DSSR

Discussion Paper No. 115

# Customer Review Analysis Using Word Embedding Model Considering Text Topics

Mirai Igarashi, P. K. Kannan
and Nobuhiko Terui

June, 2020

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

# Customer Review Analysis Using Word Embedding Model Considering Text Topics

Mirai Igarashi * P. K. Kannan † Nobuhiko Terui ‡

May 2020

**Abstract**

Customers often give feedback on their evaluations and experiences with the products and service in the form of customer reviews, and developing the technology of customer review analysis plays an important role in the modern marketing research. Existing studies on marketing have used topic models to capture the review generating behaviors, but this approach ignores the word ordering, that is, it assumes a bag-of-words, and thus cannot adequately consider the context of text even with topic models. In this study, we propose a model combining supervised topic model and word embedding model for estimating the relationship between the product attributes mentioned in the customer review and their satisfactions while capturing the context of review text customers generate. In the empirical analysis, we apply the proposed model to a real customer review data on mascara-related products on a cosmetics e-commerce site, and the results show that our model captures some interpretable topics related to mascara products and estimates their effects on satisfaction scores, for example, the "eyelash" topic mentioned in the review tends to result in high levels of satisfaction, while the "brush" topic is associated with low levels of satisfaction.

**Keywords:** Customer review analysis, Word embedding, Skip-gram model, Topic model, Supervised learning

*Mirai Igarashi is Doctoral Student, Graduate School of Economics and Management, Tohoku University, 27-1, Kawauchi Aoba-ku, Sendai, 980-8576, Japan (E-mail: mirai.igarashi.s7@dc.tohoku.ac.jp). Igarashi acknowledges a grant by JSPS KAKENHI 18J20698.

†P. K. Kannan is Professor, Robert H. Smith School of Business, University of Maryland (E-mail: pkkannan@rhsmith.umd.edu)

‡Nobuhiko Terui is Professor, Graduate School of Economics and Management, Tohoku University (E-mail: terui@tohoku.ac.jp). Terui acknowledges a grant by JSPS KAKENHI (A) 17H01001.

# 1 Introduction

With the development of e-commerce sites, it has become commonplace that consumers purchase products online and give feedback on their evaluations and experiences with the products in the form of customer reviews. Companies utilize this wealth of information to understand consumer preference structures and make use of it in a variety of marketing activities, such as product development, market analysis, and advertising strategy planning. Therefore, developing the technology of customer review analysis plays a vital role in the modern marketing research.

In the literature, modeling the consumer behavior of creating customer reviews has been studied by many researchers to reveal the preference structure behind them. Many of them adopt the topic model, or latent Dirichlet allocation (LDA, Blei et al., 2003), to model the review generating behavior (e.g., Tirunillai and Tellis, 2014). In such studies, it is assumed that when consumers write reviews, they have certain topics, or product attributes, in mind and generate text by selecting words from own vocabulary to describe their evaluations and experiences with respect to these attributes.

From the perspective of text modeling, however, LDA has a major problem of ignoring the important information of context because it applies *bag-of-words*—meaning that word ordering does not matter. The LDA takes into account whether words of interest (e.g., a product name and its evaluation such like good and bad) co-occurs in the same textual unit (e.g., a review), but it cannot understand the semantic relationships between the words. For example, if a review describes the good points of one product attribute and the bad points of another attribute, LDA regards that the good attribute co-occurs with some words used to describe the bad attribute.

On the other hand, word embedding model, or the word2vec (Mikolov et al., 2013), is a machine learning method that has a great success in the field of text modeling. The word2vec defines the probability of word generation given the surrounding words (i.e., the skip-gram model) while it projects words into a feature space. Therefore, the word2vec

can understand the word context in terms of considering the words in the window because the word2vec regards that words related to a product attribute co-occur only with their surrounding words which can be considered to qualify the words, not with words related to another attribute which are at a distance in the same document. This approach can be applied in a variety of domains from sentiment classification (Zhang et al., 2015) to item recommendations (Caselles-Dupré et al., 2018).

However, because we aims to understand preference structures behind review generating behavior, there are not many advantages of using word2vec as it is. The feature vectors of words resulting from embedding learning are usually very high dimensional, and it is impossible to interpret each dimension, as in factor analysis or principal component analysis. Therefore, even if word2vec is applied to customer review analysis, we may not know what consumers express about which attributes in their reviews.

In this study, we propose a model for review generating behaviors based on word2vec and LDA by leaning vectors with respect to not only words but also topics projected into the same feature space. The purpose of this study is to clarify the effects of product attributes mentioned in customer reviews on the customers' satisfactions while considering the contexts by combining the word embedding model and topic model.

In the following sections, first we discuss related works in the relevant body of literature. Next, we describe the model structure that combines word and topic embedding systems and regression model. In the empirical study, we apply the proposed model to a real dataset on e-commerce sites about cosmetics. Finally, we provide concluding remarks and directions for future research.

