Interpretable Perceived Topics in Online Customer Reviews for Product Management

Aijing Xing and Nobuhiko Terui
Graduate School of Economics and Management, Tohoku University

Abstract

Online customer reviews contain useful and important information, because customers tend to praise or criticize certain features or attributes of goods in their reviews. We propose a model that extracts the perceived topics from textual reviews under the restrictions of their interpretability and predictability of product satisfaction as current product evaluation and expectation as future possible demand by supervised learning. The empirical analysis on of food reviews shows that our proposed model performs better than alternative models, and it suggests product managers the necessity of improving some specific attributes to fulfill customer needs.

Keywords and Phrases: Food satisfaction, Online customer reviews, Labeled topic model, Supervised learning, User-generated content, Text data

1 Introduction

With the rise of electronic commerce, online retailers such as Amazon, Walmart, and Taobao have experienced growth in their number of users. Most online retailers have customer feedback systems to receive information, including satisfaction scores (also called product ratings) and textual reviews. Performing text analysis of online customer reviews is useful for improving a seller's marketing strategy and customer management. Company managers can mine textual information about product features from the "voice of the consumer" and hence better meet customer needs when proposing new products (Fay and Xie, 2008, Griffin and Hauser, 1993, Xie and Gestner, 2007, Yohan and Alice, 2011). Lee and Bradlow (2011) and Netzer et al. (2012) exploit text-mined semi-structured product reviews to understand market structures based on the product attributes mentioned in the reviews.

This study aims at information extraction from review data and addresses two questions about customer experience and expectation:

"What product features are mentioned in the online customer reviews posted by satisfied and dissatisfied customers?"

"Whether readers expect to find helpful information about product features in customer reviews."

To answer these questions, we extend topic model for extracting perceived topics containing product features and predict satisfaction rating and helpfulness of product expectation. Topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) assumes each customer review consists of several topics with varying probabilities. The main problem of LDA model is the lack of interpretability. Promiscuous words sometimes are clustered in a common topic, which makes the topics hard to interpret. To improve interpretability, several modifications of LDA have been proposed in terms of incorporate supervision, such as Supervised LDA (Blei and McAuliffe, 2007) and labeled LDA (Ramage et al., 2009). This paper proposes the supervised–labeled LDA model by combining

supervised LDA and labeled LDA to extract interpretable perceived latent topics about product features from online customer reviews and predict customer satisfaction and product expectation.

2. Online Customer Review

2.1. Layout Form of Online Customer Reviews

Most online retailers have a similar layout form of online customer reviews, including textual content and product rating. For instance, the Amazon review includes three parts: (i)Satisfaction Score, (ii)Textual Review, and (iii)Helpfulness Numerator. Specifically, after buying and accepting goods, customers are allowed post their feelings and experiences by writing a textual review. At the same time, these customers may grade products with "stars" ranging from 1 to 5. The reader of this online review evaluate whether this review helpful for his/her product expectation and purchase decision by clicking "Yes" or unhelpful by clicking "No".

2.2. Textual Reviews and Perceived Topics

The "Amazon Fine Food Reviews" data consists of about 500,000 food reviews posted by Amazon users up till October, 2012. Specific brand of potato chips which has largest number of customer reviews is chosen as our research objects, and it contains 564 reviews. First, we reduce the whole vocabulary and avoids unimportant words such as *I, they, for, just, cannot, the, these, that, am, is* and so on. Below are two typical examples from the online review data, with the red words kept as meaningful words :

- Customer A
  - •Helpfulness Denominator: 13. Helpfulness Numerator: 7 (13 − 6).
  - •Satisfaction Score: 1
  - •Textual Review: I just cannot understand the high praise these chips have received. I ordered the variety pack and am very disappointed. All flavors have a really weird taste to them, except for maybe the plain chips. The salt and vinegar especially tasted gross! I took one bite and had to throw the bag away. Now I'm stuck with a bunch of chips that I'll never eat.
- Customer B
  - •Helpfulness Denominator: 35. Helpfulness Numerator: 33 (35 − 2).
  - •Satisfaction Score: 5
  - •Textual Review: I want to start out by saying that i thought at first that a bag with only 120 calories and 4 grams of fat (no saturated or trans) for every 20 chips was going to taste like crap. I must say that not only was i wrong, that this is my favorite BBQ chip on the market today. They are light and you cannot taste any fat or grease after eating them. That's because they aren't baked or fried, just popped as their name suggests. These chips are very easy to dip as well. FANTASTIC PRODUCT!

