# Predicting Sales from the Language of Product Descriptions

Reid Pryzant
Stanford University
rpryzant@stanford.edu

Young-joo Chung
Rakuten Institute of Technology
yjchung@acm.org

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

## ABSTRACT

What can a business say to attract customers? E-commerce vendors frequently sell the same items but use different marketing strategies to present their goods. Understanding consumer responses to this heterogeneous landscape of information is important both as business intelligence and, more broadly, a window into consumer attitudes. When studying consumer behavior, the existing literature is primarily concerned with product reviews. In this paper we posit that textual product descriptions are also important determinants of consumer choice. We mine 90,000+ product descriptions on the Japanese e-commerce marketplace Rakuten and identify actionable writing styles and word usages that are highly predictive of consumer purchasing behavior. In the process, we observe the inadequacies of traditional feature extraction algorithms, namely their inability to control for the implicit effects of confounds like brand loyalty and pricing strategies. To circumvent this problem, we propose a novel neural network architecture that leverages an adversarial objective to control for confounding factors, and attentional scores over its input to automatically elicit textual features as a domain-specific lexicon. We show that these textual features can predict the sales of each product, and investigate the narratives highlighted by these words. Our results suggest that appeals to authority, polite language, and mentions of informative and seasonal language win over the most customers.

## CCS CONCEPTS

• **Information systems** → **Content analysis and feature selection**; • **Computing methodologies** → **Information extraction**; **Neural networks**;

## KEYWORDS

e-commerce, feature selection, neural networks, adversarial learning, natural language processing

**ACM Reference format:**
Reid Pryzant, Young-joo Chung, and Dan Jurafsky. 2017. Predicting Sales from the Language of Product Descriptions. In *Proceedings of SIGIR, Tokyo, Japan, August 2017 (SIGIR 2017 eCom),* 10 pages.

## 1 INTRODUCTION

The internet has dramatically altered consumer shopping habits. Whereas customers of physical stores can physically manipulate, test, and evaluate products before making purchasing decisions, the remote nature of e-commerce renders such tactile evaluations obsolete.

In lieu of in-store evaluation, online shoppers increasingly rely on alternative sources of information. This includes "word-of-mouth" recommendations from outside sources [9] and local product reviews [13, 18, 20]. These factors, though well studied, are only indirectly controllable from a business perspective [25, 52]. Business owners have considerably stronger control over their own product descriptions. The same products may be sold by multiple vendors, with each item having a different textual description (note that we take *product* to mean a purchasable object, and *item* to mean an individual e-commerce listing). Studying consumers' reactions to these descriptions is valuable both as business intelligence and as a new window into consumer attitudes.

The hypothesis that business-generated product descriptions affect consumer behavior (manifested in sales) has received strong support in prior empirical studies [22, 26, 34, 37, 39]. However, these studies have only used summary statistics of these descriptions (i.e. readability, length, completeness). We propose that embedded in these product descriptions are narratives that affect shoppers, which can be studied by examining the words in each description.

Our hypothesis is that product descriptions are fundamentally a kind of social discourse, one whose linguistic contents have real control over consumer purchasing behavior. Business owners employ narratives to portray their products, and consumers react accordingly according to their beliefs and attitudes.

To test this hypothesis, we mine 93,591 product descriptions and sales records from the Japanese e-commerce website rakuten.co.jp ("Rakuten"). We build models that can explain how the textual content of product descriptions impacts sales. Second, we use these models to conduct a explanatory analysis, identifying what linguistic aspects of product descriptions are the most important determinants of success.

**We seek to unearth actionable phrases that can help e-commerce vendors increase their sales regardless of what's being sold**. Thus, we want to study the effect of language on sales *in isolation*, i.e. find textual features that are untangled from the effects of pricing strategies [15], brand loyalty [17, 48], and product identity. Choosing features for such a task is a challenging problem, because product descriptions are embedded in a larger e-commerce experience that *leverages* the shared power of these confounds to market a product. For a not-so-subtle example, product descriptions frequently boast "free shipping!", overtly pointing to a pricing strategy with known power over consumer choice [19].

We develop a new text feature selection algorithm to operate in this confound-controlled setting. This algorithm makes use of a novel neural network architecture. The network uses attentional


In: J. Degenhardt, S. Kallumadi, M. de Rijke, L. Si, A. Trotman, Y. Xu (eds.): Proceedings of *the SIGIR 2017 eCom workshop, August 2017, Tokyo, Japan*, published at http://ceur-ws.org

scores over its input and an adversarial objective to select a lexicon that is simultaneously predictive of consumer behavior and controlled for confounds such as brand and price.

We evaluate our feature selection algorithm on two pools of feature candidates: morphemes obtained with the JUMAN tokenizer[1], and sub-word units obtained via byte-pair encoding ("BPE") [47]. From these pools we select features with either (1) our proposed neural network, (2) odds ratios [10], (3) mutual information [41], and (4) the features with nonzero coefficients of a L1 regularized linear regression. Our results suggest that lexicons produced by the neural model are both less correlated with confounding factors and the most powerful predictors of sales.

In summary, our contributions are as follows:

- We demonstrate that the narratives embedded in e-commerce product descriptions influence sales.
- We propose a novel neural architecture to mine features for the task.
- We discover actionable writing styles and words that have especially high influence on these outcomes.

## 2 PREDICTING SALES FROM DESCRIPTIONS

Our task is to predict consumer demand (measured in $\log(sales)$) from the narratives embedded in product descriptions. To do so, we mine features from these textual data and fit a statistical model. In this section, we review our feature-mining baselines, present our novel approach to feature-mining, and outline our statistical technique for predicting sales from these features while accounting for confounding factors like brand loyalty and product identity.

### 2.1 Feature Mining Preliminaries

We approach the featurization problem by first segmenting product descriptions into sequences of tokens, then selecting tokens from the vocabulary of tokens that are predictive of high sales. We take subsets of these vocabularies (rather than one feature per vocabulary item) because (1) we need to be able to examine the linguistic contents of the resulting feature sets, and (2) we need models that are highly generalizable, and not too closely adapted to the peculiarities of these data's vocabulary distributions.