## 2 Literature Review

In the literature, researchers proposed some approaches using interviews and questionnaires to clarify the consumer preference structure (Fischer et al., 1999; Hoeffler, 2003). However,

because these approaches are costly to implement and the obtained data is limited, it was required to use new data sources, such as information on the internet (Netzer et al., 2008).

As alternatives to the approaches using interviews and questionnaires, the use of customer reviews has been proposed. Decker and Trusov (2010) estimated the impact of product attributes that were positively or negatively mentioned in the customer reviews on customer satisfactions using a latent class Poisson regression. Archak et al. (2011) proposed a demand estimation model that captures the effects of product attributes on sales considering the heterogeneity of each word that qualifies the attributes. Companies and marketers can use the relationship between product attributes and customer satisfactions or sales for a variety of marketing activities, such as to analyze the market structure (Lee and Bradlow, 2011; Moon and Kamakura, 2017) and to improve the product search algorithm (Ghose et al., 2012).

In these studies, they extract the product attributes mentioned in the customer reviews, and then propose models to captures their relationships with objective variables such as review ratings and sales. In other words, the product attributes are regarded as just observed variables. In contrast, some studies simultaneously model the generation of customer reviews and the formation of customer satisfaction, and they often apply topic models, in which the product attributes are treated as the latent variables (e.g., Qi et al., 2016; Puranam et al., 2017; Bi et al., 2019). Topic modeling approach assumes that consumers have topics (product attributes) in their mind and their evaluations and experiences about the topics are embodied in the text and ratings of customer reviews.

However, since topic model or LDA regards the observed text information as the sets of words ignoring their order, i.e., bag-of-words, the main disadvantage of this approach is that it cannot consider the context of text (Berger et al., 2020). For example, if there are two sentences, "I prefer the macbook instead of the surface" and "I prefer the surface instead of the macbook," the LDA model considers them to be the same data, even though these reviews state completely conflicting preferences. This characteristics can be a more serious

drawback in the customer review analysis.

In the marketing literature, some studies tackle to relax the bag-of-words assumption, for example, Büschken and Allenby (2016) propose the extended LDA model that assign topics not to words independently but to sentences taking into account the within-sentence dependencies. Liu and Toubia (2018) also extend the LDA model by considering hierarchically the contexts of related documents (e.g., search queries and results), and hence their proposed model assigns topics which are related to topics in the correspondence documents rather than independently assignment.

On the other hand, this study focuses on the skip-gram model using word embedding representation (Mikolov et al., 2013) to relax the bag-of-words assumption for the more appropriate text modeling. For the generating probability of a certain word, the skip-gram considers the conditional probability given the surrounding words of the focal word. For example, if considering a problem of predicting appropriate word in a sentence, "I ... a student." ("..." is a masked word), the skip-gram defines the conditional probability, $p(... \mid \mathrm{I}, \mathrm{a}, \mathrm{student})$. Therefore, it is expected that $p(\mathrm{am} \mid \mathrm{I}, \mathrm{a}, \mathrm{student})$ is larger than $p(\mathrm{are} \mid \mathrm{I}, \mathrm{a}, \mathrm{student})$ if the skip-gram learns good embedding representations. While Büschken and Allenby (2016)'s model takes a sentence as a unit to assign the same topic to the words in the sentence, the skip-gram takes a moving window as a unit to assign topics considering the co-occurrence relationships of the words in the window.

Moreover, we focus on the word embedding approach which the word2vec (Mikolov et al., 2013), on which the proposed model is based, takes with the skip-gram model. Word embedding approach assumes an embedding vector in feature space for each word, and it can represent more flexibly the semantically meanings of words compared to other naive vector representation such as one-hot vector. The formulation of the skip-gram using word embedding approach is known as word2vec, and as described above,word2vec also considers the conditional probability of the focal word given the surrounding word. Let $\overrightarrow{w}_i$ be the word $i$'s embedding representation, word2vec defines the conditional probability as the inner product

of the word vectors, $p(\overrightarrow{w}_i \mid \overrightarrow{w}_j) = \exp(\overrightarrow{w}_i^\top \overrightarrow{w}_j) / \sum_v \exp(\overrightarrow{w}_i^\top \overrightarrow{w}_v)$. In recent years, with the development of deep learning and related technologies, there are a number of research aimed at acquiring embedding representations of not only words but also documents (Le and Mikolov, 2014) and products (Barkan and Koenigstein, 2016).

However, it is difficult to apply the word2vec approach to customer review analysis as it is because we cannot interpret each dimension of word embedding vector and linguistically understand the relationships among words. While word2vec approach may be effective in predicting text, it does not help us to understand the effects of the product attributes mentioned in customer reviews on the customers' satisfactions, which is the purpose of this study. In this study, we propose a model combining word2vec with supervised topic model by assuming embedding vectors for words and topics in the same feature space. Since words which have similar vector representations to the topic vector can be regarded to be representative of that topic, we can interpret the extracted topics. Therefore, we achieve the same objectives as the LDA model, obtaining summaries of the documents in the topic dimensions, while relaxing the bag-of-words assumption.

