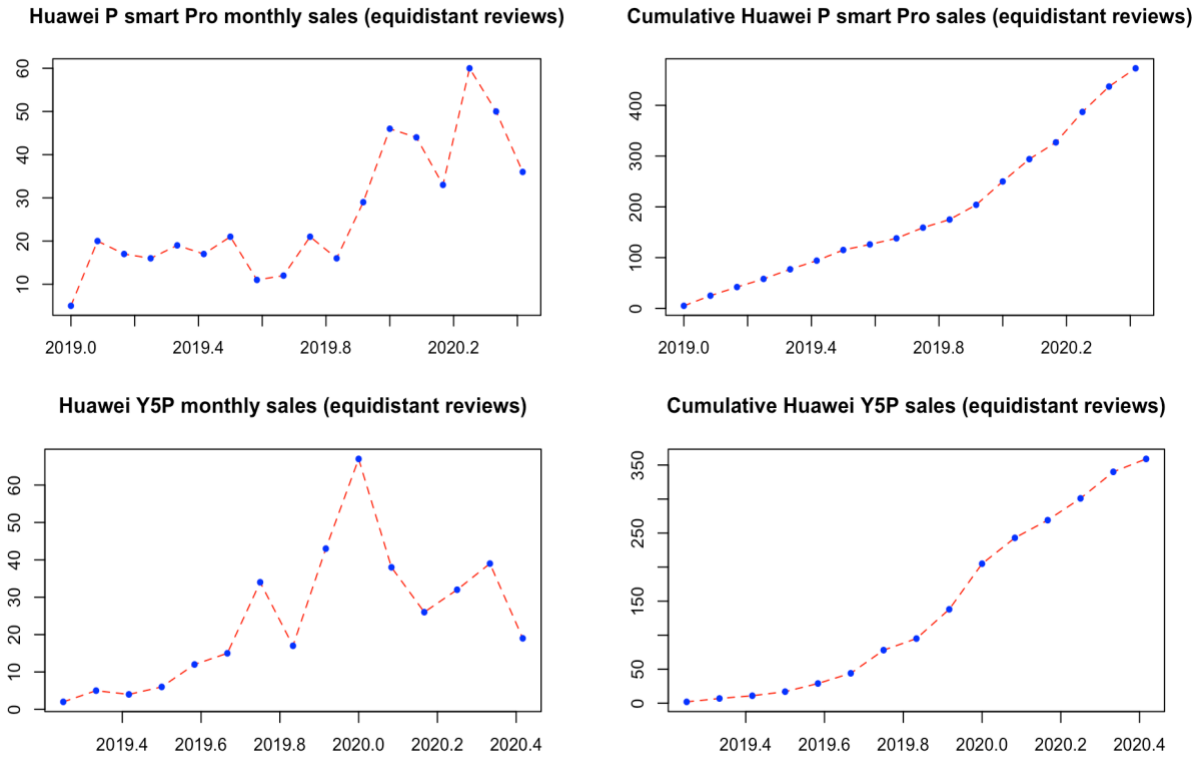


Figure 10: Estimated Monthly Sales and Cumulative Sales for Samsung



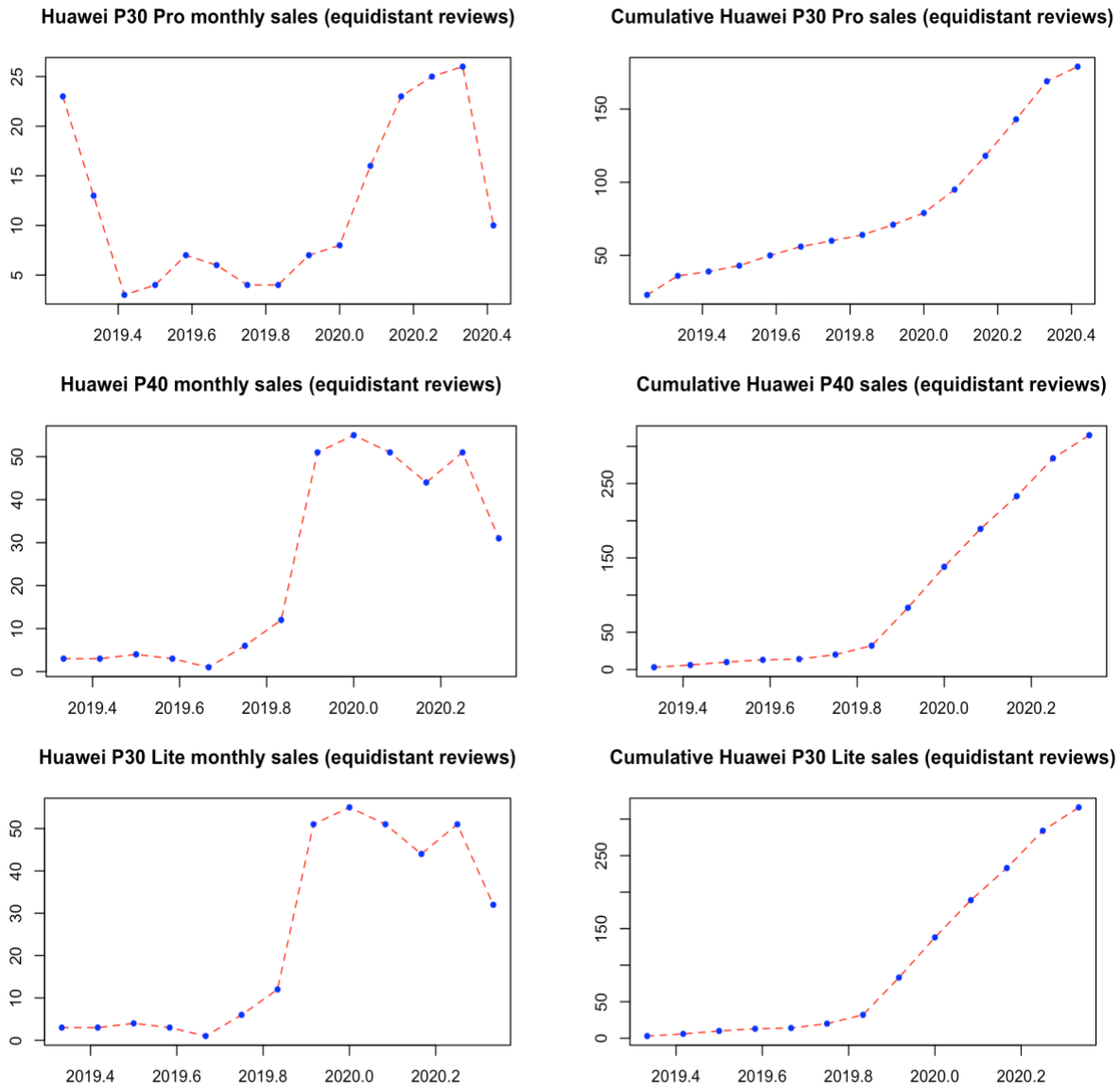
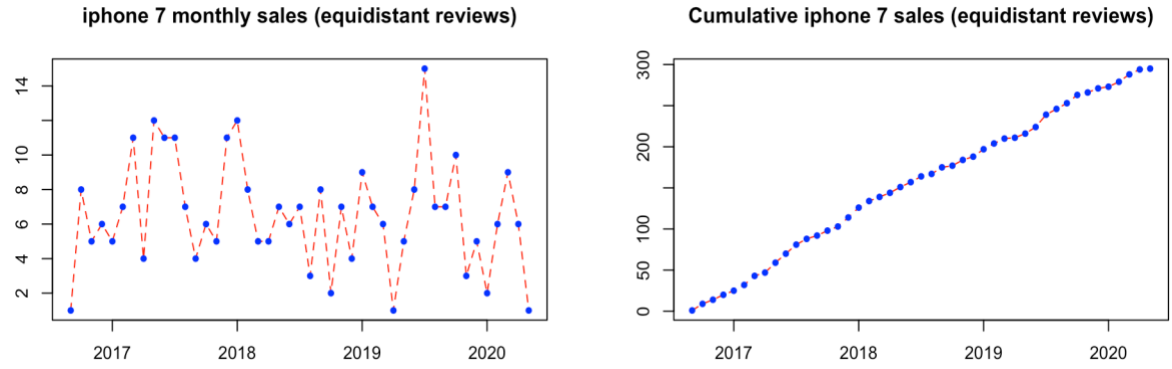


Figure 11: Estimated Monthly Sales and Cumulative Sales for Huawei



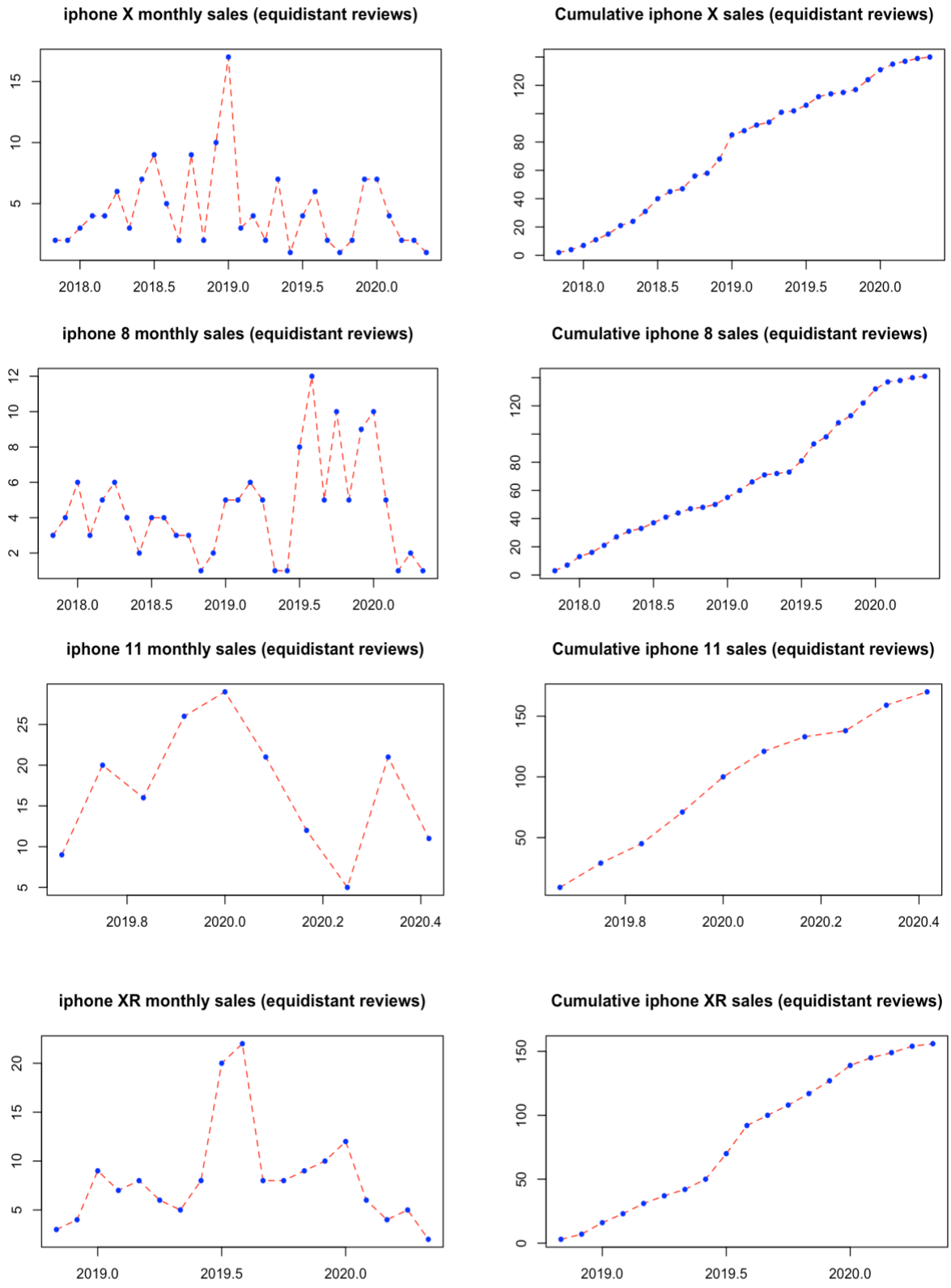


Figure 12: Estimated Monthly Sales and Cumulative Sales for Apple

4.4.2 Estimate the Coefficient of Innovator and Imitator

The Bass model (Bass, 1967) has been widely utilized in forecasting the diffusion of new products or technologies, with four important parameters including coefficient of innovation (p), coefficient of imitation (q), the capacity of the market, and the launch time of the product or technologies. Among these parameters, the values of p and q describe the interaction between previous buyers and potential adopters during the process of product adoption. As mentioned in previous sections, the Bass model considers a product is successful when the imitation coefficient is larger than the innovation coefficient ($p < q$). Conversely, a product is regarded as a failure ($p > q$).