We select predictive subsets of the data's tokenized vocabularies in four ways. Three of these (Section 2.2) are traditional feature selection methods that serve as strong baselines for our proposed method (Section 2.3).

### 2.2 Traditional Feature Mining

**Odds Ratio** (OR) finds words that are over-represented in a particular copora when compared to another (e.g. descriptions of high selling items verses those of low-selling counterparts). Formally, this is:

$$\frac{p_i/(1-p_i)}{p_j/(1-p_j)} \tag{1}$$

where $p_i$ is the probability of the word in copora $i$ (e.g high-selling descriptions) and $p_j$ is the probability of the word in copora $j$ (e.g low-selling descriptions). Note that this method requires dichotomized targets, which we discuss further in Section 3.1.

**Mutual information** (MI) is a measurement of how informative the presence of a token is to making correct classification decisions. Formally, the mutual information $MI(t, c)$ of a token $t$ and binary class $c$ is

$$MI(t,c) = \sum_{I_t \in \{1,0\}} \sum_{I_t \in \{1,0\}} P(I_t, I_c) \log \frac{P(I_t, I_c)}{P(I_t)P(I_C)} \tag{2}$$

where $I_t$ and $I_c$ are indicators on term presence and class label for a given description. Like OR, this method requires dichotomized sales targets.

**Lasso Regularization** (L1) can perform variable selection on a linear regression model [51] by including a regularization term to the least squares objective. This term penalizes the L1 norm of the model parameters:

$$\arg\min \Big\{ \sum_{i=1}^{N} \Big( y_i - \beta_0 - \sum_j \beta_j x_{ij} \Big) \Big\}, \tag{3}$$

$$subject\ to \sum_j |\beta_j| \leq \alpha \tag{4}$$

.

Where $y_i$ is the $i$th target, $\beta_0$ is an intercept, $\beta_j$ is the $j$th coefficient of the $i$th predictor $x_i$. $\alpha$ is pre-specified parameter that determines the amount of regularization. The parameter $\alpha$ can be obtained by minimizing the error in cross-validation.

### 2.3 Deep Adversarial Feature Mining

An important limitation of all the aforementioned feature selection methods is that they are incapable of selecting features that are decorrelated from confounds like brand and price. Recall from Section 1 the price-related example of "free shipping!". Consider the brand-related example of "the quality you know and love from Daison". Though effective marketing tools, these phrases leverage the power of pricing strategies and brand loyalty, factors with known power over consumers. We wish to study the impact of linguistic structures in product descriptions *in isolation*, beyond those indicators of price or branding. Thus, we consider brand, product, and price information as confounding factors that confuse the effect of language on consumers.

As a solution to this problem, we propose a novel feature-selecting neural network (RNN+/-GF), sketched in Figure 1. The model uses an attention mechanism to produce estimates for $\log(sales)$, $brand$, and $price$. We omit $product$ because it is only present in our test data; see Section 3.1 for details. During training, the model uses an adversarial objective to discourage feature effectiveness with respect to two of these prediction targets: $brand$ and $price$. That is, the model finds features that are good at predicting sales, and bad at predicting brand and price.

**Deep learning review.** Before we describe the model, we review its primary building blocks.

**Feedforward Neural Networks (FFNNs)** are composed of a series of fully connected layers, where each layer takes on the form

$$\boldsymbol{y} = f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}). \tag{5}$$

[1]JUMAN (a User-Extensible Morphological Analyzer for Japanese), http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

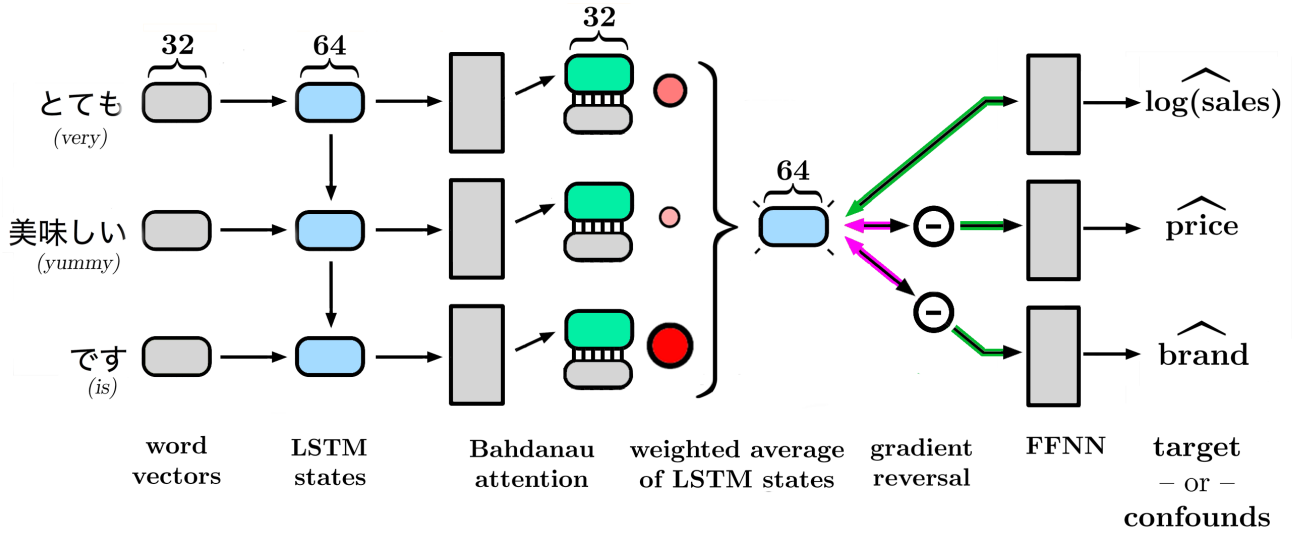


**Figure 1: An illustration of the proposed RNN+GF model operating on an example product description with three timesteps. All operations and dimensionalities are explicitly shown. Vectors are depicted as rounded rectangles, matrix multiplications as squared rectangles, and scalars as circles. Trainable parameters are grey, while dynamically computed values are colored. Gradient reversal layers multiply gradients by -1 as they backpropagate from the prediction networks to the encoder. In this example, the model attends to the description's final token the most, so that would be the most likely candidate for a generated lexicon.**

Note that $\boldsymbol{x} \in \mathbb{R}^n$ is a vector of inputs (e.g. from a previous layer), $\boldsymbol{W} \in \mathbb{R}^{y \times n}$ is a matrix of parameters, $\boldsymbol{b} \in \mathbb{R}^y$ is a vector of biases, $\boldsymbol{y} \in \mathbb{R}^y$ is an output vector, and $f(\cdot)$ is some nonlinear activation function, e.g. the ReLU: $ReLU(x) = \max\{0, x\}$.