The combination of the topic model and word2vec itself has been proposed in Moody (2016)'s LDA2vec, however, this study extends his model from the following two perspectives. First perspective is that the proposed model combines with the supervised topic model for explaining the effects of product attributes on the customer satisfactions, rather than unsupervised learning such as LDA. We can extract topics or product attributes in the reviews considering not only the text structure but also the relationships between the topics and the customer satisfaction through the supervised learning process. Second perspective is that we assume explicitly the topic assignment into the words as the augmented variable, which is more oriented to the formulation of the original LDA model. As explained in the next section, we apply different estimation procedure—the stochastic optimization and the Bayesian estimation— for estimation the embedding vectors and the regression parameters, respectively, to reduce the computational cost, and this data augmentation simplifies the

derivation of the conditional distributions and allows for the separation of the estimation methods.

# 3 Model

In this section, we describe the formulation of the proposed model in the order of the word embedding part and then the regression model part for the preference measurement. First, we take two embedding vectors, word vector and topic vector, with a fixed embedding dimension, $M$, and let $\overrightarrow{w}_i = \{w_{i1}, \ldots, w_{iM}\}$ and $\overrightarrow{t}_k = \{t_{k1}, \ldots, t_{kM}\}$ be these vectors, respectively, where $w_{im}, t_{km} \in \mathbb{R}\ \forall i, k, m$. Words and topics are projected into the same feature space. In addition, it is assumed that a topic is assigned to the word $i$ according to the topic proportion vector specific to the document to which the word $i$ belongs, $z_i \sim Categorical(\theta_{d_i})$, where $z_i$ represents the topic assignment for word $i$ and $\theta_{d_i}$ is the topic proportion vector of the document to which the word $i$ belongs.

The proposed model represents the context of the word $i$ using a context vector represented by the sum of the word vector and the topic vector corresponding the topic assignment to the word $i, \overrightarrow{c}_i = \overrightarrow{w}_i + \overrightarrow{t}_k$, if $z_i = k$. For example, the context of the word "small" in a sentence "The body of this mobile battery is small and convenient to carry out." is expressed in a positive sense with the linguistic meaning of small if it is in a document with respect to the topic of the mobile battery. In contrast, if the word "small" had been used in a sentence about a mobile display, its contextual meaning would have a negative sense, at least in modern times. The approach of constructing a context vector by the sum of the word vector and the topic vector has also been used in Moody (2016)'s LDA2vec. His model constructs a document vector by the inner product of the topic proportion and the topic vectors, and formulates the context vector as the sum of the word vector and the document vector. In contrast, this study is more oriented to the formulation of the LDA model in that the context is directly formulated as the sum of the word meanings and its topic by assuming

the latent variables of topic assignment.

The proposed model considers its surrounding words to define the probability of generating the focal word. We define $S_i$ as a multiset that contains the surrounding words of the word $i$ in the dataset. Let $I_P = \{(i,j) \mid j \in S_i\}$ and $I_N = \{(i,j) \mid j \notin S_i\}$ be the positive and negative multisets, respectively, and $I_D = I_P \cup I_N$ be the multiset of total vocabulary. Then, we define $G = \{g_{ij} \mid (i,j) \in I_G\}$, where $g_{ij}$ is a random variable whose value is taken to be 1 if $(i,j) \in I_P$ or $-1$ if $(i,j) \in I_N$. In the proposed model, this random variable indicating whether the word $i$ and the word $j$ are in the same window is assumed to follow the binomial logit model, which the probability of the variable is defined as the standard sigmoid (or logistic) function of the inner product of the context vector and the surrounding word vector, $p(g_{ij} \mid \overrightarrow{w}_i, \overrightarrow{w}_j, \{\overrightarrow{t}_k\}, z_i) = \sigma(g_{ij} \cdot \overrightarrow{c}_i^\top \overrightarrow{w}_j)$, where $\sigma(x) = 1/1 + \exp(-x)$. Because the dependent variable of the logit model ($\overrightarrow{c}_i$) is also latent variable, this formulation can be seen as the factor model by regarding $\overrightarrow{c}_i$ as factor scores of the word $i$ and $\overrightarrow{w}_j$ as factor loadings of the word $j$. However, this factor model has a constraint that factor score $\overrightarrow{c}_i$ can be decomposed into the word vector and the corresponding topic vector, $\overrightarrow{c}_i = \overrightarrow{w}_i + \overrightarrow{t}_k$.