Starting from the fourth review of the new product, we estimated the corresponding ***p, q*** values for each review according to the combination of chronological order, daily sales, and cumulative sales. Since there is a multiple relationships between the actual sales and the number of reviews filled in by customers, for example, there is only one review in a few days but the actual purchase amount within these days is x ($x \geq 1$), the values of p and q of some reviews are displayed as 'NA'. After calculating and excluding values displayed as 'NA', the obtained results of the coefficients of three brands are as follows (**Figure 13**). We can see that the number of reviews related to successful products far exceeds those related to failed products. This may be because we chose those mainstream mobile phone brands with a relatively large market share, and this means the number of reviews from the following products is relatively large compared with niche products. To a certain extent, the ratio of successful to failed product reflects the possibility of a large number of imitators for products from Samsung, Apple, and Huawei.

Success/Failure	Samsung	Apple (iphone)	Huawei
# of reviews related to success (p < q)	1466	392	811
# of reviews related to failure (p < q)	7	41	27
% of reviews related to success within own brand (p < q)	99.52 %	90.53%	96.78%
% of reviews related to failure within own brand (p < q)	4.8 %	9.47%	3.22%

Figure 13: Ratio of Reviews that Suggest the Success and Failure of the Products

4.5 Analysis of the Relationship Between Topic and Metadata in STM

According to Roberts et al., (2014), the key feature of the STM is to estimate the relationship between topics and meta. This is also recognized by Grimmer and Stewart (2013), who suggested that the relationship between these two aspects is with great importance in verifying the effectiveness of STM. In the ***stm*** package, the function named ***estimateEffect*** is used to extract the relationships and the uncertainty on overall ***K*** topics selected in the previous stage, with the output of a group of parameters in order to form a set of topics for further estimation. The prevalence used to estimate the parameters in ***estimatedEffect*** is the same as when we determined the topic in an earlier stage.

4.5.1 Overall p, q Difference on Topic Prevalence of Phone Segmentation

Since the covariate that we were interested in the case is binary: the success and failure of the product (obtained by p minus q), we set the method = ***'difference'***, covariate = "difference between p and q" to explore the change in the topic proportion from success ($p < q$) to failure ($p > q$). Combing the information obtained from topic analysis and the Bass model, we summarised and plotted the marginal effects obtained from STM on overall difference between innovation coefficients and imitation coefficients with the topic prevalence of phone segmentation.

After analysing the difference between the imitator and innovator coefficients of reviews written by purchased consumers from the three major mobile phone brands, we found an agreement concerning the importance of product features, price service and the value of consumer experience factor. As shown in the **Figure 14**, taking the call function as an example, per unit increase in the overall satisfaction is linked with approximately 1.1% increase in the discussion of specific topics 'Call Function Related'. According to Korfiatis et al. (2019), the dotted line indicates zero effect, and the possibility of a particular topic becomes more significant as the absolute value of their distance from the dashed axis increases. For example, dissatisfied product delivery ('Product Delivery'), unreasonable price setting & consumer complaint ('Price & Complaint Related'), incompatible systems ('Model & System') are vital factors that might cause product launch failure when ignored and unsolved by phone companies. On the contrary, great call function ('Call Function Related'), suitable touch experience and appropriate size ('Touch Experience & Size') are primary drivers for a successful product launch.

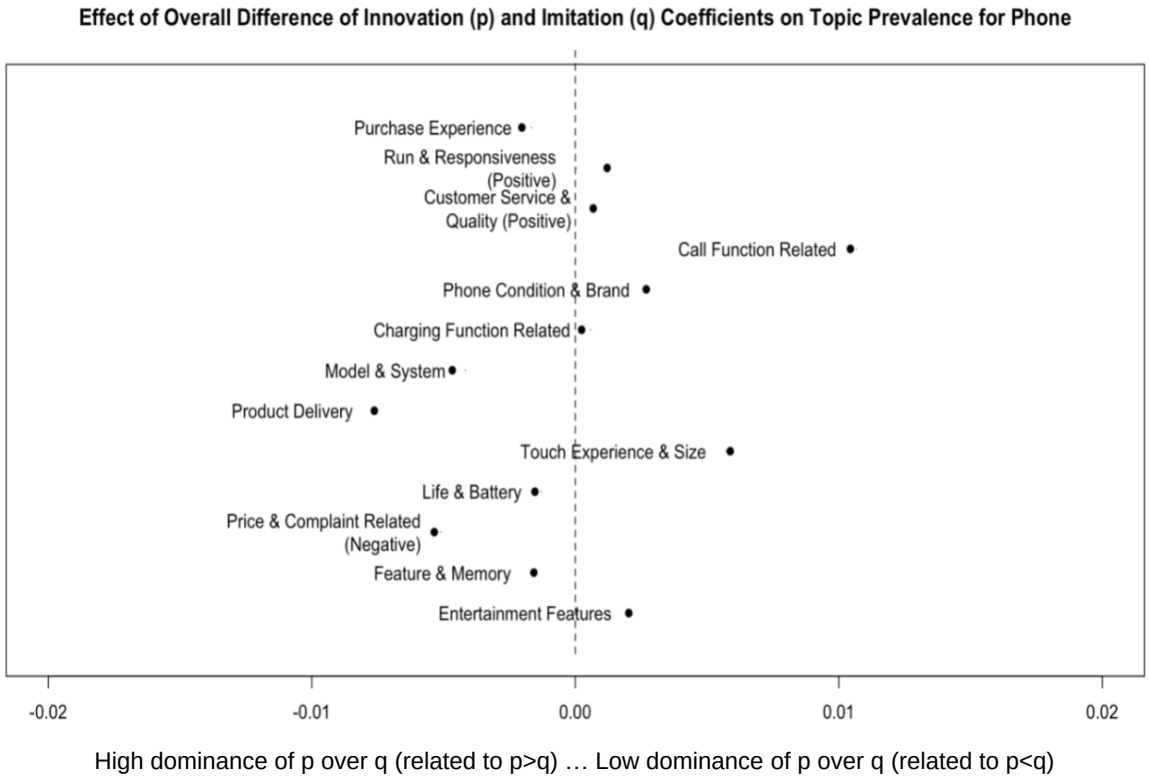


Figure 14: Marginal Effects of the Expected Proportion of Topic Prevalence Changes Based on the Ratio of the Coefficients p and q of the Bass Model (Low Dominance of p Over q and High Dominance of p Over q)

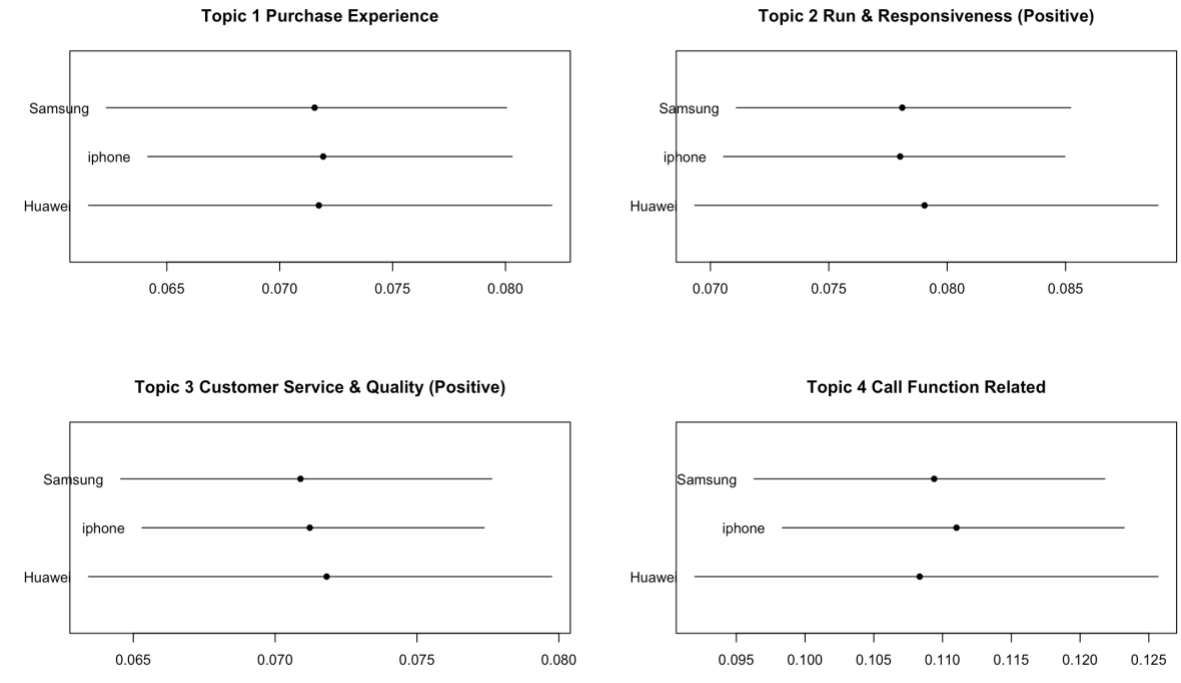
4.5.2 Overall Brand Difference on Topic Prevalence of Phone Segmentation

We were also interested in the marginal topic proportion value for each of the brand, thus we set the method = '***pointestimate***', covariate = "different brands" for estimation. As can be seen from **Figure 15**, consumers believe that three brands have similar performance in dimensions of phone purchase experience (Topic #1) and the touch screen valuation and the phone size (Topic #9).

At the same time, consumer reviews indicate that in the aspect of processing speed and responsiveness (Topic #2), Huawei's overall performance is better, while Samsung and Apple's processing performance is relatively similar. In terms of customer service purchases and product quality (Topic #3), consumers more satisfied with Huawei, while Samsung is relatively inferior. In addition, Huawei has overall advantages over Samsung and Apple in terms of storage space and external features (Topic #12) and entertainment functions (Topic #13) such as games, videos, etc. For the call-related evaluations (Topic #4) that accounted for the largest proportion of topics, Apple's overall performance was the best, with Samsung

and Huawei ranked second and third respectively. And in terms of brand recognition and phone condition (Topic #5), Apple and Huawei are relatively similar, while Samsung is slightly lower. For Samsung, the proportion of negative terms about price and consumer complaints (Topic #11) is greater than others, but it performs well in charging (Topic #6) and the life cycle of the phone (Topic #10).

Overall, if we ignore the aspect of product distribution related to Amazon, Huawei performs relatively well in other dimensions except for the call function and charging function, such as the aforementioned operating speed, memory, quality, and entertainment features. However, the call and charging function are of great importance in the overall evaluation of the mobile phone. If these two aspects can be improved, consumers may get more positive reviews from Huawei. Samsung ranks second in most dimensions, except for the worst performance in terms of price and consumer complaints. If Samsung can improve on certain aspects based on its existing dimensional advantages, it might gain more satisfaction from consumers. Apple's performance in basic functions such as call charging is relatively good, but the product life cycle is not highly evaluated. This may be due to the rapid update of Apple's new products every year and the low compatibility of the IOS.