**Recurrent Neural Networks (RNNs)** are effective tools for learning structure from sequential data [14]. RNNs take a vector $\boldsymbol{x}_t$ at each timestep. They compute a hidden state vector $\boldsymbol{h}_t \in \mathbb{R}^h$ at each timestep by applying nonlinear maps to the the previous hidden state $\boldsymbol{h}_{t-1}$ and the current input $\boldsymbol{x}_t$ (note that $\boldsymbol{h}_0$ is initialized to $\vec{0}$):

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{W}^{(hx)}\boldsymbol{x}_t + \boldsymbol{W}^{(hh)}\boldsymbol{h}_{t-1}\right). \tag{6}$$

$\boldsymbol{W}^{(hx)} \in \mathbb{R}^{h \times n}$, $\boldsymbol{W}^{(hh)} \in \mathbb{R}^{x \times h}$ are parameterized matrices. We use Long Short-Term Memory Network (LSTM) cells, a variant of the traditional RNN cell that can more effectively model long-term temporal dependencies [23].

**Attention mechanisms**. Attentional mechanisms allow neural models to focus on parts of the encoded input before producing predictions. We calculate Bahdanau-style attentional contexts [3] because these have been shown to perform well for other tasks like translation and language modeling [11, 31], and preliminary experiments suggested that this mechanism worked best for our problem setting.

Bahdanau-style attention computes the attentional context as a weighted average of hidden states. The weights are computed as follows: pass each hidden state $\boldsymbol{h}_i$ through a fully-connected neural network, then compute a dot product with a vector of parameters to produce an intermediary scalar $\hat{a}_i$ (eq. 7). Next, the $\hat{a}_i$'s are scaled by a softmax function so that they map to a distribution over hidden states (eq. 8). Finally, this distribution is used to compute a weighted average of hidden states $\boldsymbol{c}$ (eq. 9). Formally, this can be written as:

$$\hat{a}_i = \boldsymbol{v}_a^\intercal \tanh(\boldsymbol{W}_a \boldsymbol{h}_i) \tag{7}$$

$$\boldsymbol{a} = \text{softmax}(\hat{\boldsymbol{a}}) \tag{8}$$

$$\boldsymbol{c} = \sum_j a_j \boldsymbol{h}_j \tag{9}$$

**Our model.** We continue by describing our adversarial feature mining model. The process of obtaining features from the model can be thought of as a three-stage algorithm: (1) forward pass, where predictions are generated, (2) backward pass, where parameters are updated, and, after repeated iterations of 1 and 2, (3) feature selection, where we use attentional scores to elicit lexicons.

The **forward pass** operates as follows:

(1) The segmented input is fed into an LSTM to produce hidden state encodings for each timestep.
(2) We compute an attentional summary of these hidden states to obtain a single vector encoding of the input.
(3) We feed this encoding into three FFNNs. One is a regression network that tries to minimize $\mathcal{L} = ||\hat{y} - x||_2$, the squared loss between the predicted and true $\log(price)$. The second and third are classification networks, which predict a likelihood distribution over all possible labels, and are trained to minimize $\mathcal{L} = -\log p(y)$, the negative log probability of the correct class label. We attach classification networks for brand id and a dochotomization of price (see Section 3.1 for details). We dichotomized sales in this way to create a fair comparison between this method

and the baselines: other feature selection algorithms (OR, MI) are not so flexible and require dichotomized targets.

The **backward pass** draws on prior work in leveraging adversarial objective functions to match feature distributions in different settings [40]. In particular, we draw from a line of research in the style of [16], [8], and [27]. This method involves passing gradients through a *gradient reversal layer*, which multiplies gradients by a negative constant, i.e. -1, as they propagate back through the network. Intuitively, this encourages parameters to update *away* from the optimization objective.

If $\mathcal{L}_{sales}$, $\mathcal{L}_{brand}$, $\mathcal{L}_{price}$ are the regression and classification losses from each prediction network, then the final loss we are optimizing is $\mathcal{L} = \mathcal{L}_{sales} + \mathcal{L}_{brand} + \mathcal{L}_{price}$. However, when backpropagating from each prediction network to the encoder, we reverse the gradients of the networks that are predicting confounds. This means that the prediction networks still learn to predict *brand* and *price*, but the encoder is forced to learn brand- and price-invariant representations which are not useful to these downstream tasks. We hope that such representations encourage the model to attend to confound-decorrelated tokens.

The **lexicon induction** stage uses a trained model defined above to select textual features that are predictive of sales, but control for the influence of brand and price. This stage operates as follows:

(1) Generate predictions for each test item, but rather than saving those predictions, save the attentional distribution over each source sequence.
(2) Standardize these distributions. For each input $i$, standardize the distribution over timesteps $\mathbf{p}^{(i)}$ by computing

$$\mathbf{z}^{(i)} = \frac{\mathbf{p}^{(i)} - \boldsymbol{\mu}_{\mathbf{p}}^{(i)}}{\boldsymbol{\sigma}_{\mathbf{p}}^{(i)}} \tag{10}$$

(3) Merge these standardized distribution over each input sequence. If there is a word collision (i.e. we observe the same token in multiple input sequences and the model assigned each observation a different z-score), take the max of those words' z-scores.
(4) Select the $k$ tokens with highest z-scores. This is our induced lexicon.