Therefore, the likelihood of the word embedding part is provided as follows:

$$p(G, \{\overrightarrow{w}_i\}, \{\overrightarrow{t}_k\}, Z \mid \theta) = \prod_{i=1}^{V} \left\{ p(z_i \mid \theta_{d_i}) \prod_{j=1, i \neq j}^{V} \sigma(g_{ij} \cdot \overrightarrow{c}_i^\top \overrightarrow{w}_j) \right\}, \tag{1}$$

where $V$ is the number of total vocabulary in the corpus. In the proposed model, the embedding vectors are trained to maximize the probability of the inner product of the context vector and the surrounding word vector, so that word vectors which their corresponding words often belong to the same window in a dataset will have similar values. Therefore, we can obtain vector representations taking into account the co-occurrence of words within window, that is, the context, while LDA uniformly considers the co-occurrence of words in a document.

However, the computation cost of the likelihood (1) is very high because the number of

total vocabulary is usually over thousands and the total cost of the likelihood is its square. In this study, according to the approach proposed by Mikolov et al. (2013), we also apply negative sampling technique to approximate the likelihood (1) with the following computable formulation:

$$p(G, \{\overrightarrow{w}_i\}, \{\overrightarrow{t}_k\}, Z \mid \theta) \approx \prod_{i=1}^{V} \left\{ p(z_i \mid \theta_{d_i}) \prod_{j \in S_i} \sigma(\overrightarrow{c}_i^{\top} \overrightarrow{w}_j) \times \prod_{n \sim P_n(w)}^{N} \sigma(-\overrightarrow{c}_i^{\top} \overrightarrow{w}_n) \right\}, \quad (2)$$

where $P_n(w)$ is the noise distribution as a free parameter, and we choose the unigram distribution raised to the 3/4 rd power according to the Mikolov et al. (2013)'s suggestion. The number of negative samples, $N$, is suggested to be $5-20$ for a small dataset and $2-5$ for a large dataset. In machine learning research, since they use a huge dataset containing millions and sometimes billions words for training, our dataset is small in comparison to them and we determine 15 words for the number of negative samples.

Next, we construct the prediction model for the customer review ratings using the ordered probit model, which is the similar formulation to Büschken and Allenby (2016). Recalling that the topic assignments is assumed to follow the categorical distribution of the topic proportion, $z_i \sim Categorical(\theta_{d_i})$, the topic proportion $\theta_d$ represents the summary of the product attributes mentioned in the review $d$ in the dimensions of topics and works as dependent variables for explaining the customer satisfaction in the proposed model. Let $y_d$ be the satisfaction score of the review $d$, which follows the ordered probit model:

$$\begin{aligned} y_d &= r \quad \text{if } \tau_{r-1} \leq y_d^* < \tau_r \\ y_d^* &= \sum_{k=1}^{K} \beta_k \theta_{dk} + \epsilon_d, \qquad \epsilon_d \sim N(0,1), \end{aligned} \quad (3)$$

where the thresholds $\{\tau_r\}$ work for realizing discrete satisfaction scores $(y_d)$ through the latent continuous variable $(y_d^*)$, and the both sides of thresholds, $\tau_0$ and $\tau_R$, are set to $-\infty$ and $\infty$, respectively. Since the proposed model contains no intercept term and the fixed

variance parameter in the regression, the model can be identified for all remaining thresholds parameters.

Therefore, the likelihood of the regression part for the customer satisfaction is provided as follows:

$$p(Y, \beta, \tau \mid \theta) = \left\{ \prod_{d=1}^{D} p(y_d \mid y_d^*, \tau) p(y_d^* \mid \theta_d, \beta) \right\} \times p(\beta), \tag{4}$$

where $p(\beta)$ is the prior distribution for the regression coefficients and the definition is explained in the appendix. Under the assumption of the conditional independence of likelihood (1) and (4) when the topic distributions are given, the full joint likelihood of the proposed model is obtained by the product of equations (1) and (4) multiplied by the prior density for the topic distribution, which is assumed to be the Dirichlet prior $p(\theta_d \mid \theta_0) \sim Dirichlet(\theta_0)$ in this study.

In estimation procedure of the proposed model, we take a hybrid approach combining the Markov Chain Monte Carlo (MCMC) sampling method and the gradient-based stochastic optimization using Adam (Kingma and Ba, 2015). The proposed estimation procedure produces point estimates for two embedding vectors through the stochastic optimization, and then apply the Metropolis-Hastings sampling for the topic distributions and the Gibbs sampling for the remaining parameters given the embedding vector estimates every for updating via the optimization. The estimation procedure and the settings of analysis in the empirical study are explained in more detail in the appendix.

# 4 Empirical Results

## 4.1 Data

In the empirical study, we use product reviews on the website `www.sephora.com` which is an e-commerce site primarily dedicated to cosmetics, and they were collected by the authors in

January 2020. This dataset consists of 8,551 customer reviews on 52 mascara products of 25 brands generated from January 1, 2019 to December 31, 2019. Prior to data analysis, the reviews were preprocessed including the removal of symbols, substituting with lower-case letters, and the removal of reviews consisting of less than 10 words. As a result, the number of words used in the final dataset is 376,033, and the number of unique words is 7,853. The number of words in the review is from 10 words to 287 words, and the histogram distribution is right-skewed. The average of the number of words in the review is 44.0 words per review, the median is 37 words, and the standard deviation is 29.5.