Based on the two typical examples, we suppose that product features are related with customer satisfaction and product expectation, specifically:

- Experienced customers praise or complain about this brand of chip by talking about topics about product features that include flavor, taste, ingredients, packing, cooking and consumption experience, private sentiments, and so on.

- The readers of reviews mostly have no consumption experience of the chip, and these novice customers expect to search for important product information from the reviews written by experienced customers. If the review contains informative topics of product features, the readers will recognize the review as helpful.

In order to find the possible food features in the reviews, we read the subset of the customer reviews and observe the descending sort of word frequencies of the whole vocabulary. These feature-based words are about flavor, taste, packing, weight, food ingredient, cooking method, and purchase. We combine synonymic topics and propose five perceived topics: (1) Flavor and Taste, (2) Packing, (3) Healthy, (4) Money, and (5) Ingredient.

3. Model

3.1 Supervised Learning Model with Labeled Topics

For the interpretability of the extracted topics in the text analysis, we first employ a labeled topic model. "Labeled" means that some labeled words are fixed to their respective topics in order to distinguish meanings among latent topics. Second, we use other information sources for the model to be supervised, where "supervised" means that the observed numeric data (or ordered categorical data) work as dependent variable, and the frequencies of topics work as covariates. In this article, we the model by combining two kinds of models and call it supervised–labeled LDA (SL–LDA) in the following.

Let $y_d = (y_{SAd}, y_{HNd}, y_{HDd})$ denote the vector of three continuous dependent variables of customer satisfaction, helpfulness numerator and helpfulness denominator in document *d.* The satisfaction reflects consumer's current evaluation based on their past experience, and we assume that helpfulness measures imply their interest and expectation on the product when they purchase in the future,

Given the word-topic assignment z_i in document d, $\sum_{i=1}^{N_d} \#\{z_i = k\}$ is the number of words assigned to topic k, and it works as explanatory variables after logarithm transformation. We define the supervised model by multivariate regression:

$$y_d = X_d \gamma + \delta_d, \quad d = 1,...,D, \tag{1}$$

where $X_d = [1, x_{1,d}\ x_{2,d}\ x_{3,d}\ x_{4,d}\ x_{5,d}]$, and we define the covariate

$$x_{k,d} = \log\left(\sum_{i=1}^{N_d} \#\{z_i = k\} + 1\right) \tag{2}$$

where N_d is the number of words of document d. We assume the error term follows multivariate normal distribution, $\delta_d \sim N_3\left(\mathbf{0}, diag\left(\sigma_1^2, \sigma_2^2, \sigma_3^2\right)\right)$.

Figure 1. Graphical model

The graphical model of SL-LDA is shown in Figure 1. Let $k \in \{1, \ldots, K\}$ and $d \in \{1, \ldots, D\}$, where K and D mean the number of topics and documents (customer reviews), respectively, and N_d stands for the number of words in document d. α is a K-dimensional vector acting as the Dirichlet prior for each row of θ, where θ is the $D \times K$ dimensional document-topic matrix defining mixing proportion. Denote $\beta = (\beta_1, \beta_2 ... \beta_V)$ and $\beta_k^* = (\beta_{\lambda_{1,k}}, \beta_{\lambda_{2,k}} ... \beta_{\lambda_{vk}} ... \beta_{\lambda_{VK}})$ as V and V_k dimensional vectors, respectively, where V is the size of the total vocabulary. V_k is the size of vocabulary $\{v_1^k, v_2^k, ... v_\lambda^k, ... v_{Vk}^k,\}$ for topic k, and $V_k < V$. Λ_k is defined as the transformation matrix for the reducing dimension of topic k. $\phi = \{\phi_1, \phi_2, ... \phi_k ... \phi_K\}$, where ϕ_k means the V_k dimensional topic-vocabulary vector. The term w = (w_1, w_2, ... w_{Nd}) is the occurred words of document d. z = (z_1, z_2, ... z_{Nd}), z_i is the word-topic assignment of word w_i.