## 2.4 Using Features to Predict Sales

Once we have mined textual features from product descriptions, we need a statistical model that accounts for the effects of confounding variables like product identity and brand loyalty in predicting the sales of each item. We use a mixed-effects model, a type of hierarchical regression that assumes observations can be explained with two types of categorical variables: fixed effect variables and random effect variables [7].

We model textual features as fixed effects. We take the product that each item corresponds to and the brand selling each item as random effects. Thus, we force the model to assume that product and brand information is decorrelated from everything else, and we expect to observe the explanatory power of text features without the influence of brand or product. Note that the continuous nature of the "price" confound precludes our ability to model it (Section 3.1).

We proceed with a formal description of our mixed-effects model. Let $y_{ijk}$ be the log(*sales*) of item $i$, which is product $j$ and sold by brand $k$. The description for this item is written as $\mathbf{x}_{ijk}$, and each $x_{ijk}^{(h)} \in \mathbf{x}_{ijk}$ is the $h^{\text{th}}$ feature of this description. With these definitions, we can write our mixed-effects model model as

$$y_{ijk} = \beta_0 + \sum_h \beta_h x_{ijk}^{(h)} + \gamma_j + \alpha_k + \epsilon_{ijk} \tag{11}$$

$$\gamma_j \sim \mathcal{N}(0, \sigma_\gamma^2) \tag{12}$$

$$\alpha_k \sim \mathcal{N}(0, \sigma_\alpha^2) \tag{13}$$

$$\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{14}$$

where $\gamma_j$ and $\alpha_k$ are the random effects of product and brand, respectively, and $\epsilon_{ijk}$ is an item-specific effect, i.e. this item's deviation from the mean item sales.

Nakagawa and Schielzeth [44] introduced the marginal and conditional $R^2$ ($R_m^2$ and $R_c^2$) as summary statistics of mixed-effects models. Marginal $R_m^2$ is the $R^2$ of the *textual effects only*. It reports the proportion of variance in the model's predictions can be explained with fixed effects variables $x_{ijk}^{(h)}$. It is written as;

$$R_m^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}, \tag{15}$$

$$\sigma_f^2 = var\left(\sum_h \beta_h x_{ijk}^{(h)}\right). \tag{16}$$

Conditional $R_c^2$ is the $R^2$ of the *entire model* (text + product + brand). It conditions on the variances of the random factors we are controlling for (product and brand):

$$R_c^2 = \frac{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\alpha^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}. \tag{17}$$

# 3 EXPERIMENTS

We now detail a series of experiments that were conducted to evaluate the effectiveness of each feature set, and, more generally, to test the hypothesis that narratives embedded in product descriptions are indeed predictive of sales.

## 3.1 Product and Sales Data

We obtained data on e-commerce product descriptions, sales, vendors, and prices from a December 2012 snapshot of the Rakuten marketplace[2]. We focused on items belonging to two product categories: chocolate and health. These two categories are both popular on the marketplace, but their characteristics are different. There is more variability among chocolate products than health products; many vendors are boutiques that sell handmade goods. Health vendors, on the other hand, are often large providers of pharmaceuticals goods, sometimes wholesale.

We segment product descriptions two ways. First, we tokenize descriptions into morphological units (morphemes) with the JUMAN tokenizer[3]. Second, we break descriptions into frequently occurring

[2]Please refer to https://rit.rakuten.co.jp/opendata.html for details on data acquisition.
[3]Using JUMAN (a User-Extensible Morphological Analyzer for Japanese), http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

sub-word units [4]. From here on we refer to the morpheme features as "morph", and sub-word features as "BPE".

Details of these data can be found in Table 1. Notably, the ratio of the size of vocabulary (unique keywords) to the size of tokens (occurrence of keywords) in the chocolate category is twice as large as that of the health category as listed in (%) in Table 1. This implies that product descriptions in the chocolate category are written with more diverse language.

Recall that some feature selection algorithms (OR, MI) require dichotomized prediction targets. Thus, we dichotomized the data on log(*sales*), taking the top-selling 30% and bottom-selling 30% as positive and negative examples, respectively. Our textual features were selected using these dichotomized data.

In order to evaluate mixed-effects regression models on these data, we consider the vendor selling an item as its "brand identifier" (vendors have unique branding on the Rakuten platform). We also need to know what product each item corresponds to, something not present in the data. Thus, we hand-labeled 2,131 items with product identifiers and separated these into a separate dataset for testing (Table 2). Our experimental results are reported on this test data set.

**Table 1: Characteristics of the Rakuten data. These data consist of 93,591 product descriptions, vendors, prices, and sales figures.**

| | Chocolate | Health |
|---|---|---|
| # items | 32,104 | 61,487 |
| # vendors | 1,373 | 1,533 |
| # morph tokens | 5,237,277 | 11,544,145 |
| # BPE tokens | 6,581,490 | 16,706,646 |
| # morph vocab (%) | 18,807 (0.36%) | 20,669 (0.18%) |
| # BPE vocab (%) | 16,000 (0.24%) | 16,000 (0.10%) |

**Table 2: Characteristics of the test data. Product identifiers were manually assigned to these data for evaluation.**

| | Chocolate | Health |
|---|---|---|
| # items | 924 | 1207 |
| # products | 186 | 50 |
| # vendors | 201 | 384 |
| avg. # items per product | 4 | 9 |
| (min, max) | (2, 26) | (2, 134) |

## 3.2 Experimental Protocol

All deep learning models were implemented using the Tensorflow framework [1]. In order to obtain features from the proposed RNN+GF model, we conducted a brief hyperparameter search on a held-out development set. This set consisted of of 2,000 examples randomly drawn from the pool of training data. The final model used 32-dimensional word vectors, an LSTM with 64-dimensional hidden states, and 32-dimensional intermediate Bahdanau vectors as described in Figure 1. Dropout at a rate of 0.2 was applied to the input of each LSTM cell. We optimized using Adam, a batch size of 128, and a learning rate of 0.0001 [30]. All models took approximately three hours to reach convergence on a Nvidia TITAN X GPU.

The L1 regularization parameter $\alpha$ was obtained with the scikit-learn library [45] by minimizing the error in the four-fold cross validation on training set.