The review includes not only the text but also the satisfaction (rating) scores for the products observed on a five-point scale. The frequency of the rating scores is 917, 736, 1,009, 1,751, and 4,143 in the order of 1 to 5. This positively skewed (J-shaped) characteristics is well-known in many previous studies (e.g., Xiao et al., 2016).

## 4.2 Results

First, we determine the number of topics by comparing models with varying numbers through model selection using the criterion, where in this study, we use the log marginal density (LMD) calculated by the method of Newton and Raftery (1994). Table 1 reports the LMDs for the rating scores in the range of the number of topics from 5 to 10, and the model with 10 topics is the largest. In the following analysis, we use the proposed model with 10 topics.

Next, we interpret each dimension of the topics by interpreting the coherency of word meanings from the words associated with the topic. In the proposed model, since we estimate the embedding vectors of words and topics in the same feature space, we can consider the word whose word vector is the closest to the topic vector as the most related word to that topic. Table 2 displays the top 20 words corresponding to the closest word vectors to each topic vector. The words for each topic provide some coherent descriptions of the mascara products, for example, topic 1 is a description of flaking such that discusses how hard the mascara is to peel off, as evidenced by the use of words such as “flakes,” “hour,” and

"smudges." Topic 3 talks about contents of the mascaras because it consists of "volume," "length(en)," and "color." Additionally, topic 9 describes the attribute of brush (such as "brush" and "bristle"). Figure 1 shows the proportion of topics for which the topic distributions ($\theta$) is aggregated for all the customer reviews. These topics 1, 3, and 9 are mentioned in many reviews. Topic 2, 7, and 10 seem to be less discussed in the reviews but they are a collection of words associated with reviews written by the consumers who received the complimentary samples. These three topics are closely related each other in the point of the semantically meaning, but they are extracted as separate topics.

Finally, we discuss the estimation results of regression coefficients and thresholds, and Table 3 provides the posterior mean estimates and the 95% highest posterior density (HPD). The threshold parameters ($\tau_r$) indicate that an approximate 0.50 increase in the latent continuous rating is associated with a one-point increase in the observed discrete rating. For example, if the proportion that a review mentions about the topic "Eyelash" increase by one unit, the expected change in the latent rating is 0.538, translating to an almost one-point increase in customer satisfaction.

Table 3 also provides some interesting findings related to the regression coefficients. The coefficients of topics 3 and 5 for the satisfaction score are positively estimated with significance, and this findings indicates that satisfied customers are more likely to talk about "contents" and "eyelash" topics. In contrast, the coefficients of topics 1 and 9 are negatively estimated, which show that customers tend to be dissatisfied with the product attributes of "flaking" and "brush". Another topics 2, 7, and 10, which concern the critique of the use of product by those who have been offered free samples, are related to a little negative score or have no significant relationship. In addition, topic 6, which may appear in the reviews by the customers who have tried and bought the trial products such as smaller sizes, is negatively estimated. This topic can be considered to be in similar situation to the complimentary topics in terms of the reviews by first-time buyer, except that they do not allow to offer the free samples. Thus, the proposed model can estimate the impact of not only the

Table 1: Model comparison using log marginal density

| Number of topics | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|
| Log marginal density | -1492.66 | -1303.82 | -1383.23 | -1316.52 | -2226.84 | -1259.65 |

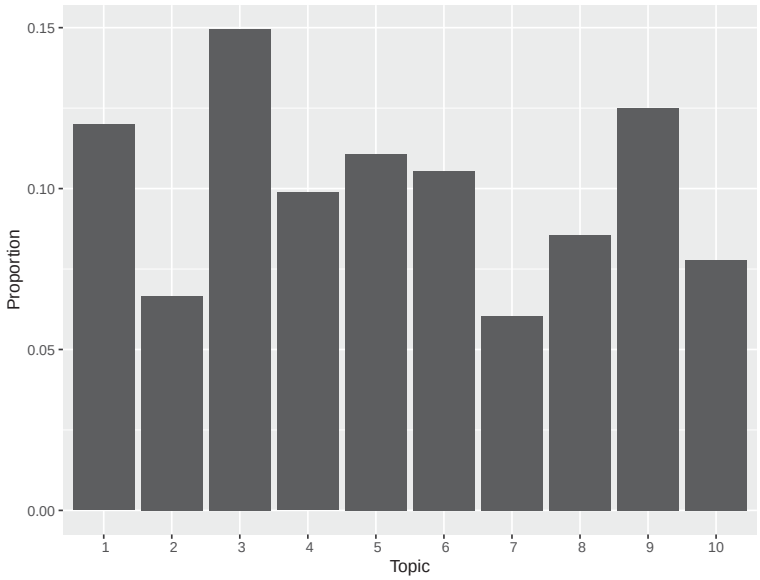


Figure 1: The proportion of topic distribution

product attributes, such as eyelash and brush, but also the situation in purchasing, such as complimentary and trying, on the customer satisfaction.