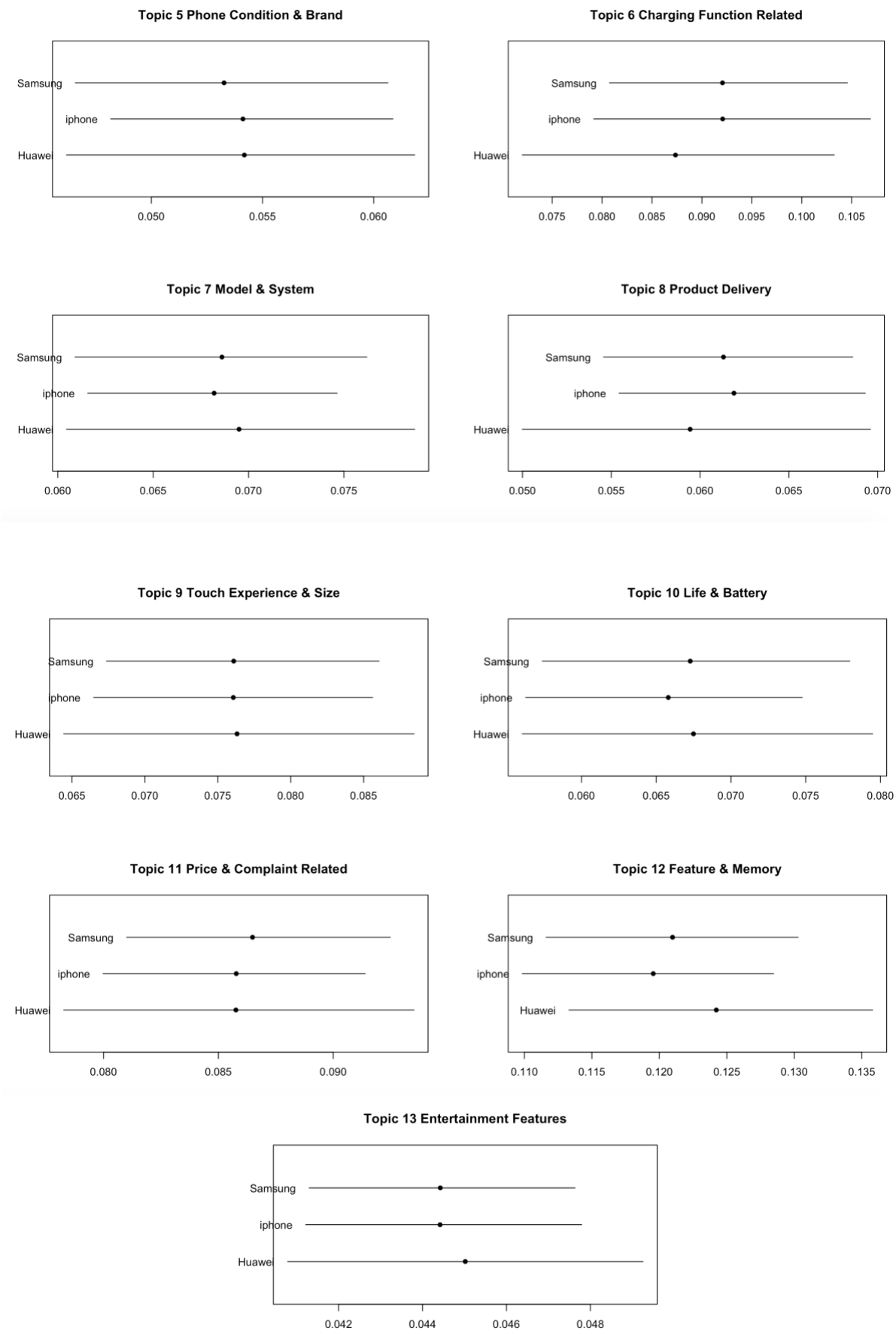


Figure 15: The Mean Topic Proportions of Topics for Each Brand

4.5.3 Extra Information & Perceptual Map

Values related to expected topic proportions can be obtained from previous steps, and it is shown in the following **Table 3** and plotted by the method of perceptual maps.

Table 3: The Proportion of Expected Topics Under the Three Mobile Phone Brands

#	Topic	Huawei	Apple	Samsung
1	Purchase Experience	0.0691	0.0680	0.0685
2	Run & Responsiveness (Positive)	0.0807	0.0833	0.0816
3	Customer Service & Quality (Positive)	0.0741	0.0740	0.0737
4	Call Function Related	0.1069	0.1057	0.1062
5	Phone Condition & Brand	0.0523	0.0518	0.0517
6	Charging Function Related	0.0877	0.0816	0.0870
7	Model & System	0.0694	0.0705	0.0695
8	Product Delivery	0.0608	0.0583	0.0609
9	Touch Experience & Size	0.0680	0.0688	0.0675
10	Life & Battery	0.0626	0.0642	0.0632
11	Price & Complaint Related (Negative)	0.0930	0.0939	0.0940
12	Feature & Memory	0.1294	0.1340	0.1301
13	Entertainment Features	0.0461	0.0468	0.0460

To compare the deviation estimation of impendence between product quality dimensions between Samsung, Huawei, and Apple, it is considered that these dimensions are orthogonal. In other words, this indicates that all brands have the same possibility of displaying product quality dimensions. Referring to **Figure 16**, some interesting findings can be drawn concerning the positioning of mobile phone brands in terms of product quality. For example, the topic concerning the purchase experience is orthogonal to most of the other dimensions of focus. This is relatively possible since consumers may pay more attention to the performance of the phone itself when measuring the phone quality. Some functional-related topics such as running and operating speed, call quality, and charging status may with a certain degree of correlation. It is suggested that Apple, Samsung, and Huawei are with their advantages in different dimensions. Thus, these companies can refer to the quality characteristics of their products in different dimensions as shown in the online reviews and make a comprehensive

consideration in the design of the new products. This might help to gain more consumer recognition of their new products and thus occupy a dominant position in the phone market.

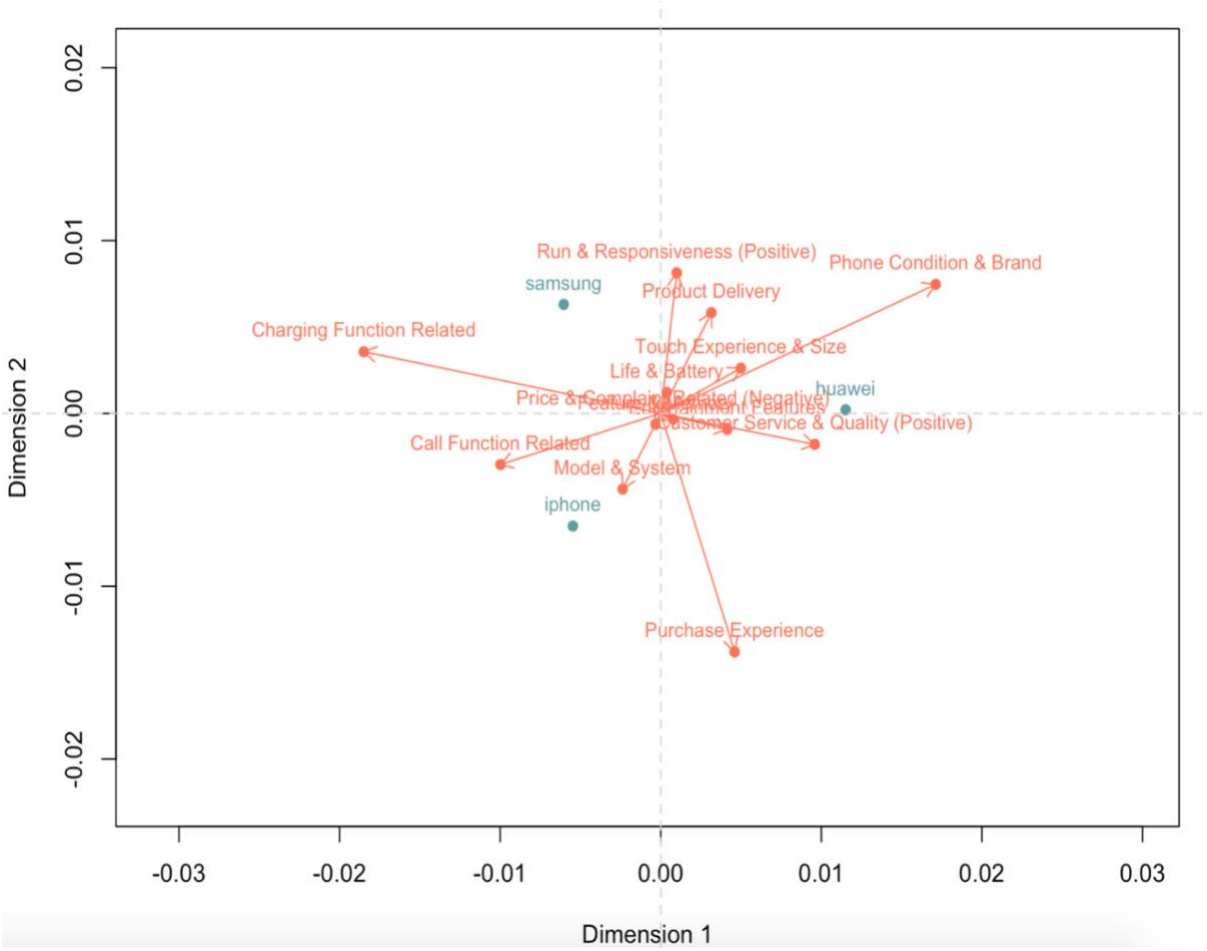


Figure 16: Market Positions and Satisfaction of Product Quality for Leading Phone Brands

4.6 Estimation on Temporal Topics Change

We estimated the proportion of topic change over time and selected several representative topics for analysis (the rest are in the **Appendix B**). Since each product is with a different release date, we set the x-axis to the time interval (unit: day) of each comment from the first comment, and the y-axis to the proportion of the estimated topics. The dotted line represents the upper and lower limits. This estimate is based on 15 products under 3 brands, and the overall period time is 1146 days.

We found that around 300 days (about 1 year) after the first product review, the proportion of most topics fluctuated up and down. Specifically, the proportion of topics such as charging function, screen touch experience & size, and entertainment features declined in around 200

to 300 days and increased in around 300 to 400 days. In contrast, the proportion of topics related to the life cycle and battery condition, memory and storage space increase at around 200 to 300 days, and then decrease around 300 to 400 days.

Among these different topics, the expected proportion of Topic 'Model & System' increases gradually with a relatively slow increasing rate. This might indicate that consumers pay more attention to model versions and operating systems of phones as time goes by. According to **Figure 17**, the main increase in the proportion of the topics is concentrated in the 800 to 1,200 days after the current new product is released. This suggests that consumers might expect their phones with upgraded model versions or operating systems in a period time after product launch. Thus, we could observe that to enhance user experience and gaining a larger market share, different mobile phone manufacturing companies release new phone models regularly and are willing to upgrading the current operating system with the latest version.

Besides, the expected proportion of Topic 'Life & Battery' showed a downward trend after small fluctuations in the first 800 days. This may indicate that consumers focus on product endurance and the length of service life during the initial introduction and growth phase of the product life cycle. However, as time goes by, the product might gradually enter the maturity stage (around 800 days or equivalent to about 2 years). As suggested by (Ng, 2019), the average mobile phone usage time of users in the UK in 2018 was about 27.7 months. At around two years after purchasing, consumers' expectations for the aspect of batteries and products are reduced, including change their phones to buy a new one and lower their expectations for battery capabilities. This leads to the decrease in the percentage of expected topic proportion.