In all of our experiments, we analyzed the log(*sales*) of an item as a function of textual description features. We used mixed-effects regression to model the relationship between these two entities. We included linguistic features obtained by the methods of Section 2.2 and 2.3 as fixed effect variables, and the confounding product/vendor identifiers in the test set as random effect variables. We used the "lme4" package in the R software environment v. 3.3.3 to perform these analyses [6]. To evaluate feature effectiveness and goodness of fit, we obtained conditional and marginal $R^2$ values with the "MuMIn" R package [5]. We also performed t-tests to obtain significance measurements on the model's fitted parameters. For this we obtained degrees of freedom with Satterthwaite approximations [46] with the "lmerTest" R package[32].

In addition to keywords, we experimented with two additional types of features: description length in number of keywords and part-of-speech tags obtained with JUMAN.

## 3.3 Experimental Results

**Influence of narratives.** Figure 2 depicts the performance of mixed-effects regression models fitted with the top 500 features from each approach. Overall, these results strongly support the hypothesis that narrative elements of product descriptions are predictive of consumer behavior. Adding text features to the model increased its explanatory power in all settings. The marginal $R^2_m$'s of each approach are listed on Table 3. The RNN+GF method selected features superior in both marginal and conditional $R^2$. This implies that it could select features that perform well in both isolated and confound-combined settings.

To investigate whether the high performance of RNN+GF features is simply a phenomenon of model capacity, we compared RNN+GF and one of the best-performing baselines, that of the lasso. We varied the number of features each algorithm is allowed to select and compared the resulting conditional $R^2$ values, finding that RNN+GF features are consistently on-par with or outperform that of the lasso, regardless of feature count as shown in Figure 3.

**Effect of gradient reversal** To determine the role of gradient reversal in the efficacy of the RNN+GF features, we conducted an ablation test, toggling the gradient reversal layer of our model and observing the performance of the elicited features. From Table 4, it is apparent that the confound-invariant representations encouraged by gradient reversal lead to more effective features being selected. Apart from summary statistics, this observation can be seen in the features themselves. For example, one of the highest scoring morphemes without gradient reversal was 無料

[4]Using https://github.com/google/sentencepiece

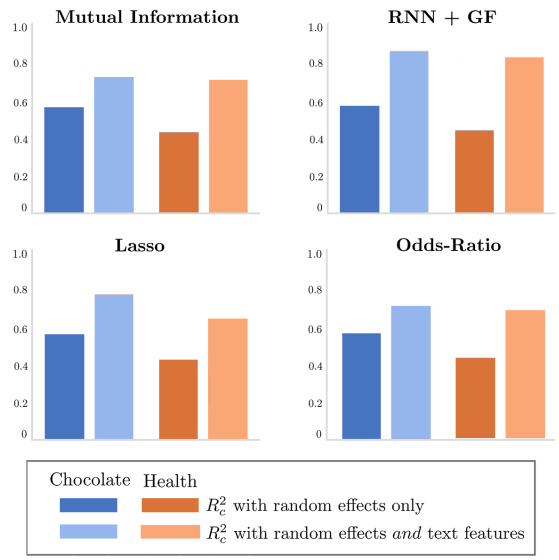


**Figure 2: Conditional $R^2$ of random effects only models (brand + product) and full models (brand + product + keywords + POS + BPE tokens) from Table 3. Including textual features in mixed effect regressions improves predictive power regardless of dataset and feature selection method features provide the largest gains. Morpheme tokens yielded similar results.**

**Table 3: The explanatory power of random effect confounds (brand, product), text (BPE features, description length, and POS tags), and the combination of confounds and text. Marginal and conditional $R^2$ are depicted where appropriate. The RNN+GF-selected features appear superior with and without confounds ($R_c^2$ and $R_m^2$). Morpheme features yielded similar results.**

| Chocolate | | | | | |
|---|---|---|---|---|---|
| Model features | $R^2$ type | L1 | MI | OR | RNN+GF |
| confounds only | *conditional* | 0.57 | 0.57 | 0.57 | 0.57 |
| text only | *marginal* | 0.58 | 0.53 | 0.49 | **0.60** |
| confounds + text | *conditional* | 0.78 | 0.73 | 0.71 | **0.81** |
| Health | | | | | |
| Model Features | $R^2$ type | L1 | MI | OR | RNN+GF |
| confounds only | *conditional* | 0.44 | 0.44 | 0.44 | 0.44 |
| text only | *marginal* | 0.40 | 0.40 | 0.36 | **0.44** |
| confounds + text | *conditional* | 0.65 | 0.71 | 0.69 | **0.78** |

(*"free"*). The RNN+GF features, on the other hand, are devoid of words relating to brand/vendor/price.

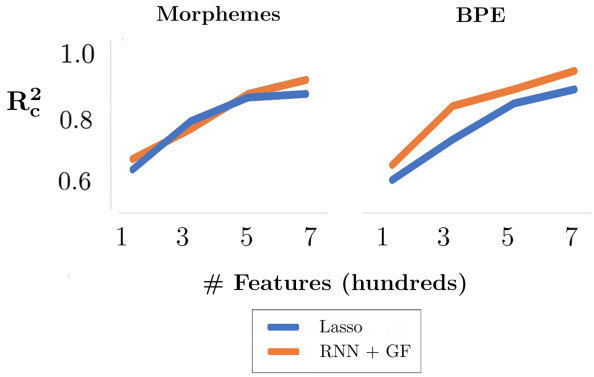


**Figure 3: Conditional $R^2$ ($R_c^2$) of the model trained varying numbers of of morpheme/BPE features. Despite being decorrelated from the random effects of brand and price, RNN+GF features are competitive with that of the lasso regardless of token type and feature set size.**

**Table 4: Gradient reversal ablation and its impact on conditional $R^2$. The confound-invariance encouraged by the adversarial objective helps downstream regressions.**

| | Chocolate | | Health | |
|---|---|---|---|---|
| | BPE | morph | BPE | morph |
| +GF | 0.81 | 0.81 | 0.78 | 0.75 |
| -GF | 0.76 | 0.75 | 0.64 | 0.69 |

**Comparison of different feature mining strategies.** To investigate whether the proposed method successfully discovered features that are simultaneously explanatory of sales and untangled from the confounding effects of product, brand, and price, we computed the correlations between BPE tokens selected by different methods and these non-linguistic confounds. For each feature set, the average per-feature Cramer's V was computed for product and brand, while the average per-feature point-biserial correlation coefficient was computed for price. Our results indicate that the RNN+GF features are less correlated with these confounds than any other method (Table 5).