# 5 Conclusion

We introduced a model for capturing review generating behaviors based on word embedding approach and topic modeling by leaning feature vectors with respect to not only words but also topics or product attributes projected into the same feature space. The proposed model allows us to consider the order of words in the text and understand the context, which the traditional topic modeling approach for the customer review analysis do not take into account. To estimate the effect of the product attributes mentioned in the review text on customer satisfaction, we connect the proportion of the extracted topics in the review with

Table 2: Top 20 words in the proposed model

| Topic 1 "Flaking" | Topic 2 "Complimentary" | Topic 3 "Contents" | Topic 4 "Eyebrows" | Topic 5 "Eyelash" |
|---|---|---|---|---|
| eyes | complimentary | volume | weeks | eyelashes |
| off | encourage | and | results | lashes |
| under | review | gives | serum | look |
| raccoon | influenster | does | using | makes |
| water | purposes | length | eyebrows | long |
| day | testing | clump | see | straight |
| flakes | received | doesn | noticed | short |
| eye | free | great | difference | them |
| hours | honest | buildable | months | naturally |
| hour | opinions | mascara | grown | curled |
| face | voxbox | not | brows | thin |
| smudges | product | adds | grow | and |
| skin | item | color | now | fake |
| morning | iliabeauty | the | growth | longer |
| throughout | opinion | separates | saw | asian |
| flaking | complementary | lengthens | week | curl |
| makeup | influenstervoxbox | lengthening | month | blonde |
| burn | sampling | but | lash | are |
| eyelids | promotional | lashes | been | have |
| rubbing | test | builds | consistent | false |
| **Topic 6 "Trying"** | **Topic 7 "Complimentary"** | **Topic 8 "Too faced"** | **Topic 9 "Brush"** | **Topic 10 "Complimentary"** |
| reviews | recived | better | brush | complimentary |
| size | jus | sex | the | influenster |
| try | complimentary | than | wand | voxbox |
| sephora | purposes | mascaras | bristles | free |
| was | monsieurbigmascara | tried | comb | received |
| this | testing | drugstore | get | from |
| bought | review | camera | tube | testing |
| sample | purpose | cheaper | itself | test |
| pat | complimentry | faced | big | sent |
| buy | complementary | ysl | formula | purposes |
| buying | received | dior | applicator | review |
| mini | influenster | benefit | gets | gift |
| box | honest | years | apply | recieved |
| but | sampling | jacobs | clumpy | product |
| the | product | best | hard | complementary |
| mascara | free | nars | and | access |
| write | voxbox | ever | tip | holidayvoxbox |
| trial | courtesy | lancome | too | reviewing |
| play | from | holy | mess | vox |
| beauty | exchange | tarte | dries | via |

Table 3: The estimated coefficients of the topic regression (bold numbers indicate the significance of the posterior mean from the 95% HPD test)

| Parameter | Posterior Mean | 95% HPD Interval |
|---|---|---|
| $\tau_1$ | $\mathbf{-1.408}$ | [-1.446, -1.373] |
| $\tau_2$ | $\mathbf{-1.025}$ | [-1.041, -1.007] |
| $\tau_3$ | $\mathbf{-0.628}$ | [-0.647, -0.606] |
| $\tau_4$ | $\mathbf{-0.074}$ | [-0.081, -0.066] |
| $\beta_1$ (Flaking) | $\mathbf{-0.331}$ | [-0.458, -0.198] |
| $\beta_2$ (Complimentary) | 0.005 | [-0.160, 0.178] |
| $\beta_3$ (Contents) | $\mathbf{0.331}$ | [0.209, 0.450] |
| $\beta_4$ (Eyebrows) | $-0.018$ | [-0.150, 0.117] |
| $\beta_5$ (Eyelash) | $\mathbf{0.538}$ | [0.397, 0.674] |
| $\beta_6$ (Trying) | $\mathbf{-0.453}$ | [-0.585, -0.304] |
| $\beta_7$ (Complimentary) | $-0.064$ | [-0.233, 0.120] |
| $\beta_8$ (Too faced) | 0.119 | [-0.029, 0.272] |
| $\beta_9$ (Brush) | $\mathbf{-0.947}$ | [-1.070, -0.817] |
| $\beta_{10}$ (Complimentary) | $\mathbf{-0.204}$ | [-0.369, -0.046] |

the satisfaction rating score as the ordered probit regression. In the empirical study, we apply the proposed model for the e-commerce review dataset on the mascara products, and find some interpretable topics, such as "eyelash" and "brush", and estimate the their effects on satisfaction scores.