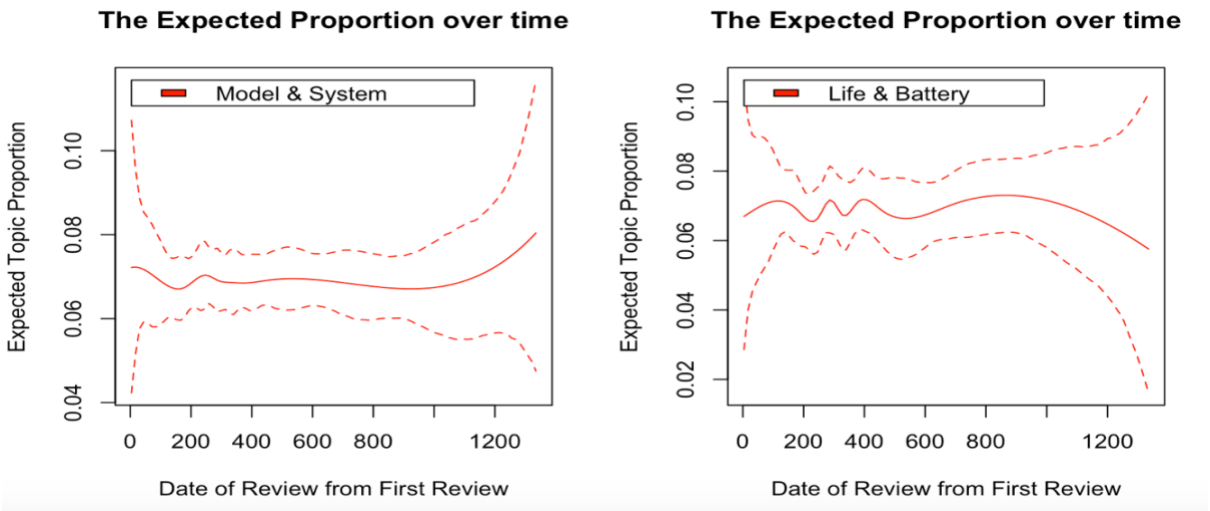


Figure 17: The Expected Proportions of Topics Over Time (95% Confidence Intervals)

4.7 Result Analysis of Regression Model

Linear regression was applied for regression analysis, which performs well in examining the relationship between a scalar response and explanatory variables. We estimated the importance of each predictor variable to affect the success and failure of the product based on the parameters of the regression. The baseline model was designed as exploring the relationship between the difference between ***p*** and ***q*** and the product rating score, and we gradually added different topic proportions as new predictors. The contributions of the four topic proportions of relatively high marginal effects on the dependent variable (p, q difference) from multiple testing could partly explain whether the product launch is regarded as successful or not. After calculating the interpretation related parameter R_2 and the significance of the corresponding parameters by rating scores and topic proportions under the three major mobile phone brands, we found that the forecasting model of is relatively ideal (**Figure 18**) compared with that of Samsung (**Figure 19**) and Apple (**Figure 20**).

For Huawei, in the gradual superposition of different topic proportions, four topics are chosen for final regression based on their marginal impact on the success of the product. These topics are labelled as Customer Service & Quality (Positive) (Topic #3), Touch Experience & Size (Topic #9), Charging Function Related (Topic #6), and Purchase Experience (Topic #1). As shown in **Figure 18**, topics about comprehensive sales and after-sales services related to the quality of mobile phones, as well as comfortable touch experience with suitable phone sizes will promote the success of new products. At the same time, the review mentioned the charger and the charging capacity is also positively related to product success. However, if the purchase experience is mentioned in the review at the same time, it would have a negative impact on the success of the product to a certain extent as indicated by the negative coefficient. Besides, the relatively important or dominant independent variables were marked with *****, such as Topic #3, Topic #6, Topic #1 and the rating score. In addition, we found that the ratings are positively related to Huawei's success. The higher the rating, the more likely the product is to succeed. For Apple and Samsung, the rating may be less important when combining with these four topics, since without ***** behind.

For Samsung and Apple (**Figure 19, Figure 20**), the four above mentioned topics proportion and rating score might be with low marginal effects to their initial product launch. Apart from these four topics, we also added other topic proportions separately to these two brands. However, when adding one or two topics on the baseline model, we did not find a remarkable topic with their proportion considered of high marginal effect. This may indicate that there may be other uncategorized topics that are more dominant if adopted a different topic models; or

for these two brands, besides the influence of the topic, there are other variable factors that are closely related to the success of the product.

At the same time, as concluded from the regression model that the marginal impact of changes in the expected proportion of topics in online rating and their corresponding reviews on the product success of Apple and Samsung is not obvious, but the impact on Huawei is relatively significant. When reviews are mentioned more about satisfying sales services related to mobile phone quality, touch experience with suitable phone size, and charging method and charging capacity, the probability of product success in Huawei will increase. However, if the purchase experience is mentioned simultaneously, it will have a negative impact on product success to a certain extent.

Regression Result for Huawei

	Dependent variable:				
	p_q_diff				
	(1)	(2)	(3)	(4)	(5)
rating	0.008**	0.008**	0.008**	0.008**	0.008**
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
topic_3		0.359***	0.555***	0.704***	1.042***
		(0.133)	(0.164)	(0.200)	(0.244)
topic_9			0.308**	0.411**	0.289
			(0.150)	(0.170)	(0.177)
topic_6				0.141	0.224**
				(0.108)	(0.113)
topic_1					-0.526**
					(0.219)
Constant	-0.047**	-0.081***	-0.121***	-0.154***	-0.149***
	(0.019)	(0.022)	(0.030)	(0.039)	(0.039)
Observations	816	816	816	816	816
R2	0.005	0.014	0.019	0.021	0.028
Adjusted R2	0.004	0.012	0.016	0.017	0.022
Residual Std. Error	0.111 (df = 814)	0.111 (df = 813)	0.111 (df = 812)	0.111 (df = 811)	0.110 (df = 810)
F Statistic	4.429** (df = 1; 814)	5.860*** (df = 2; 813)	5.327*** (df = 3; 812)	4.424*** (df = 4; 811)	4.712*** (df = 5; 810)
Note:					*p<0.1; **p<0.05; ***p<0.01

Figure 18: Regression Results of Huawei's Reviews

Regression Result for Samsung

	Dependent variable:				
	p_q_diff				
	(1)	(2)	(3)	(4)	(5)
rating	0.001	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
topic_3		-0.037	-0.059	-0.064	-0.053
		(0.064)	(0.079)	(0.097)	(0.117)
topic_9			-0.031	-0.034	-0.038
			(0.066)	(0.076)	(0.079)
topic_6				-0.004	-0.002
				(0.050)	(0.052)
topic_1					-0.016
					(0.102)
Constant	-0.015**	-0.011	-0.007	-0.006	-0.006
	(0.007)	(0.009)	(0.013)	(0.018)	(0.018)
Observations	1,403	1,403	1,403	1,403	1,403
R2	0.001	0.001	0.001	0.001	0.001
Adjusted R2	-0.0002	-0.001	-0.001	-0.002	-0.003
Residual Std. Error	0.063 (df = 1401)	0.063 (df = 1400)	0.063 (df = 1399)	0.063 (df = 1398)	0.063 (df = 1397)
F Statistic	0.722 (df = 1; 1401)	0.526 (df = 2; 1400)	0.426 (df = 3; 1399)	0.321 (df = 4; 1398)	0.262 (df = 5; 1397)
Note:					*p<0.1; **p<0.05; ***p<0.01

Figure 19: Regression Results of Samsung's Reviews

Regression Result for Apple

	Dependent variable:				
	p_q_diff				
	(1)	(2)	(3)	(4)	(5)
rating	-0.003	-0.003	-0.003	-0.003	-0.003
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
topic_3		0.009	-0.143	-0.285	-0.081
		(0.217)	(0.274)	(0.338)	(0.387)
topic_9			-0.303	-0.425	-0.591
			(0.333)	(0.374)	(0.404)
topic_6				-0.122	-0.088
				(0.171)	(0.174)
topic_1					-0.411
					(0.378)
Constant	0.016	0.015	0.050	0.082	0.098
	(0.022)	(0.030)	(0.049)	(0.067)	(0.068)
Observations	392	392	392	392	392
R2	0.001	0.001	0.003	0.004	0.007
Adjusted R2	-0.002	-0.004	-0.005	-0.006	-0.006
Residual Std. Error	0.120 (df = 390)	0.120 (df = 389)	0.120 (df = 388)	0.120 (df = 387)	0.120 (df = 386)
F Statistic	0.313 (df = 1; 390)	0.157 (df = 2; 389)	0.380 (df = 3; 388)	0.413 (df = 4; 387)	0.567 (df = 5; 386)
Note:					*p<0.1; **p<0.05; ***p<0.01

Figure 20: Regression Results of Apple's Reviews

5. Conclusion

5.1 Contributions

This dissertation aims to provide insights into the success or failure of initial product launches in the consumer durables market segment. Such target is realized by applying the structural topic model on the online reviews analysing the changes of the topic proportion discussed and combining that with the results from Bass model, specifically referring to calculating the parameters related to the success of the product for the published reviews, to make further regression predictions. Previous research areas have mostly focused on exploring the use of the Bass model in combination with the user ratings of online reviews and the emotional polarity of reviews to predict overall product sales. Compared with previous research, our dissertation is a meaningful exploration that a model for predicting the success and failure of the newly launched product is proposed by combing the Topic model (STM) and the Bass model.

From the 13 quality dimensions extracted from amazon online review that cover three mainstream phone brands of Apple, Huawei, and Samsung, we found the top three topics that account for the largest proportion of all these sampled reviews are the memory, storage space-related features, call-function evaluation, and complaints about price tags and cost performance. Among the expected proportion of the overall success of the product, features related to the call function, touch experience, and size are the two most important dimensions when consumers share their reviews. On the contrary, the expected proportion of features related to distribution logistics and delivery, complaints about price scale, and models and operating systems are considered of poor performance that linked to the overall failure of the product. Over a period of time after usage, consumers may pay more attention to models and operating systems, while their concerns about service life and battery performance have decreased.

In terms of sampling, the cumulative purchase volume of Samsung is relatively the largest, and the overall rating of Huawei is the highest. In general, Huawei has a relatively higher satisfaction with customer purchases and after-sales services and has a relatively good reputation and brand recognition. However, further improvements are needed in the basic functions of Huawei's mobile phones, such as call and charging functions. Relatively speaking, Apple performs better in terms of the basic functions, but it may require more effort and cost in terms of memory capacity and operating speed. In addition, the high pricing level of Apple

might affect its brand recognition to a certain extent. Samsung meets consumer expectations in terms of touch feel and size to a certain extent, but it ranks second in most indicators.

5.2 Implications

This dissertation is applicable to market analysts and product managers from companies. By utilizing the Structural Topic Models throughout the product life cycle, market analysts or project managers could continuously monitor the user-generated feedback related to the dynamic consumers perception of product quality dimensions. Such perception helps to grasp the orientation and demand of the market on the basis of understanding consumer preferences and tendencies. Through the analysis of different quality dimensions, companies could maintain a leading position in the market, by enhancing their competitive advantages in the dominant dimension of products and make improvements on shortcomings.

5.3 Limitations

This dissertation has a set of limitations that can be solved and improved in future research. First, we mainly analyzed the online reviews of the mobile phone segmentation from the three mainstream brands on Amazon in the UK. Our further study could collect and compare data in multiple review aggregators to obtain insights, such as Amazon UK or eBay. Secondly, we may be relatively subject to determining the topic labels due to the lack of professional suggestions on durable goods. In the future study, we can seek the opinions of experts in the consumer product research market to define topic labels. Third, this dissertation only studies reviews of mobile phone products. Future research may apply this combination of the topic model and the Bass model on other consumer durables. Fourth, due to the lack of real sales data as there is a certain linear relationship between reviews and sales, we did not cross-validate the regression model. Future research can combine accurate sales data for regression and modelling.

References

Airoldi, E. and Bischof, J. (2016) Improving and Evaluating Topic Models and Other Models of Text. Journal of the American Statistical Association, 111(516), 1381–1403.

Archak, N., Ghose, A. and Ipeirotis, P. G. (2011) Deriving the pricing power of product features by mining consumer reviews. Management Science, 57(8), 1485–1509.

Bass, F. (1969) A new product growth for model consumer durables. Management Science. 15 (5): 215–227. doi:10.1287/mnsc.15.5.215.

Bass, F., Krishnan, T. and Jain, D. (1994) Why the Bass Model Fits without Decision Variables. Marketing Science. 13 (3): 203–223.