**Table 5: Average association strengths between each BPE token set and non-linguistic factors. The RNN+GF features are the least correlated with these confounding factors. Morpheme tokens yielded similar results.**

| | L1 | MI | OR | RNN+GF |
|---|---|---|---|---|
| *product* | 0.55 | 0.57 | 0.55 | **0.38** |
| *brand* | 0.58 | 0.54 | 0.57 | **0.42** |
| *price* | 0.08 | 0.08 | 0.08 | **0.07** |

Examining the keywords selected by different methods suggests the same story as Table 5. Morpheme features with high importance

values are listed in Table 6. Note that the RNN+GF approach was the only method that did not select any keywords correlated with product, brand, or price. Additionally, every method except RNN+GF selected pecan (ピーカン・ペカン). *Lalala's pecan chocolate* is one of the most popular products on the marketplace. Although it is understandable that these tokens contribute to sales, they are product-specific and thus not generalizable. On the other hand, RNN+GF gave high scores to location-related words. Similar tendencies were observed in the health category. BPE tokens, though not listed, followed similar patterns.

### 3.4 Analysis

**Influential words.** To investigate the influence of keywords on sales, we performed t-tests on the coefficients of mixed-effects models trained with RNN+GF-selected features (both morphemes and BPE). We found out that influential descriptions generally contained words in the following four categories:

- **Informativeness** This includes informative appeals to logos with language other than raw product attributes (i.e. brand name, product name, ingredients, price, and shipping). Words like "family size" (ファミリーサイズ), "package design" (パッケージデザイン), "souvenir" (お土産), delimiters of structured information ("】【", "★", "●"), and indicators of detail ("x2", "70%", etc.) belong to this category.
- **Authority** This includes appeals to authority, in the form of authoritative figures or long-standing tradition. Words such as "staff" (スタッフ), "old-standing shop" (老舗), and "doctor" (お医者様) belong to this category.
- **Seasonality** These words suggest seasonal dependencies. Words such as "Christmas" (クリスマス), "Mother's day" (母の日), and "year-end gift" (歳暮) belong to this category. Note that words related to out-of-season events had low influence on sales.
- **Politeness** These expressions show politeness, respectfulness, and humbleness. Honorific Japanese (special words and conjugations reserved for polite contexts) such as "ing" (しております), "will do" (致します), "receive" (いただく) belong to this category.

The following are two differing descriptions of the exact same product. Words with high coefficients are shown in bold.

> Royce's chocolate has become a **standard** Hokkaido **souvenir**. They are packaged one by one so your hands won't get dirty! Also, our **staff** recommends this product!
> 北海道のお土産で定番品となっているロイズ. 手が汚れないように1本ずつパッケージされているのもありがたい! 当店 スタッフもおすすめするロイズの自信作です！
>
> Four types of nuts: almonds, cashews, pecans, macadamia, as well as cookie crunch and almond puff were packed carefully into each chocolate bar. This item is shipped with a refrigerated courier service during the **summer**.
> アーモンド、カシュー、ペカン、マカダミアの4種類のナッツとクッキークランチやアーモンドパフを一本のチョコレートバーにぎっしり詰め込みました。こちらは夏期クール便発送商品です。

The item with the former description was preferred by customers. It contains words suggestive of authority ("standard", "staff"), informativeness ("package", "souvenir"), and concern for the customer while the latter description is primarily concerned with ingredients.

**Influential part-of-speech tags.** We found a large number of adjectives and adverbs in our influential word lists. This agrees with the influential word categories mentioned previously, because adjectives and adverbs can be indicative of informativeness. We found that adjectives were more frequently influential in the chocolate category while adverbs were more common in the health category. Adjectives describing additional information such as "loved"(大好きだ), "healthy"(健康だ), and "perfect for"(ぴったりだ) had high coefficients in the chocolate category. Adverbs describing symptoms or effect such as "irritated"(イライラ) and "vigorously" (ガンガン) appeared in the health category.

## 4 RELATED WORK

In using large-scale text mining to characterize the behavior of e-commerce consumers, we draw on a large body of prior work in the space. Our inspiration comes from research on (i) unearthing the drivers of purchasing behavior in e-commerce, (ii) modeling the relationship between product presentations and business outcomes, and (iii) text mining and feature discovery in a confound-controlled setting.

There is an extensive body of literature on the progenitors of e-commerce purchasing behavior. Classic work in psychology has shown that human and judgment and behavior influenced by persuasive rhetoric [12, 49]. When our notions of human behavior are narrowed to purchasing decisions on the internet, despite the extreme diversity of online shoppers [38], prior work suggests that vendor-disseminated information exhibits a strong persuasive influence. In fact, vendor-disseminated information affects purchase likelihood just as much as user-generated information like word-of-mouth reviews [9]. The work of [22] incorporated vendor-disseminated product information into a model of customer satisfaction, a precursor of purchasing behavior [4]. Similar work has shown that product presentation (which entails textual descriptions) has a significant impact on perceived convenience [26] and credibility [36].