However, additional research is needed in the three aspects: First aspect is developing as marketing model through a consideration word sentiments and product brands. Referring to the existing literature, it is important for estimation of their impact on customer satisfaction to consider what emotions are used in conjunction with the product attributes and how much the unique influence of the product brand is, which are not taken into account in this study. Second aspect is validation of the advantages by taking a word embedding approach through model comparison. This study has only proposed a model for customer review analysis using word embedding and has not yet conducted comparison analysis to show an advantage over other existing models. Third aspect is the expansion of the empirical analysis to generalize the findings. In order to generalize the results such as those found in this empirical study,

it is necessary to analyze multiple product categories, in addition to the mascara category data conducted in this study.

# References


Archak, N., Ghose, A., and Ipeirotis, P. G. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, 57(8):1485–1509, 2011.

Barkan, O. and Koenigstein, N. ITEM2VEC: Neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2016.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., and Schweidel, D. A. Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing*, 84(1):1–25, 2020.

Bi, J.-W., Liu, Y., Fan, Z.-P., and Zhang, J. Wisdom of crowds: Conducting importance-performance analysis (IPA) through online reviews. *Tourism Management*, 70:460–478, 2019.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

Büschken, J. and Allenby, G. M. Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6):953–975, 2016.

Caselles-Dupré, H., Lesaint, F., and Royo-Letelier, J. Word2vec applied to recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 352–356, New York, NY, USA, 2018. ACM.

Decker, R. and Trusov, M. Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 27(4):293–307, 2010.

Fischer, G. W., Carmon, Z., Ariely, D., and Zauberman, G. Goal-Based Construction of Preferences: Task Goals and the Prominence Effect. *Management Science*, 45(8):1057–1075, 1999.

Ghose, A., Ipeirotis, P. G., and Li, B. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, 31(3): 493–520, 2012.

Hoeffler, S. Measuring Preferences for Really New Products. *Journal of Marketing Research*, 40(4):406–420, 2003.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, 2015.

Le, Q. and Mikolov, T. Distributed Representations of Sentences and Documents. In *Proceedings of the 31 st International Conference on Machine Learning*, pages 1188–1196, 2014.

Lee, T. Y. and Bradlow, E. T. Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5):881–894, 2011.

Liu, J. and Toubia, O. A Semantic Approach for Estimating Consumer Content Preferences from Online Search Queries. *Marketing Science*, 37(6):930–952, 2018.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed Representations

of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

Moody, C. E. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. 2016.

Moon, S. and Kamakura, W. A. A picture is worth a thousand words: Translating product reviews into a product positioning map. *International Journal of Research in Marketing*, 34(1):265–285, 2017.

Netzer, O., Toubia, O., Bradlow, E. T., Dahan, E., Evgeniou, T., Feinberg, F. M., Feit, E. M., Hui, S. K., Johnson, J., Liechty, J. C., Orlin, J. B., and Rao, V. R. Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters*, 19(3-4):337–354, 2008.

Newton, M. A. and Raftery, A. E. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994.

Puranam, D., Narayan, V., and Kadiyali, V. The effect of calorie posting regulation on consumer opinion: A flexible latent dirichlet allocation model with informative priors. *Marketing Science*, 36(5):726–746, 2017.

Qi, J., Zhang, Z., Jeon, S., and Zhou, Y. Mining customer requirements from online reviews: A product improvement perspective. *Information and Management*, 53(8):951–963, 2016.

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

Xiao, S., Wei, C. P., and Dong, M. Crowd intelligence: Analyzing online product reviews for preference measurement. *Information and Management*, 53(2):169–182, 2016.

Zhang, D., Xu, H., Su, Z., and Xu, Y. Chinese comments sentiment classification based on word2vec and SVMperf. *Expert Systems with Applications*, 42(4):1857–1863, 2015.

# Appendix

## A Estimation Procedure

In this appendix, we describe the details of estimating procedure for the proposed model. In this study, we apply a hybrid approach combining the gradient-based stochastic optimization and Markov Chain Monte Carlo (MCMC) sampling method. First, we describe the algorithm of the optimization for embedding vector parameters, which is known as Adam (Kingma and Ba, 2015). The objective loss function to minimize is defined as the negative mean of log likelihood:

$$L(\phi) = -\frac{1}{V}\sum_{i=1}^{V}\left\{\sum_{j\in S_i}\log\sigma\left(\overrightarrow{c}_i^{\top}\overrightarrow{w}_j\right) + \sum_{n\sim P_n(w)}^{N}\log\sigma\left(-\overrightarrow{c}_i^{\top}\overrightarrow{w}_n\right)\right\}, \tag{A.1}$$

where $\phi$ represents the set of parameters for embedding vectors. Let $f_h^{(s)} = \partial L(\phi^{(s)})/\partial\phi_h$ be the gradient of the loss function with respect to $h$-th parameter at $s$-th iteration, and we define two momentum terms as