Bischof, J. and Airoldi, E. (2012) Summarizing topical content with word frequency and exclusivity. Proceedings of the 29th International Conference on Machine Learning (ICML-12), ICML '12, pp. 201–208.

Blei, D. and Lafferty, J. (2007) A correlated topic model of Science. The Annals of Applied Statistics, 1(1), pp.17-35.

Blei, D., Ng, A. and Jordan, M. (2003) Latent Dirichlet allocation. Journal of Machine Learning Research, 3(2003), 993–1022.

Chevalier, J. and Mayzlin, D. (2006) The Effect of Word of Mouth on Sales: Online Book Reviews. Journal of Marketing Research, 43(3), pp.345-354.

Cui, G., Lui, H. and Guo, X. (2012) The Effect of Online Consumer Reviews on New Product Sales. International Journal of Electronic Commerce, 17(1), pp.39-58.

Das, S., (2017) Data Science: Theories, Models, Algorithms, And Analytics. [online] GitHub. Available at: <https://srdas.github.io/MLBook/productForecastingBassModel.html#symbolic-math-in-r> [Accessed 7 July 2020].

Dellarocas, C., Zhang, X. and Awad, N. (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. Journal of Interactive Marketing, 21(4), pp.23-45.

Fan, Z., Che, Y. and Chen, Z. (2017) Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. Journal of Business Research, 74, pp.90-100.

Fitzsimmons, J., Douglas, E., Steffens, P. and Thomond, P. (2007) Marketing Buzz: Towards a Framework for Entrepreneurs. In Proceedings AGSE Entrepreneurship Exchange, QUT, Brisbane.

Geletta, S., Follett, L. and Laugerman, M. (2019) Latent Dirichlet Allocation in predicting clinical trial terminations. BMC Medical Informatics and Decision Making, 19(1).

Godes, D. and Mayzlin, D. (2004) Using online conversations to study word of mouth communication. Marketing Science, 23(4), 545–560.

Grimmer, J. and Stewart, B. (2013) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), pp.267-297.

Guo, Y., Barnes, S. and Jia, Q. (2017) Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management, 59, pp.467-483.

Hu, N., Liu, L. and Zhang, J. (2008) Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. Information Technology and Management, 9(3), pp.201-214.

Huang, P., Lurie, N. and Mitra, S. (2009) Searching for experience on the web: An empirical examination of consumer behavior for search and experience goods. Journal of Marketing, 73(2), pp.55–69.

Kanungsukkasem, N. and Leelanupab, T. (2019) Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction. IEEE Access, 7, pp.71645-71664.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P. and Sagiadinos, V. (2019) Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. Expert Systems with Applications, 116, pp.472-486.

Lee, E. and Shin, S. (2014) When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo. Computers in Human Behavior, 31, pp.356-366.

Lee, J., Jung, S. and Park, J. (2017) The role of entropy of review text sentiments on online WOM and movie box office sales. Electronic Commerce Research and Applications, 22, pp.42-52.

Li, M., Huang, L., Tan, C. and Wei, K. (2011) Assessing the Helpfulness of Online Product Review: A Progressive Experimental Approach. In Proceedings of PACIS.

Lilien, G.L., Rangaswamy, A. and Bruyn, A. (2007) The Bass Model: Marketing Engineering Technical Note. [online] Faculty.washington.edu.
Available at: http://faculty.washington.edu/sundar/NPM/BASS-Forecasting%20Model/Bass%20Model%20Technical%20Note.pdf> [Accessed 8 July 2020].

Liu, Y. (2006) Word-of-mouth for movies: Its dynamics and impact on box office revenue. Journal of Marketing, 70(3), 74–89.

Liu, Y., Bi, J. and Fan, Z. (2017) Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. Information Fusion, 36, pp.149-161.

Massiani, J. and Gohs, A. (2015) The choice of Bass model coefficients to forecast diffusion for innovative products: An empirical investigation for new automotive technologies. Research in Transportation Economics, 50, pp.17-28.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. and McCallum, A. (2011) Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 262-272). Association for Computational Linguistics. Chicago

Ng, A. (2019) Smartphone Users Are Waiting Longer Before Upgrading — Here's Why. [online] CNBC. Available at: <https://www.cnbc.com/2019/05/17/smartphone-users-are-waiting-longer-before-upgrading-heres-why.html> [Accessed 10 August 2020].

Norton, J. and Bass, F. (1987) A Diffusion Theory Model of Adoption and Substitution for Successive Generations of High-Technology Products. Management Science. 33 (9): 1069–1086

Onishi, H. and Manchanda, P. (2012) Marketing activity, blogging and sales. International Journal of Research in Marketing, 29(3), pp.221-234.

Rémy, P. (2018) Amazon Multi Language Reviews Scraper. [online] GitHub. Available at: <https://github.com/philipperemy/amazon-reviews-scraper> [Accessed 5 June 2020].

Rdrr.io. (2019) Label Topics Stm: Estimation of The Structural Topic Model. [online] Available at: <https://rdrr.io/cran/stm/man/labelTopics.html> [Accessed 17 August 2020].

Roberts, M., Stewart, B. and Airoldi, E. (2016) A Model of Text for Experimentation in the Social Sciences. Journal of the American Statistical Association, 111(515), pp.988-1003.

Roberts, M., Stewart, B. and Tingley, D. (2019) stm: An R Package for Structural Topic Models. Journal of Statistical Software, 91(2).

Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B. and Rand, D. (2014). Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58(4), pp.1064-1082.

Tellis, G. J., Yin, Y. and Niraj, R. (2009) Does Quality Win? Network Effects Versus Quality in High-Tech Markets. Journal of Marketing Research, 46 (April), 135–49.

Tirunillai, S., and Tellis, G. J. (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. Journal of Marketing Research, 51(4), 463–479.

Appendix A - Top 7 Words Associated with Each Topic

Topic 1 Top Words:

Highest Prob: happy, product, purchase, mobile, replacement, smartphone, reasonable
FREX: happy, product, purchase, mobile, reasonable, smartphone, replacement
Lift: happy, product, purchase, mobile, reasonable, smartphone, replacement
Score: reasonable, happy, product, purchase, mobile, replacement, smartphone

Topic 2 Top Words:

Highest Prob: money, love, nice, galaxy, photo, transfer, smart
FREX: love, transfer, money, smart, responsive, switch, nice
Lift: transfer, round, smart, responsive, love, switch, daughter
Score: transfer, love, money, nice, daughter, galaxy, responsive

Topic 3 Top Words:

Highest Prob: quality, amazing, perfect, service, quickly, item, sound
FREX: quality, perfect, service, amazing, customer, quickly, item
Lift: customer, quality, service, perfect, quickly, amazing, item
Score: customer, quality, amazing, service, perfect, item, quickly

Topic 4 Top Words:

Highest Prob: issue, review, recognition, headphone, device, video, update
FREX: issue, finger, review, people, speaker, headphone, update
Lift: reason, finger, slow, people, speaker, reader, issue
Score: slow, issue, review, speaker, headphone, video, finger

Topic 5 Top Words:

Highest Prob: please, brand, condition, wife, decent, earlier, extremely
FREX: please, condition, wife, earlier, extremely, scratch, brand
Lift: extremely, scratch, earlier, please, standard, wife, condition
Score: standard, please, brand, condition, wife, scratch, earlier

Topic 6 Top Words:

Highest Prob: time, day, seller, charge, hour, week, charger
FREX: day, week, couple, charge, month, hour, charger
Lift: charger, half, month, return, couple, day, week
Score: couple, day, seller, charge, week, hour, charger

Topic 7 Top Words:

Highest Prob: model, fine, storage, upgrade, highly, user, expensive
FREX: model, highly, lite, storage, expensive, upgrade, system
Lift: cheaper, lite, model, system, highly, expensive, pack
Score: system, model, highly, lite, fine, storage, upgrade

Topic 8 Top Words:

Highest Prob: delivery, absolutely, impress, change, perfectly, language, upgrade
FREX: delivery, absolutely, language, instructions, impress, perfectly, package
Lift: instructions, absolutely, delivery, language, package, perfectly, italian
Score: original, delivery, language, absolutely, change, impress, perfectly

Topic 9 Top Words:

Highest Prob: size, light, picture, hand, display, easily, button
FREX: hand, button, edge, size, protector, light, touch

Lift: edge, mode, power, button, protector, hand, home
Score: home, button, size, hand, light, edge, touch

Topic 10 Top Words:

Highest Prob: battery, life, design, speed, run, wireless, comment
FREX: battery, life, speed, run, design, wireless, comment
Lift: battery, life, speed, run, wireless, design, comment
Score: life, battery, design, speed, wireless, run, comment

Topic 11 Top Words:

Highest Prob: price, fast, time, quick, fault, super, market
FREX: fast, price, complaint, quick, market, fault, space
Lift: complaint, fast, price, lovely, space, market, colour
Score: complaint, price, fast, quick, time, super, fault

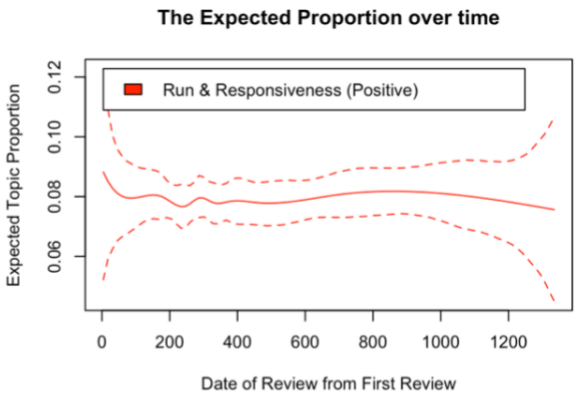
Topic 12 Top Words:

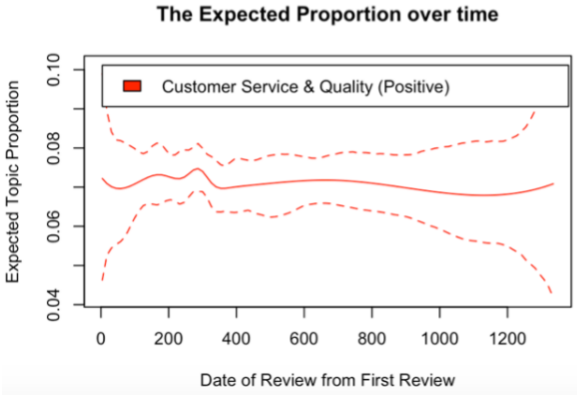
Highest Prob: camera, easy, excellent, feature, fantastic, brilliant, worth
FREX: easy, excellent, fantastic, camera, worth, brilliant, feature
Lift: worth, easy, fantastic, excellent, memory, brilliant, camera
Score: worth, easy, excellent, camera, feature, fantastic, brilliant

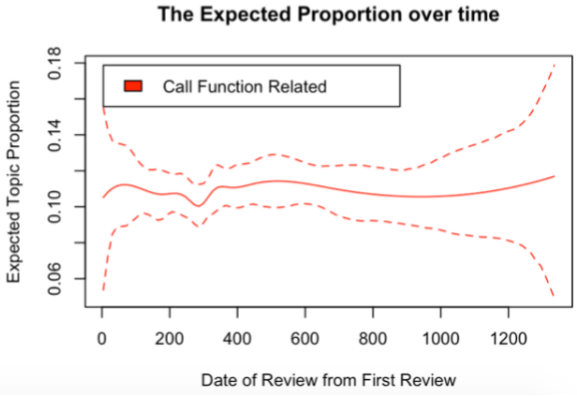
Topic 13 Top Words:

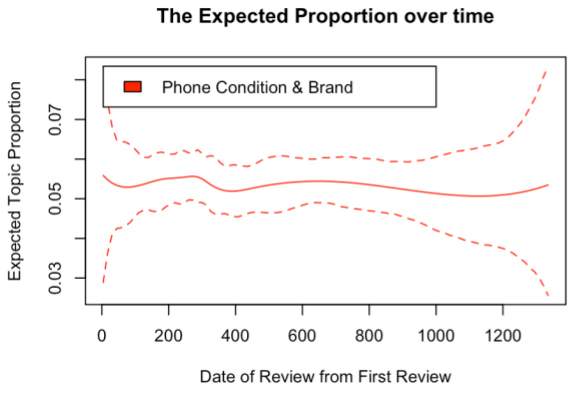
Highest Prob: screen, previous, cheap, performance, lot, game, miss
FREX: cheap, performance, previous, miss, screen, lot, game
Lift: miss, performance, cheap, lot, previous, game, screen
Score: miss, screen, performance, previous, cheap, game, lot