We also draw from prior research concerned with mining e-commerce information and predicting sales outcomes. Most of the work in this space is concerned with product *reviews*, not *descriptions*. [18] and [2] mined product reviews for textual features that are predictive of economic outcomes. This research used summary statistics of review text like length, Flesch-Kincaid readability scores [29], or, in the paradigm of [24], cluster membership in a semantic embedding space. Similar to us, [33] used product reviews to generate a domain-specific lexicon. However, this lexicon was used to predict sentiment, and then sales was predicted from sentiment. Some research has incorporated information from textual descriptions, but the best of these authors knowledge, the effect of descriptions alone is not studied. [42] used human subjects to illicit preferences

**Table 6: The highest-scoring morpheme tokens according to each feature selection algorithm. Tokens relating to confounds like brand, vendor or price are denoted with an asterisk. RNN+GF is the only method that avoided such tokens.**

| Chocolate | | | |
|---|---|---|---|
| **Lasso** | **Mutual Information** | **Odds-ratio** | **RNN+GF** |
| *小川 *(vendor address)* | 高温 *(hot)* | ペカン *(pecan)* | 神戸 *(kobe)* |
| *商店 *(vendor name)* | 株式 *(Co. Ltd)* | 百貨店 *(store dept.)* | 説明 *(description)* |
| 送信 *(send)* | 詳細だ *(detailed)* | ピーカン *(pecan)* | フランス *(france)* |
| さまざまだ *(various)* | *ロイズコンフェクト *(name)* | 新宿 *(shinjuku)* | オーストラリア *(australia)* |
| *有料 *(charge)* | *ロイズ *(brand name)* | 名人 *(master)* | タイ *(thailand)* |
| ショ糖 *(sucrose)* | 温度 *(temperature)* | 玉露 *(gyokuro)* | イタリア *(italy)* |
| 同時に *(simultaneous)* | 以下 *(under)* | *ララ *(product name)* | 老舗 *(long-standing shop)* |
| 制限 *(limit)* | セット *(set)* | 伴う *(come along)* | ハワイ *(hawaii)* |
| *買い得 *(bargain)* | 常温 *(room temp.)* | 会議 *(award name)* | ミルキー *(milky)* |
| ピーカン *(pecan)* | 保存 *(preserve)* | 会頭 *(award name)* | 蒜山 *(hiruzen)* |
| **Health** | | | |
| **Lasso** | **Mutual Information** | **Odds-ratio** | **RNN+GF** |
| 倍数 *(bulk unit)* | 消費 *(consumption)* | *アウトレット *(discount outlet)* | ダイエット *(weight loss)* |
| ビック *(big)* | *爽快 *(vendor name)* | アラゴナイト *(aragonite)* | 確認 *(confirmation)* |
| *淀川 *(vendor address)* | 見る *(see)* | ソマチット *(somatid)* | オレンジ *(orange)* |
| *アウトレット *(discount outlet)* | ブラウザ *(brower)* | ダントツ *(the very best)* | 予告 *(notice)* |
| *爽快 *(vendor name)* | 相談 *(consult)* | *アース *(brand name)* | 商品 *(product)* |
| 支店 *(branch)* | 形状 *(shape)* | *コリー *(product name)* | 注文 *(order)* |
| 地区 *(district)* | 対応 *(support)* | 筋骨 *(bones)* | 入金 *(payment)* |
| 鹿児島 *(kagoshima)* | ネット *(internet)* | ランナー *(runner)* | サプリ *(supplement)* |
| *スカルプ *(product name)* | 取り寄せる *(stock)* | *ガレノス *(brand name)* | 説明 *(explanation)* |
| くだもの *(fruit)* | 合す *(mix)* | 内外 *(inside and outside)* | ます *(is (formal))* |

between descriptions and actual products, but did not compare between descriptions. [53] tagged product descriptions with sentiment information and used this alongside review information to predict sales. Similarly, [21] and [54] used description labellings and summary statistics alongside other features to predict purchasing intent. Importantly, none of the prior work in this space seeks to untangle the influence of confounding hidden variables (e.g. brand loyalty, pricing strategies) from mined features.

Another body of research we draw from is that concerned with text mining and lexicon discovery in a confound-controlled setting. Using odds ratios to select features and hierarchical regression to determine their importance is a canonical technique in the computational linguistics literature [19, 28]. In general, alternative feature mining methods for downstream regression or classification tasks are rarely explored. [50] began with a set of hand-compiled corpora, then ran t-tests to prune these corpora of insignificant keywords. [43] developed a neural architecture that picks out keywords from a passage. However, this group did not use an attention mechanism to pick these words, and the model was developed for summarization applications. In the e-commerce literature, work alternatives to odds-ratio still rely on uncontrolled co-occurrence statistics [35].

## 5 CONCLUSION

In this paper, we discovered that that seasonal, polite, authoritative and informative product descriptions led to the best business outcomes in Japanese e-commerce.

In making these observations, we presented a statistical method that infers consumer demand from e-commerce product descriptions. We showed for the first time that words in the embedded narratives of product descriptions are important determinants of sales, even when accounting for the influence of factors like brand loyalty and item identity.

In the process, we noted the inadequacies of traditional text feature-selection algorithms, namely their ability to select features that are decorrelated from these factors. To this end we presented a novel neural network feature selection method. The features generated by this model are both high-performance and confound-decorrelated.

There are many directions for future work. These include extending our feature selectors to the broader setting of generalized lexicon induction, and applying our statistical models to e-commerce markets in other consumer cultures.

## ACKNOWLEDGMENTS

We are grateful to David Jurgens and Will Hamilton for their advice.

# REFERENCES


[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Nikolay Archak, Anindya Ghose, and Panagiotis G Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science* 57, 8 (2011), 1485–1509.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)* (2015).

[4] Billy Bai, Rob Law, and Ivan Wen. 2008. The impact of website quality on customer satisfaction and purchase intentions: Evidence from Chinese online visitors. *International journal of hospitality management* 27, 3 (2008), 391–402.

[5] Kamil Bartoń. 2013. MuMIn: Multi-model inference. R package version 1.9. 13. *The Comprehensive R Archive Network (CRAN), Vienna, Austria* (2013).

[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48.

[7] Douglas M Bates. 2010. lme4: Mixed-effects modeling with R. (2010).

[8] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, and others. 2007. Analysis of representations for domain adaptation. *Advances in neural information processing systems* 19 (2007), 137.

[9] Barbara Bickart and Robert M Schindler. 2001. Internet forums as influential sources of consumer information. *Journal of interactive marketing* 15, 3 (2001), 31–40.

[10] J Martin Bland and Douglas G Altman. 2000. The odds ratio. *Bmj* 320, 7247 (2000), 1468.