$$m_h^{(s)} = \gamma_1 \times m_h^{(s-1)} + (1-\gamma_1)\times f_h^{(s)}, \quad m_h^{(0)} = 0 \tag{A.2}$$

$$v_h^{(s)} = \gamma_2 \times v_h^{(s-1)} + (1-\gamma_2)\times \left(f_h^{(s)}\right)^2, \quad v_h^{(0)} = 0 \tag{A.3}$$

$$\hat{m}_h^{(s)} = \frac{m_h^{(s)}}{1-\gamma_1^s} \tag{A.4}$$

$$\hat{v}_h^{(s)} = \frac{v_h^{(s)}}{1-\gamma_2^s}, \tag{A.5}$$

where $\gamma_1$ and $\gamma_2$ are coefficients used for computing running averages of the gradient and its square, and they are set to be 0.9 and 0.999, respectively, in this empirical study. Finally,

we update the focal parameter value using the following update rule,

$$\phi_h^{(s+1)} = \phi_h^{(s)} - \frac{\eta}{\sqrt{\hat{v}_h^{(s)}} + \epsilon} \times \hat{m}_h^{(s)}, \tag{A.6}$$

where $\eta$ is a learning rate parameter and set to be 0.001, and $\epsilon$ is a small value added to the denominator to improve numerical stability and set to be $10^{-8}$. Updating the embedding vectors via Adam optimization, we obtain their point estimates for each iteration. Next, given these embedding vector estimates, we conduct the MCMC sampling for the remaining parameters.

The conditional probability density and the sampling equations for the remaining parameters, the topic assignment $z_i = k$, the latent continuous rating $y_d^*$, regression coefficients $\beta$, and the thresholds $\tau_r$, are given as:

$$p(z_i = k \mid \{\overrightarrow{w}_i\}, \{\overrightarrow{t}_k\}, \theta_{d_i}) \propto \theta_{d_i} \times \prod_{j \in S_i} \sigma\left(\overrightarrow{c}_i^{\top} \overrightarrow{w}_j\right) \times \prod_{n \sim P_n(w)}^{N} \sigma\left(-\overrightarrow{c}_i^{\top} \overrightarrow{w}_n\right) \tag{A.7}$$

$$y_d^* \mid y_d, \theta_d, \beta, \tau \sim N\left(\sum_{k=1}^{K} \beta_k \theta_{dk}, 1\right), \text{ truncated to } (\tau_{r-1}, \tau_r] \text{ if } y_d = r \tag{A.8}$$

$$\beta \mid Y^*, \theta \sim N\left(\mu_\beta, \Sigma_\beta\right), \quad \Sigma_\beta = \left(\sum_{d=1}^{D} \theta_d \theta_d^{\top} + b_0^{-1} \cdot I\right)^{-1}, \quad \mu_\beta = \Sigma_\beta \left(\sum_{d=1}^{D} y_d^* \theta_d\right) \tag{A.9}$$

$$\tau_r \mid Y, Y^*, \tau_q \sim U[\tau_-^*, \tau_+^*], \quad r = 1, \ldots, R-1, q \neq r$$

$$\tau_-^* = \max(\max\{y_d^*; y_d = r\}, \tau_{r-1}), \quad \tau_+^* = \min(\min\{y_d^*; y_d = r+1\}, \tau_r), \tag{A.10}$$

where $b_0$ is a hyper parameter of the precision for the coefficients and set to be 0.01. Then, we employ the independence chain Metropolis-Hastings algorithm to estimate the topic distribution because the joint conditional density of $\theta_d$ is given by the product of the truncated normal density for $y_d^*$, the multinomial density of the topic assignments, and the Dirichlet density for the prior distribution, which the constant term of this posterior density is unknown. In the procedure of the sampling, we generate the candidate $\theta_d^{cand}$ from the proposal

density which is the Dirichlet distribution of the posterior density of the latent Dirichlet allocation model, $Dirichlet(N_{d\cdot} + \theta_0)$, where $N_{d\cdot}$ is a vector of the number of words to which each topic is assigned, and $\theta_0$ is a hyper parameter of the Dirichlet prior distribution and set to be 0.8. As a result of this generation for the candidate, all elements in the Metropolis acceptance ratio $\alpha$ cancel out, except for the likelihood component of the regression model. Therefore, the Metropolis ratio is given by the ratio of the truncated normal distribution,

$$\alpha = \frac{p(y_d^* \mid y_d, \theta_d^{cand}, \beta, \tau)}{p(y_d^* \mid y_d, \theta_d, \beta, \tau)}. \tag{A.11}$$

In the empirical study, we repeat the above updating process and the MCMC process 10,000 times, and then we use the 5,000 samples excluding the burn-in samples to calculate the posterior means and the intervals of the highest posterior density.