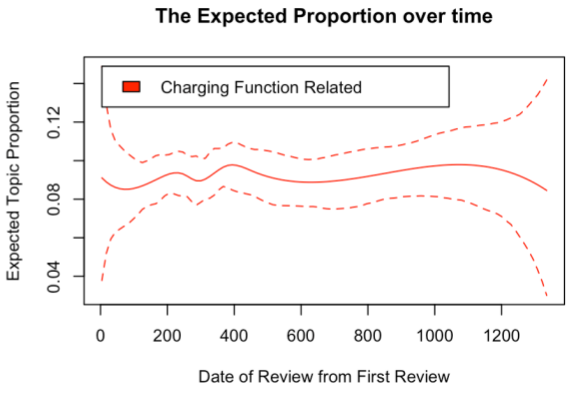
Appendix B - Estimated Proportion of Topic Change over Time

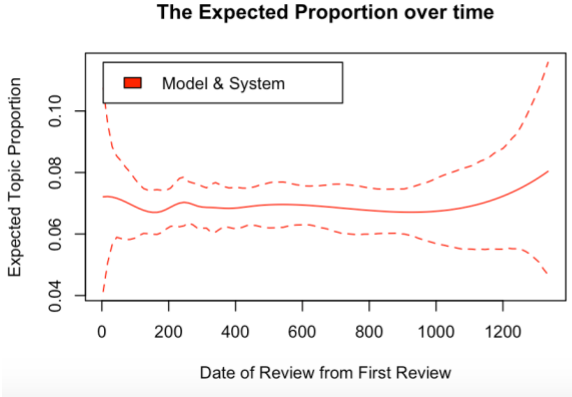


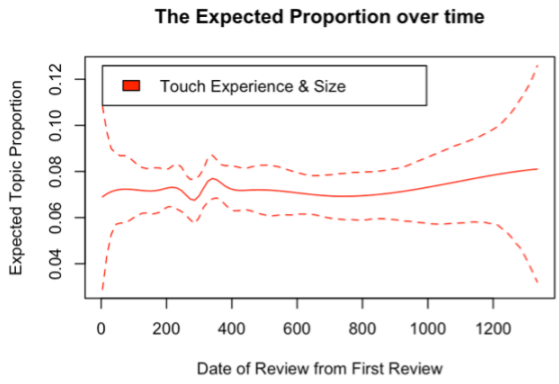


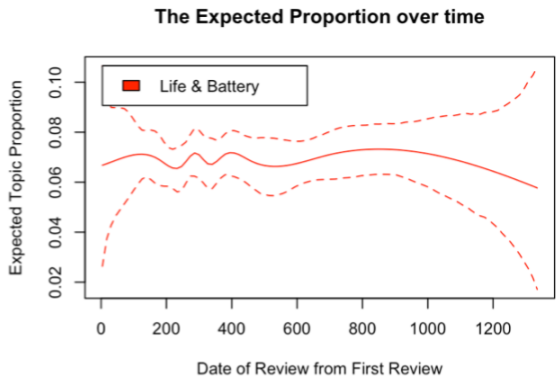


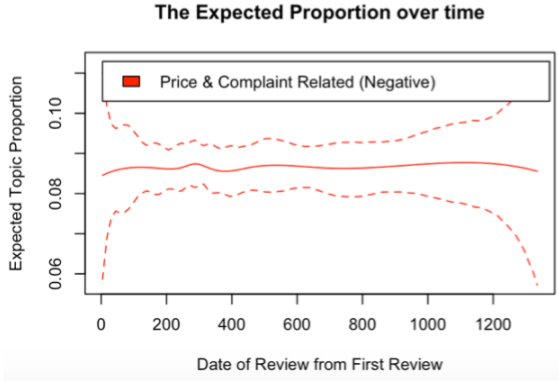


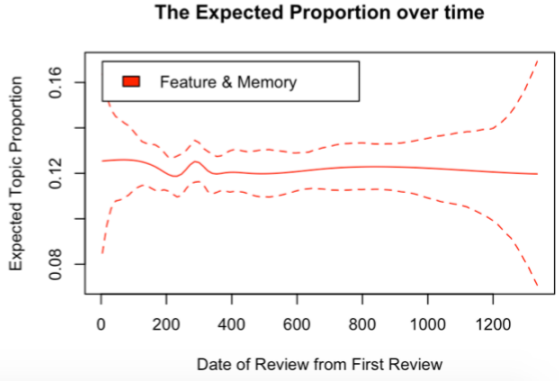


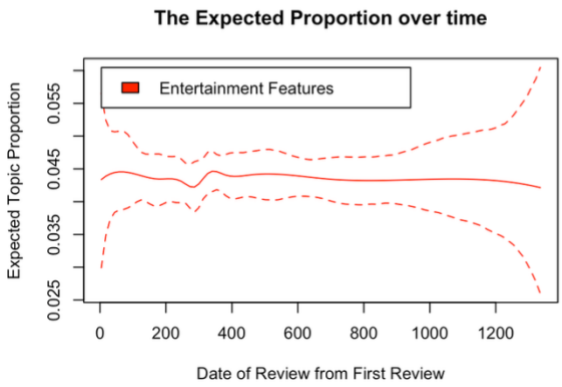



The Expected Proportion over time
Touch Experience & Size
Expected Topic Proportion
0.04
0.06
0.08
0.10
0.12
0
200
400
600
800
1000
1200
Date of Review from First Review


The Expected Proportion over time
Life & Battery
Expected Topic Proportion
0.02
0.04
0.06
0.08
0.10
0
200
400
600
800
1000
1200
Date of Review from First Review


The Expected Proportion over time
Price & Complaint Related (Negative)
Expected Topic Proportion
0.06
0.08
0.10
0
200
400
600
800
1000
1200
Date of Review from First Review


The Expected Proportion over time
Feature & Memory
Expected Topic Proportion
0.08
0.12
0.16
0
200
400
600
800
1000
1200
Date of Review from First Review


The Expected Proportion over time
Entertainment Features
Expected Topic Proportion
0.025
0.035
0.045
0.055
0
200
400
600
800
1000
1200
Date of Review from First Review

Appendix C - R Source Code

Step 1: Extract Online Reviews

```
# Use python to extract product reviews by applying the method of Amazon Multi Language Reviews Scraper

# Available at https://github.com/philipperemy/amazon-reviews-scraper
```

Step 2: Format Conversion

```
# Transfer json (download via GitHub python link) to rjson format

# Then transfer them into data frame for further process

library("rjson")
library("dplyr")

# newly updated files
json_sumsungs8 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/samsung/s8_20170331.json"))
json_sumsungs9 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/samsung/s9_20180225.json"))
json_sumsunga10 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/samsung/a10_B07S4TBRNJ.json"))
json_sumsunga40 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/samsung/a40_20190503.json"))
json_sumsunga70 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/samsung/a70_20190503.json"))

json_huawei_psmart <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/huawei/P_Smart_Pro 20200420.json"))
json_huawei_y5p <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/huawei/Y5p_20200103.json"))
json_huawei_p40 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/huawei/p40_20200103.json"))
json_huawei_p30pro <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/huawei/p30_20190326.json"))
json_huawei_p30lite <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/huawei/p30lite_20190425.json"))

json_iphone7 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/iphone_new/B01LW1VU9E_iphone7.json"))
json_iphone8 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/iphone_new/B075V3RKP3_iphone8.json"))
json_iphone11 <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/iphone_new/B07XRPFJ14_iphone11.json"))
json_iphonex <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/iphone_new/B076GV5GXF_iphonex.json"))
json_iphonexr <- fromJSON(file = paste0(getwd(),"/amazon-reviews-scraper-master/iphone_new/B07JZVDRKD_iphonexr.json"))
```

```
# Take one of the data sets Samsung s8 as an example
all_length_list <- length(json_sumsungs8 )

df <- data.frame(author_url=character(),
                 body=character(),
                 product_id=character(),
                 rating=character(),
                 review_date=character(),
                 review_url=character(),
                 title=character(),
                 stringsAsFactors = FALSE)

for (i in 1:all_length_list){

  record = json_sumsungs8 [[i]]
  if (is.null(record$author_url) == FALSE &&
      is.null(record$body) == FALSE &&
      is.null(record$product_id) == FALSE &&
      is.null(record$rating) == FALSE &&
      is.null(record$review_date) == FALSE){


    df[nrow(df) + 1,] = c(record$author_url,
                          record$body,
                          record$product_id,
                          record$rating,
                          record$review_date,
                          record$review_url,
                          record$title)
  }
}

# Get information about review length (from the original review set)
sample_length <- sapply(strsplit(df$body, " "), length)
avg_length <- mean(sample_length)

# Get information about review rating (from the original review set)
df$rating <- as.numeric(df$rating)
avg_rating <- mean(df$rating)

save(df,file=paste0(getwd(),"/json_sumsunga8.rda"))
```

Step 3: Pre-processing

```
# Clean & preprocessing
library(textclean)
library(qdap)
library(tidytext)
library(dplyr)
library(stringr)
library(textcat)
```

```
# For content
df$body <- iconv(df$body,"latin1", "ASCII", sub="")

## Cleaning up the language
df$language <- textcat(df$body)
df <- df %>% filter(language == "english")

# Remove numbers
df$body <- gsub('[[:digit:]]+','',df$body)

# Remove pounds sign
df$body <-gsub('£','',df$body)

# Remove punctuation
df$body <- gsub('[[:punct:]]+',' ', df$body)

# Remove abbreviations
df$body <- replace_abbreviation(df$body)
df$body <- replace_contraction(df$body)

# Remove people's names
df$body <- replace_names(df$body)

# Replace ordinal
df$body <- replace_ordinal(df$body)

# Replace internet slang
df$body <- replace_internet_slang(df$body)

# Make word in lowercase
df$body <- tolower(df$body)

# Remove empty line and other space
df$body <- gsub('\n',' ', df$body)
df$body <- gsub('[^ -~]', ' ', df$body)

df_backup <- df

# Stop word will be also removed before token (see later)

# Remove stop word
data("stop_words")
special_words <- c("phone","samsung","huawei","iphone","apple")
cus_stop_words <- rbind(stop_words,tibble(word=special_words,lexicon="cust
om"))


# Extract date via regular expression
df <- df %>% mutate(date = str_extract(review_date,
                   "[0-9]{1,2}\\s(January|February|March|April|May|June|
                   July|August|September|October|November|December)\\s[0-9
]{4}"))

df <-df %>% mutate(betterdate = as.Date(date,format("%d %B %Y")))
```

```
# Add reviewer id from (review) author_url
ept_vector <- c()

for (i in 1:length(df$author_url)){


  ept_vector[i] = str_split(str_split(df$author_url,"/")[i][[1]][4],"\\.")
[[1]][3]


}

df <- cbind(df,ept_vector)

# Change column names
colnames(df)[11] <- c("reviewer_id")
colnames(df)[10] <- c("review_dates")
colnames(df)[2] <- c("review_content")

# Prepare for token
df_for_token <- df %>% select(product_id,reviewer_id,review_dates,review_c
ontent,rating)

# Length of review (number of words in review after pre-processing)
afterpre_length <- sapply(strsplit(df_for_token$review_content, " "), leng
th)
afterpre_avg_length <- mean(afterpre_length)


save(df_for_token,file=paste0(getwd(),"/df_samsungs8_token.rda"))
save(df,file=paste0(getwd(),"/df_samsungs8_clean.rda"))
```