[11] Denny Britz, Anna Goldie, Thang Luong, and Quoc Le. 2017. Massive Exploration of Neural Machine Translation Architectures. *arXiv preprint arXiv:1703.03906* (2017).

[12] Shelly Chaiken, Mark P Zanna, James M Olson, and C Peter Herman. 1987. The heuristic model of persuasion. In *Social influence: the ontario symposium*, Vol. 5. Hillsdale, NJ: Lawrence Erlbaum, 3–39.

[13] Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43, 3 (2006), 345–354.

[14] Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14, 2 (1990), 179–211.

[15] Richard Friberg, Mattias Ganslandt, and Mikael Sandström. 2001. *Pricing strategies in e-commerce: Bricks vs. clicks.* Technical Report. IUI working paper.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17, 59 (2016), 1–35.

[17] David Gefen. 2002. Customer loyalty in e-commerce. *Journal of the association for information systems* 3, 1 (2002), 2.

[18] Anindya Ghose and Panagiotis G Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering* 23, 10 (2011), 1498–1512.

[19] Anindya Ghose and Arun Sundararajan. 2006. Evaluating pricing strategy using e-commerce data: Evidence and estimation challenges. *Statist. Sci.* (2006), 131–142.

[20] David Godes and Dina Mayzlin. 2004. Using online conversations to study word-of-mouth communication. *Marketing science* 23, 4 (2004), 545–560.

[21] Dennis Herhausen, Jochen Binder, Marcus Schoegel, and Andreas Herrmann. 2015. Integrating bricks with clicks: retailer-level and channel-level outcomes of online–offline channel integration. *Journal of Retailing* 91, 2 (2015), 309–325.

[22] Chin-Fu Ho and Wen-Hsiung Wu. 1999. Antecedents of customer satisfaction on the Internet: an empirical study of online shopping. In *Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on.* IEEE, 9–pp.

[23] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[24] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 168–177.

[25] Nan Hu, Jie Zhang, and Paul A Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (2009), 144–147.

[26] Ling Jiang, Zhilin Yang, and Minjoon Jun. 2013. Measuring consumer perceptions of online shopping convenience. *Journal of Service Management* 24, 2 (2013), 191–214.

[27] Fredrik D. Johansson, Uri Shalit, and David Sontag. 2016. Learning Representations for Counterfactual Inference. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16).* JMLR.org, 3020–3029.

[28] Dan Jurafsky, Victor Chahuneau, Bryan R Routledge, and Noah A Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* 19, 4 (2014).

[29] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Technical Report. DTIC Document.

[30] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference for Learning Representations* (2014).

[31] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[32] Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2015. Package 'lmerTest'. *R package version* 2 (2015).

[33] Raymond YK Lau, Wenping Zhang, Peter D Bruza, and Kam-Fai Wong. 2011. Learning domain-specific sentiment lexicons for predicting product sales. In *e-Business Engineering (ICEBE), 2011 IEEE 8th International Conference on.* IEEE, 131–138.

[34] Eun-Ju Lee and Soo Yun Shin. 2014. When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo. *Computers in Human Behavior* 31 (2014), 356–366.

[35] Thomas Lee and Eric T Bradlow. 2007. Automatic construction of conjoint attributes and levels from online customer reviews. *University Of Pennsylvania, The Wharton School Working Paper* (2007).

[36] Ziqi Liao and Michael Tow Cheung. 2001. Internet-based e-shopping and consumer attitudes: an empirical study. *Information & management* 38, 5 (2001), 299–306.

[37] Moez Limayem, Mohamed Khalifa, and Anissa Frini. 2000. What makes consumers buy from Internet? A longitudinal study of online shopping. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 30, 4 (2000), 421–432.

[38] Ying Liu, Hong Li, Geng Peng, Benfu Lv, and Chong Zhang. 2015. Online purchaser segmentation and promotion strategy selection: evidence from Chinese E-commerce market. *Annals of Operations Research* 233, 1 (2015), 263–279.

[39] Gerald L Lohse and Peter Spiller. 1998. Quantifying the effect of user interface design features on cyberstore traffic and sales. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM Press/Addison-Wesley Publishing Co., 211–218.

[40] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.* ACM, 641–647.

[41] Christopher D Manning, Hinrich Schütze, and others. 1999. *Foundations of statistical natural language processing.* Vol. 999. MIT Press.

[42] Deborah Brown McCabe and Stephen M Nowlis. 2003. The effect of examining actual products or product descriptions on consumer preference. *Journal of Consumer Psychology* 13, 4 (2003), 431–439.

[43] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep Keyphrase Generation. *Annual Meeting of the Association for Computational Linguistics* (2017).

[44] Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4, 2 (2013), 133–142.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[46] Franklin E Satterthwaite. 1946. An approximate distribution of estimates of variance components. *Biometrics bulletin* 2, 6 (1946), 110–114.

[47] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany.*

[48] Srini S Srinivasan, Rolph Anderson, and Kishore Ponnavolu. 2002. Customer loyalty in e-commerce: an exploration of its antecedents and consequences. *Journal of retailing* 78, 1 (2002), 41–50.

[49] Brian Sternthal, Ruby Dholakia, and Clark Leavitt. 1978. The persuasive effect of source credibility: Tests of cognitive response. *Journal of Consumer research* 4, 4 (1978), 252–260.

[50] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* ACL, Baltimore, Maryland, 175–185.

[51] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[52] Lou W Turley and Ronald E Milliman. 2000. Atmospheric effects on shopping behavior: a review of the experimental evidence. *Journal of business research* 49, 2 (2000), 193–211.

[53] Hui Yuan, Wei Xu, Qian Li, and Raymond Lau. 2017. Topic sentiment mining for sales performance prediction in e-commerce. *Annals of Operations Research*

(2017), 1–24.
[54] Cai-Nicolas Ziegler, Lars Schmidt-Thieme, and Georg Lausen. 2004. Exploiting semantic product descriptions for recommender systems. In *Proceedings of the 2nd ACM SIGIR Semantic Web and Information Retrieval Workshop*. 25–29.