The core of our modeling is to sample the probabilities with which the word w_i is assigned to topic k, and the probabilities are inferred from the following conditional posterior density :

$$p(z_i = k \mid z_{-i}, \theta, \phi_k, y_d, w_i,) \propto p(z_i = k \mid z_{-i}, \theta) p(w_i = v_\lambda^k \mid z_i = k, z_{-i}, \phi_k) p(y_d \mid z_i = k, z_{-i}, \gamma, \delta) \quad (3)$$

3.2 Transformation Matrix and Labeled Words

To make topics interpretable, some meaningful and frequently-used *labeled words* are chosen from the total vocabulary and fixed in respective perceived topics. At the same time, the remaining words which have no definite meaning are called as *non-labeled words* and assigned to topics with specific probabilities. For example, in Figure 2, given five topics about product features and total vocabulary of $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$, we assume that word v_1 works as the labeled word of Topic 1 and it is thus excluded from the other four topics. Other labeled words $v_2, v_3, v_4,$ and v_5 are specified as the labeled words for the remaining topics. The non-labeled words v_6 and v_7, can be assigned to any topic with certain probabilities; hence, the vocabulary of Topic 1 is $\{v_1, v_6, v_7\}$, which is the subset of total vocabulary of $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$.

Figure 2. Labeled and non-labeled words for respective topics

To build the vocabulary for topic k from total vocabulary, we imitate and simplify the methodology from labeled LDA (Ramage et al. 2009). We first generate the topic's *transformation matrix* $\Lambda^{(k)}$ ($V_k \times V$) conditional on $\{\lambda_1, \lambda_2 .. \lambda_{Vk}\}$, where $\{\lambda_1, \lambda_2 .. \lambda_{Vk}\}$ is the sequence number of probable words. The goal of "Labeled" is making the asymmetric vocabularies for each topic through $\Lambda^{(k)}$, in other words, we exclude other topics' labeled

words from the objective topic's vocabulary. Given the parameter vector β^*_k, we sample ϕ_k, a topic-vocabulary vector and obeys the Dirichlet distribution:

$$\phi_k \sim \mathrm{Dir}(\cdot \mid \beta^*_k) \tag{4}$$

Similar to collapsed Gibbs sampling of the labeled LDA model (Ramage et al. 2009), we sample the probabilities in which the topic assignment z_i of w_i is assigned to topic k according to the conditional posterior density:

$$p(z_i = k \mid z_{-i}, \alpha, \beta^*_k, w) \propto \frac{\left(n^{-i}_{d,k} + \alpha_k\right)}{n_d + \alpha^T \times 1} \frac{\left(n^{-i}_{k,w=v^k_\lambda} + \beta^*_{\lambda_{vk}}\right)}{n_k + \beta^{*T}_k \times 1} \tag{5}$$

where $w_i = v^k_\lambda$ means the word w_i is chosen from vocabulary $\{v^k_1, v^k_2, v^k_3, ..v^k_\lambda ..\}$ of topic k. The term $n^{-i}_{d,k}$ means, except for z_i in document d, the number of words assigned to topic k. The term $n^{-i}_{k,\ w=v^k_\lambda}$ means, given the topic assignment k, the number of times that v^k_λ is chosen. In addition, if w_i is the labeled word of topic k like v_1 in Topic 1, we ignore the equation (5) and assign w_i to topic k directly because no Dirichlet prior exists for the other topic assignments of labeled word w_i.

3.3 Model Estimation

The conditional probability density of topic assignment $z_i = k$ is given as

$$p(z_i = k \mid w, y, z_{-i}, \alpha, \beta^*, \gamma, \delta) \propto \frac{\left(n^{-i}_{d,k} + \alpha_k\right)}{n_d + \alpha^T \times 1} \frac{\left(n^{-i}_{k,w=v^k_\lambda} + \beta^*_{\lambda_{vk}}\right)}{n_k + \beta^{*T}_k \times 1} p(y_d \mid z_{d,i} = k, \gamma, \delta) \tag{6},$$

where $i = 1, 2, 3 \ldots, N$.

Then, the MCMC process by Gibbs sampling is as follows:

(i) Initiate the topic assignment z of both labeled and non-labeled words.

(ii) For each document $d = 1, 2 \ldots D$, and for each word w_i, $i = 1, 2, \ldots, N_d$. If w_i is a non-labeled word, for topic $k = 1, 2, \ldots, K$, assign the topic based on (6), else directly assign the labeled word w_i to a certain topic according to step (i).

(iii) Given the topic assignment Z, sample the following posterior mean of the Dirichlet distribution:

$$\theta_{d,k} = \frac{n_{d,k} + \alpha_k}{\sum_K \left(n_{d,k} + \alpha_k\right)} \qquad \phi_{k,w=v^k_\lambda} = \frac{n_{k,w=v^k_\lambda} + \beta^*_{\lambda_{vk}}}{\sum_{V_k} \left(n_{k,w=v^k_\lambda} + \beta^*_{\lambda_{vk}}\right)} \tag{7}$$

(iv) Sample $p(\gamma, \delta \mid z, y)$ with Bayesian linear regression and return to step (ii).