Step 4: Tokenization

```
library(udpipe)
library(tidytext)
library(ggplot2)
library(qdap)
library(wordcloud)

load(file=paste0(getwd(),"/df_isamsungs8_token.rda"))

# Check if same author writes different reviews
# Which means one reviewer write more than one reviews
duplicated(df_for_token$reviewer_id)
df_for_token$reviewer_id[duplicated(df_for_token$reviewer_id)]

# Sort records in order of comment date
# Create a review id
df_for_token <- df_for_token %>% arrange(review_dates)
df_for_token <- df_for_token %>% mutate(review_id = row_number())
```

```
# Extract year and month information
df_for_token <- df_for_token %>% mutate(year = str_extract(df_for_token$re
view_dates,"[0-9]{4}"))
df_for_token <- df_for_token %>% mutate(month = format(df_for_token$review
_dates, "%m"))
# str_extract(df_for_token$review_dates,"(?<=-)[0-9]{2}(?=-)") -- other re
gex for month


# Token for words
tokens_all <-df_for_token%>%
  unnest_tokens(word,review_content) %>%
  count(word,review_id) %>%
  anti_join(cus_stop_words)

tokens_all$length_chars <- nchar(tokens_all$word)

# Plot for nchar distribution
hist(tokens_all$length_chars,breaks = 30,main="Samsung s8 review character
 length plot")

# Filter the word length should >= 4
tokens_all <- tokens_all %>%
  filter(length_chars >=4 & length_chars <=15)

# Udpipe
language <- udpipe_download_model(language="english",overwrite = F)
ud_model <- udpipe_load_model("english-ewt-ud-2.4-190531.udpipe")

# Select Noun, Adj, Adv
udpipe_annotate(tokens_all$word,
                doc_id = tokens_all$review_id,
                object = ud_model) %>%
  as.data.frame() %>%
  filter(upos %in% c("NOUN","ADV","ADJ")) %>%
  select(doc_id,lemma) %>%
  group_by(doc_id) %>%
  summarise(annotated_review = paste(lemma, collapse = " "))  -> annotated
_iphonexr_review_word

get_word <- strsplit(annotated_iphonexr_review_word$annotated_review," ")
get_word <- unlist(get_word)
vec<- unique(get_word)

tokens_all <- tokens_all %>%
  filter(word %in% vec)


# Check tf-idf
# Get histogram
review_tokens_all <- tokens_all %>%
    count(word,review_id) %>%
    bind_tf_idf(word,review_id,n)
```

```
review_tokens_all <- review_tokens_all %>%
                              filter(idf>0.02)

hist(review_tokens_all$tf_idf,breaks = 200,main="SamsungS8 TF-IDF plot")

# Find the most important word & plot
tokens_top_10 <- review_tokens_all %>%
                     group_by(word) %>%
                     summarise(total=n()) %>%
                     arrange(desc(total)) %>%
                     top_n(10)

ggplot(tokens_top_10, aes(x=reorder(word,total), y=total)) +
  geom_bar(stat="identity") +
  theme_bw() +
  coord_flip() +
  theme(axis.title.y = element_blank()) +
  labs(y="count", title="Top10 most important words from Samsungs8 reviews
")

# Wordcloud for most important words
wordcloud(tokens_top_10$word, tokens_top_10$total,  scale = c(1.5,.65),
          random.order = F,
          random.color = F,
          rot.per=.15)
          # colors="green4")
```

Step 5: Bass model Application

```
library(fpp2)

# Still take Samsung s8 as an example
# Rank reviews by review written date
df_bass_samsungs8 <- df_for_token %>%
             arrange(review_dates)

# Assumption: assume X consumers buy the product and an equidistant sample
 of them posts a review
# Regard each review as a sale transaction
df_bass_samsungs8 <- df_bass_samsungs8  %>% mutate(amount=1)

# Time series preparation
amount_per_day <- df_bass_samsungs8  %>% count(review_dates)


# Plot review (equidistant sale)data in quarter
df_month_year <- amount_per_day %>%
                     mutate(month = format(review_dates, "%m"), year = fo
rmat(review_dates, "%Y")) %>%
                     group_by(year, month) %>%
                     summarise(total = sum(n))
```

```r
par(mfrow = c(2, 2))

# Assumption: check the first review as the product release day (input as yyyy/mm)
sales=ts(df_month_year$total,start=c(2017,4),freq=12)

# Plot sales per month
plot(sales,type = "l", lty=2, col="red",ylab="",xlab="")
points(sales, pch=20, col="blue")
title("Samsung s8 monthly sales (equidistant reviews)")


# Plot cumulative sales
Y=cumsum(sales)
Y=ts(Y,start = c(2017,4),freq=12)

plot(Y,type = "l", lty=2, col="red",ylab="",xlab="")
points(Y, pch=20, col="blue")
title("Cumulative Samsung s8 sales (equidistant reviews)")


# Assumption: when estimate the p,q coefficient by using the Bass model, start from the fourth review
# Reason: There are three parameters in the parameter estimation model of the Bass model

# “From” should be the date of the first review
# “To" should be the date of the last review
date_set <- seq(from=as.Date("2017-04-24"), to=as.Date("2020-06-14"), by = 1)
date_set <- as.data.frame(date_set)

new_df <- data.frame()
new_df <- left_join(date_set,df_bass_samsungs8, by=c("date_set" = "review_dates")) %>%
                    select(date_set,amount,review_id)
new_df[is.na(new_df)] <- 0

per_df <- aggregate(new_df$amount, by=list(new_df$date_set), sum)

colnames(per_df) <- c("review_dates", "day_sale")

# Mutate cumulative sales data
per_df <- per_df %>% mutate(cum_sale = cumsum(day_sale))

# Add s(day) into data frame as a new column
per_df <- per_df %>% mutate(day_count = row_number())
```

```
# The Bass model
# Calculate p/q value
coef_pq <- data.frame(p=rep(0,each=3),q=rep(0,each=3))
Ycum = per_df$cum_sale


for (i in 4:length(Ycum)){
  cum_sales <- Ycum[1:i]

  # Fit bass regression and compute m,p,q
  # Create the lag value for y
  sales = per_df$day_sale[1:i]
  Ylag=c(0,cum_sales[1:(length(cum_sales)-1)])
  Ysq = Ylag**2

  # Run regression
  out = lm(sales~Ylag + Ysq)
  a=out$coef[1]
  b=out$coef[2]
  c=out$coef[3]

  mplus<- (-b+sqrt(b**2-4*a*c))/(2*c)
  mminus <- (-b-sqrt(b**2-4*a*c))/(2*c)
  m <- max(mplus,mminus)

  p = a/m
  q = b + a/m

  coef_pq <- rbind(coef_pq,c(p,q))

}

# Bind coefficient of pand q with per_df together
per_df_back <- per_df
per_df <- per_df %>% mutate(p=coef_pq$p, q=coef_pq$q)

# Successful products (p <q) and the failed products (p> q)
pq_success <- per_df %>% filter(p<q)

pq_fail <- per_df %>% filter(p>q)
```