4. Empirical Application

Our research object is a brand of potato chips, whose data set has 564 customer reviews. First of all, we need to choose the labeled words for perceived topics mentioned in section 2.2 appropriately. We select the following labeled words for their respective

perceived topics:

- **Flavor** and **Taste**: flavor, flavors, sweet, salt, taste, pepper, vinegar, tastes, tasty, delicious
- **Packing**: size, bag, box, bags
- **Healthy**: calories, calorie, fat, grease, greasy, healthy, health, unhealthy, nutrition, nutritious
- **Money**: price, money, cheap, expensive
- **Ingredient**: oil, flour, ingredient, ingredients, potato, sugar, starch, flakes

4.1 Word Classification

After MCMC process of Gibbs sampling, which takes 4000 iterations, we illustrate the word classifications of the top 15 words in Table 1, where we can clearly see that the word classification of SL–LDA has better interpretation than Supervised–LDA (SLDA). Generally, all five topics of SL-LDA contains corresponding words.

Table 1. Word Classification of SL–LDA and SLDA

4.2 Supervised Learning Model: Multivariate Regression

Besides the word classification, our other research purpose is to predict the dependent variables of customer satisfaction, helpfulness numerator and helpfulness denominator through the interpretable perceived topics about product features. We use DIC as goodness of fit to compare the models in Table 2, and HPD region to check the significance of estimated coefficients in Table 3. SL-LDA is the best model in our research. Specifically, the estimates of SL-LDA shows that "Satisfaction Scores" is positively affected by the topics of "Flavor and Taste" and "Health," and negatively affected by "Money and Buying" and "Ingredient."

Table 2. Model Comparison of DIC

Table 3. Estimates of Gamma in Linear Regression (0.1 HPD region test)

5. Discussion and Conclusion

According to the estimated coefficients of SL-LDA, "Satisfaction Score" is significantly affected by four topics. Specifically, dissatisfied customers are more likely to talk about "Flavor and Taste" and "Health," while satisfied customers usually talk about "Money and Buying" and "Ingredient." We suggest that the manufacturer of these potato chips improve the attributes of flavor, taste, and health, such as raising awareness and advertising their nutritive function, so that the needs of dissatisfied customers will be fulfilled. In addition, reviews including the topic "Health" are regarded as helpful by readers of online customer reviews, while the topic "Ingredient" weakens helpfulness. We deduce that novice customers highly expect the healthiness information from reviews, whereas they pay less attention to the reviews about ingredient issues, possibly because these readers are impatient to read the terminology. Consequently, retailers need to improve the healthy attributes of their products and advertise healthy attributes to meet novice customers' expectations.

References

Blei DM, Ng, AY, Jordan MI (2003) Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3(January):993–1022.

Blei DM, McAuliffe J (2007) Supervised topic models. *Adv. Neural Inf. Process Syst.* 21:121–128.

Chen Y, Xie J (2008) Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Sci.* 54(3):477–491.

Fay S, Xie J (2008) Probabilistic goods: A creative way of selling products and services. *Marketing Sci.* 27(2):674–690.

Griffin A, Hauser JR (1993) The voice of the customer. *Marketing Sci.* 12(1):1–27.

Lee T, Bradlow E (2011) Automated marketing research using online customer reviews. *J. Marketing Res*. 48(5):881–894.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci*. 31(3):521–543.

Ramage D, Hall D, Nallapati R, Manning CD (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proc. Conf. Empir. Methods Nat. Lang. Process* 1:248–256.

Xie J, Gestner E (2007) Service escape: Profiting from customer cancellations. *Marketing Sci*. 26(1):18–30.

Table 1. Word Classification of SL–LDA and SLDA

SL-LDA				
Flavor and Taste	**Packing**	**Health**	**Money and Buying**	**Ingredients**
chips	bag	calories	price	potato
flavor	bags	fat	save	rice
like	box	healthy	money	ingredients
taste	snack	chip	after	oil
good	chips	low	subscribe	potatoes
salt	size	calorie	expensive	flour
popchips	popchips	greasy	reading	corn
love	weight	snack	average	natural
great	great	chips	reviews	starch
really	single	crunch	thembr	ingredient
flavors	love	crunchy	helps	cake
tried	eat	grams	thicker	list
much	small	sodium	shipping	flakes
try	per	great	decided	sugar
bbq	ounce	salty	seems	seasoning
SLDA				
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
snack	sour	chips	weight	flour
calories	cream	flavor	watchers	rice
makes	onion	like	points	ingredients
perfect	cheddar	taste	chips	corn
bag	after	potato	portion	oil
popchips	original	good	love	starch
relatively	garlic	love	bags	ingredient
afternoon	reading	popchips	great	natural
gives	parmesan	bag	bag	flakes
each	idea	great	single	sugar
mine	reviews	salt	low	safflower
shipments	thicker	chip	serve	sodium
compared	helps	really	pop	seasoning
boxes	preference	tried	control	list
tasty	bagbr	flavors	popchips	preservatives

Table 2.Model Comparison of DIC

Model	DIC
LDA + Multivariate Linear Regression	2707.46
SLDA	2611.38
SL–LDA	186.66
SL–LDA with Independent Linear Regression	1559.92 + 1605.84 + 1603.81

Table 3. Estimates of Gamma in Linear Regression (0.1 HPD region test)

Explanatory Variable	Response Variable	LDA + Multivariate Linear Regression	SLDA	SL-LDA	SL-LDA with Independent Linear Regression
Intercept		−0.104081	−0.16475	0.17948	0.17365
γ1		***−0.082277****	***−0.45349****	***−0.35696****	***−0.17229****
γ2	Satisfaction	0.025689	−0.27023	−0.26295	***0.18504****
γ3	Scores	−0.047175	0.14893	***−0.16791****	***0.25785****
γ4		0.016622	−0.0499	***0.12293****	−0.03308
γ5		0.202006	−0.09601	***0.11739****	−0.06994
Intercept		−0.465789	−0.02932	***0.17587****	***−0.33906****
γ1		***−0.013201****	−0.0463	0.0261	***0.11597****
γ2	Helpfulness	0.152468	−0.0601	−0.02763	0.03422
γ3	Numerator	0.048836	0.04158	***0.26881****	−0.03415
γ4		0.007903	***0.14977****	−0.02411	−0.01212
γ5		0.096529	0.09486	***−0.10661****	0.05451
Intercept		−0.296761	***0.23187****	−0.05071	−0.2654
γ1		0.030989	0.10702	−0.01633	***0.11932****
γ2	Helpfulness	0.101681	0.0237	0.05048	−0.02704
γ3	Denominator	***0.039018****	−0.06036	−0.07703	***−0.10624****
γ4		−0.047418	0.04784	0.05149	0.0375
γ5		0.021323	0.04359	0.06881	0.06569

Figure 1. Graphical model

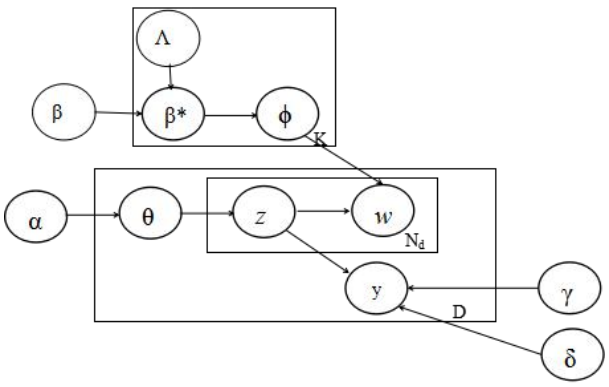


Figure 2. Labeled and non-labeled words for respective topics

$$\begin{array}{l} \\ Topic\ 1, \beta_1^* \\ Topic\ 2, \beta_2^* \\ Topic\ 3, \beta_3^* \\ Topic\ 4, \beta_4^* \\ Topic\ 5, \beta_5^* \end{array} \begin{array}{c} \begin{array}{ccccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 \end{array} \\ \begin{bmatrix} \beta_1 & \times & \times & \times & \times & \beta_6 & \beta_7 \\ \times & \beta_2 & \times & \times & \times & \beta_6 & \beta_7 \\ \times & \times & \beta_3 & \times & \times & \beta_6 & \beta_7 \\ \times & \times & \times & \beta_4 & \times & \beta_6 & \beta_7 \\ \times & \times & \times & \times & \beta_5 & \beta_6 & \beta_7 \end{bmatrix} \end{array}$$