Step 5: STM

```
library(tm)
library(stm)
library(stargazer)

# Linked with the Step 4 Tokenization
data_stm <- annotated_iphonexr_review_word
data_stm$doc_id <- as.integer(data_stm$doc_id)
```

```
data_stm_1 <- data_stm %>%
              left_join(df_bass_samsungs8,by=c("doc_id"="review_id")) %>%
              mutate(brand = "samsungs8")

# Linked with the results of the Step 5 Bass model
per_df <- per_df %>% na.omit() %>% filter(p !=0 | q!=0)

data_stm_2 <- data_stm_1 %>%
                  left_join(per_df,by="review_dates") %>% na.omit() %>%
                  mutate(p_q_diff = p-q)

samsungs8<- data_stm_2
save(samsungs8, file = paste0(getwd(),"/samsungs8.rda"))


# Aggregated all dataset together
data_samsung <- bind_rows(s8,s9,a10,a40,a70) %>%
                        mutate(general_brand = "samsung")
data_huawei <- bind_rows(psmart,y5p, p40,p30lite,p30pro) %>%
                        mutate(general_brand = "huawei")
data_iphone <- bind_rows(iphone7,iphone8, iphone11, iphonex, iphonexr) %>%

                        mutate(general_brand = "iphone")


data_brands <- bind_rows(data_samsung,data_huawei, data_iphone)



# Set new document ID
data_brands <- data_brands %>%
                mutate(id = paste0(brand,"_",doc_id)) %>%
                select(-doc_id)

colnames(data_brands_v2)[18] <- c("doc_id")



# Double check for stop words
cus_words <- c("huawei","samsung","iphone","phone","apple")
processed<- textProcessor(data_brands$annotated_review,
                          metadata = data_brands,
                          customstopwords = cus_words,
                          stem = F)


# Set threshold
threshold <- round(1/100 * length(processed$documents),0)

out <- prepDocuments(processed$documents,
                     processed$vocab,
                     processed$meta,
                     lower.thresh = threshold)
```

```
# Store doc, vocab, and meta
docs <- out$documents
vocab <- out$vocab
meta <- out$meta

# Separate dataset for different brands
samsung_data <- meta[meta$general_brand == "samsung",]
huawei_data <- meta[meta$general_brand == "huawei",]
iphone_data <- meta[meta$general_brand == "iphone",]

# Plot for rating score under different brand
avg_rating_samsung <- samsung_data  %>%
  group_by(year,month) %>%
  summarise(mean_rating = mean(rating,na.rm=T)) %>%
  arrange(year,month) %>%
  na.omit()

avg_rating_huawei<- huawei_data  %>%
  group_by(year,month) %>%
  summarise(mean_rating = mean(rating,na.rm=T)) %>%
  arrange(year,month) %>%
  na.omit()

avg_rating_iphone<- iphone_data  %>%
  group_by(year,month) %>%
  summarise(mean_rating = mean(rating,na.rm=T)) %>%
  arrange(year,month) %>%
  na.omit()

# Visualize the rating score results
samsung_plot <- ggplot(avg_rating_samsung,aes(x=paste(year, month), y=mean
_rating,group=1))+
  geom_point()+geom_line()+geom_smooth()+theme_bw()+
  theme(panel.grid=element_blank(), axis.text.x = element_blank(), panel.b
order = element_blank(),
        axis.line = element_line(size=1, colour = "black")) +
  scale_y_continuous(breaks = seq(1, 5, 0.5)) + labs(x = "Year", y = "Aver
age Rating")+ggtitle("Samsung")

huawei_plot <- ggplot(avg_rating_huawei,aes(x=paste(year, month), y=mean_r
ating,group=1))+
  geom_point()+geom_line()+geom_smooth()+theme_bw()+
  theme(panel.grid=element_blank(), axis.text.x = element_blank(), panel.b
order = element_blank(),
        axis.line = element_line(size=1, colour = "black")) +
  scale_y_continuous(breaks = seq(1, 5, 0.2)) + labs(x = "Year", y = "Aver
age Rating")+ggtitle("Huawei")

iphone_plot <- ggplot(avg_rating_iphone,aes(x=paste(year, month), y=mean_r
ating,group=1))+
  geom_point()+geom_line()+geom_smooth()+theme_bw()+
  theme(panel.grid=element_blank(), axis.text.x = element_blank(), panel.b
order = element_blank(),
        axis.line = element_line(size=1, colour = "black")) +
```

```
  scale_y_continuous(breaks = seq(1, 5, 0.2)) + labs(x = "Year", y = "Aver
age Rating")+ggtitle("iphone")



# SearchK to find the optimal number of topics
# First attempt:
k1=seq(from=2,to=40,by=2)
kresult_1 <- searchK(out$documents, out$vocab, K=k1,
                     N = floor(0.1 *length(out$documents)),
                     prevalence = ~factor(rating) + p_q_diff + factor(gene
ral_brand) +s(day_count),
                     cores = 1,data=out$meta)
plot(kresult_1)

# Second attempt (narrow the range, and shorten the step):
k4=seq(from=6,to=26,by=1)
kresult_2 <- searchK(out$documents, out$vocab, K=k4,
                     N = floor(0.1 *length(out$documents)),
                     prevalence = ~factor(rating) + p_q_diff + factor(gene
ral_brand) + s(day_count),
                     cores = 1,data=out$meta)
plot(kresult_2)


# Find optimal number of topics -- 13

# STM select models
topic_13 <-  stm::selectModel(documents = out$documents,
                              vocab = out$vocab,
                              K = 13,
                              prevalence =~factor(rating) + p_q_diff + fac
tor(general_brand) + s(day_count),
                              # max.em.its = 80,
                              data = out$meta,
                              # reportevery=3,
                              # seed = 123,
                              # # gamma.prior = "L1",
                              # sigma.prior = 0.7,
                              init.type = "Spectral",
                              ngroups = 5)

# Check semantic coherence & exclusivity
plotModels(topic_13, pch = c(1:13),legend.position="bottomleft")

# Summarize the optimal model
# Choose model 3 with relatively best performance
summary(topic_13$runout[[3]])
select_stm<- topic_13$runout[[3]]

# Demonstrate the top words in the given model
# Top 7 words associated with each topic
labelTopics(select_stm,n=7)
```

```r
# Check & plot topic correlation
plot(topicCorr(select_stm))

topic_cor <- topicCorr(select_stm,method = "simple")
plot.topicCorr(topic_cor,vlabels = paste0(c(1:13),": ",topic_labels),verte
x.color = "orange",
               # vertex.label.dist=-1,
               layout=igraph::layout.kamada.kawai,
               vertex.label.cex =0.7)

# Use plot.STM to compare the topic proportion
stm_model_summary<- summary(select_stm)
plot.STM(select_stm,type = "summary",xlim=c(0,0.3),labeltype = "frex",n=7)


# Label the topic & get topic proportions
convergence_theta <- as.data.frame(select_stm$theta)
topic_proportions <- colMeans(convergence_theta)
topic_labels <- c(
  "Purchase Experience",
  "Run & Responsiveness (Positive)",
  "Customer Service & Quality (Positive)",
  "Call Function Related",
  "Phone Condition & Brand",
  "Charging Function Related",
  "Model & System",
  "Product Delivery",
  "Touch Experience & Size",
  "Life & Battery",
  "Price & Complaint Related (Negative)",
  "Feature & Memory",
  "Entertainment Features")


colnames(convergence_theta) <-  paste0("topic_",1:13)
for_analysis <- cbind(out$meta,convergence_theta)

# Combine topics with corresponding proportion
un_table_towrite_labels <- data.frame()
for(i in 1:length(stm_model_summary$topicnums)){

  row_here <- tibble(topicnum= stm_model_summary$topicnums[i],
                     topic_label = topic_labels[i],
                     # topic_label = paste(unsupervised_results$frex[i,1:3
],
                     #                     collapse = ", "),
                     proportion = 100*round(topic_proportions[i],4),
                     frex_words = paste(stm_model_summary$frex[i,1:7],
                                        collapse = ", "))
  un_table_towrite_labels <- rbind(row_here,un_table_towrite_labels)
}
```

```r
# Arrange by topic number
un_table_towrite_labels %>% arrange(topicnum)


# Arrange by topic proportion
# With more insights
un_table_towrite_labels %>% arrange(desc(proportion))

# Proportion of topics (one of the topics)
round(100*mean(select_stm$theta[,10]),2)

# Perspective statistics
toplot <- for_analysis %>% group_by(general_brand) %>% summarise(
  mtopic1 = mean(topic_1),
  mtopic2 = mean(topic_2),
  mtopic3 = mean(topic_3),
  mtopic4 = mean(topic_4),
  mtopic5 = mean(topic_5),
  mtopic6 = mean(topic_6),
  mtopic7 = mean(topic_7),
  mtopic8 = mean(topic_8),
  mtopic9 = mean(topic_9),
  mtopic10 = mean(topic_10),
  mtopic11 = mean(topic_11),
  mtopic12 = mean(topic_12),
  mtopic13 = mean(topic_13)
) %>%as.data.frame()

colnames(toplot) <- c("general_brand",topic_labels)
rownames(toplot) <- toplot$general_brand
toplot$general_brand <- NULL

plot(anacor(toplot), pch = 16, arrows = c(F, T),
     main = NULL, xlim = c(-0.025, 0.025), ylim = c(-0.015, 0.015))


# Sub-plot related to perspective statistics
perceptual_map_plot <- toplot %>%
  select('Call Function Related','Touch Experience & Size',
         'Phone Condition & Brand','Product Performance & Game')
perceptual_map_plot2 <- toplot %>%
  select('Price & Complain Related (Negative)', 'Run & Responsiveness (Pos
itive)',
         'Customer Service & Quality (Positive)', 'Product Delivery', 'Mod
el & System')

perceptual_map_plot3 <- toplot %>%
  select('Purchase Experience','Charging Function Related',
         'Life & Battery','Secondary Feature')

plot(anacor(perceptual_map_plot), pch = 16, arrows = c(F, T),
     main = NULL, xlim = c(-0.01, 0.01), ylim = c(-0.001, 0.001))

plot(anacor(perceptual_map_plot2), pch = 16, arrows = c(F, T),
```

```
     main = NULL,xlim = c(-0.025, 0.025), ylim = c(-0.01, 0.01))

plot(anacor(perceptual_map_plot3), pch = 16, arrows = c(F, T),
     main = NULL,xlim = c(-0.03, 0.03), ylim = c(-0.005, 0.005))




# Estimate effect
graphics.off()
effects_prep<- estimateEffect(1:13 ~factor(rating) + p_q_diff + factor(gen
eral_brand) + s(day_count),
                              # prior= 1e-5,
                              stmobj = select_stm,
                              meta = out$meta,
                              uncertainty = "Global")


# p_q difference based on overall brand level
par(mfrow = c(1, 1))
plot(effects_prep, covariate = "p_q_diff", # here p_q_diff means value of
p-q
     topics = c(1:13),
     model = select_stm_2, method = "difference",
     cov.value1 = "-1", cov.value2 = "1",
     ci.level = 0,
     xlab = "p>q  ....... p<q",
     xlim = c(-0.015,0.015),
     main = "Effect of Overall Difference of Innovation (p) and Imitation
(q) Coefficients on Topic Prevalence for Phone",
     custom.labels =topic_labels,
     labeltype = "custom")



# Marginal topic proportion for each of the brand
# "pointestimate" estimates mean topic proportions for each value of the
covariate.

par(mfrow = c(2, 2))

plot(effects_prep, covariate = "general_brand", topics = 1,
     model = select_stm_2, method="pointestimate",
     main = "Topic 1 Purchase Experience", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep_2, covariate = "general_brand", topics = 2,
     model = select_stm_2, method="pointestimate",
     main = "Topic 2 Run & Responsiveness", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep_2, covariate = "general_brand", topics = 3,
     model = select_stm_2, method="pointestimate",
     main = "Topic 3 Customer Service & Quality (Positive)", labeltype = "
```

```
custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep_2, covariate = "general_brand", topics = 4,
     model = select_stm_2, method="pointestimate",
     main = "Topic 4 Call Function Related", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 5,
     model = select_stm_2, method="pointestimate",
     main = "Topic 5 Phone Condition & Brand", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 6,
     model = select_stm_2, method="pointestimate",
     main = "Topic 6 Charging Function Related", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 7,
     model = select_stm_2, method="pointestimate",
     main = "Topic 7 Model & System", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 8,
     model = select_stm_2, method="pointestimate",
     main = "Topic 8 Product Delivery", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 9,
     model = select_stm_2, method="pointestimate",
     main = "Topic 9 Touch Experience & Size", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 10,
     model = select_stm_2, method="pointestimate",
     main = "Topic 10 Life & Battery", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 11,
     model = select_stm_2, method="pointestimate",
     main = "Topic 11 Price & Complaint Related", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 12,
     model = select_stm_2, method="pointestimate",
     main = "Topic 12 Feature & Memory", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))

plot(effects_prep, covariate = "general_brand", topics = 13,
     model = select_stm_2, method="pointestimate",
     main = "Topic 13 Entertainment Features", labeltype = "custom",
     custom.labels = c("Samsung","iphone","Huawei"))
```

```
# How topics change with time
monthseq <- seq(from=min(as.Date(out$meta$review_dates,format="%B-%Y")),
                to=max(as.Date(out$meta$review_dates,format="%B-%Y")),
                by="month")

review_time <- as.Date(out$meta$review_dates,format="%B-%Y")
earliest_date <- min(as.Date(out$meta$review_dates,format="%B-%Y"))
latest_date <- max(as.Date(out$meta$review_dates,format="%B-%Y"))

date_diff <- as.numeric(latest_date) -as.numeric(earliest_date)

par(mfrow=c(2,3),
    oma = c(5,4,0,0) + 0.1,
    mar = c(2,0,2,0) + 0.1)

for(i in 1:13){

  plot(effects_prep, covariate = "day_count",
       topics = c(i),
       model = select_stm, method = "continuous",
       xaxt='n',
       xlab="Date of Review （from first product review)",
       main = "The Expected Proportion over time",
       printlegend = F,
       # linecol = "black",
       # labeltype = "custom")

  # axis(1,at=seq(from=1,
  #               to=date_diff,
  #               by=30),labels=monthseq)
}
```

Step 6: Regression model

```
# Adding a regression
analysis_samsung <- for_analysis %>% filter(general_brand =="samsung")
analysis_huawei <- for_analysis %>% filter(general_brand =="huawei")
analysis_iphone <- for_analysis %>% filter(general_brand =="iphone")

# Samsung
lm_samsung_baseline = lm(p_q_diff ~ rating, data = analysis_samsung)
lm_samsung_1 = lm(p_q_diff ~ rating + topic_3, data = analysis_samsung)
lm_samsung_2 = lm(p_q_diff ~ rating + topic_3 +topic_9, data = analysis_sa
msung)
lm_samsung_3 = lm(p_q_diff ~ rating + topic_3 +topic_9 + topic_6, data = a
nalysis_samsung)
lm_samsung_4 = lm(p_q_diff ~ rating + topic_3 +topic_9 + topic_6 + topic_1
, data = analysis_samsung)


stargazer::stargazer(lm_samsung_baseline, lm_samsung_1, lm_samsung_2, lm_s
```

```
amsung_3, lm_samsung_4,
                     type = "text",
                     title = "Regression Result for Samsung")

# try with relatively large proportion
# lm_samsung = lm(p_q_diff ~ rating + topic_2 + topic_3 + topic_4 + topic_
6 + topic_11 + topic_12, data = analysis_samsung)
# summary(lm_samsung)


# Huawei
lm_huawei_baseline = lm(p_q_diff ~ rating, data = analysis_huawei)
lm_huawei_1 = lm(p_q_diff ~ rating + topic_3, data = analysis_huawei) # to
pic3 important  0.359122
lm_huawei_2 = lm(p_q_diff ~ rating + topic_3 + topic_9, data = analysis_hu
awei)
lm_huawei_3 = lm(p_q_diff ~ rating + topic_3 + topic_9 +topic_6 , data = a
nalysis_huawei)
lm_huawei_4 = lm(p_q_diff ~ rating + topic_3 + topic_9 +topic_6 +topic_1,
data = analysis_huawei)


stargazer::stargazer(lm_huawei_baseline,lm_huawei_1,lm_huawei_2,lm_huawei_
3,lm_huawei_4,
                     type = "text",
                     title = "Regression Result for Huawei")


# Apple
lm_iphone_baseline = lm(p_q_diff ~  rating, data = analysis_iphone)
lm_iphone_1 = lm(p_q_diff ~  rating + topic_3, data = analysis_iphone)
lm_iphone_2 = lm(p_q_diff ~  rating + topic_3 +topic_9, data = analysis_ip
hone)
lm_iphone_3 = lm(p_q_diff ~  rating + topic_3 +topic_9+ topic_6, data = an
alysis_iphone)
lm_iphone_4 = lm(p_q_diff ~  rating + topic_3 +topic_9+ topic_6 + topic_1,
 data = analysis_iphone)


stargazer::stargazer(lm_iphone_baseline, lm_iphone_1,lm_iphone_2, lm_iphon
e_3, lm_iphone_4,
                     type = "text",
                     title = "Regression Result for Apple")
```