Identifying Customer Needs from User-Generated Content

by

Artem Timoshenko

and

John R. Hauser

April 2017

Artem Timoshenko is a PhD student at the MIT Sloan School of Management, Massachusetts Institute of Technology, E62-584, 77 Massachusetts Avenue, Cambridge, MA 02139, (617) 803-5630, atimoshe@mit.edu.

John R. Hauser is the Kirin Professor of Marketing, MIT Sloan School of Management, Massachusetts Institute of Technology, E62-538, 77 Massachusetts Avenue, Cambridge, MA 02139, (617) 253-2929, hauser@mit.edu.

We thank John Mitchell, Steven Gaskin, Carmel Dibner, Andrea Ruttenberg, Patti Yanes, Kristyn Corrigan and Meaghan Foley for their help and support. We thank Regina Barzilay, Clarence Lee, Daria Dzyabura, Dean Eckles, Duncan Simester, Evgeny Pavlov, Guilherme Liberali, Theodoros Evgeniou, and Hema Yoganarasimhan for helpful comments and discussions. We thank Ken Deal and Ewa Nowakowska for suggestions on earlier versions of this paper. This paper has benefited from presentations at the 2016 Sawtooth Software Conference in Park City Utah, the MIT Marketing Group Seminar, and presentations at Applied Marketing Science, Inc. and Cornerstone Research, Inc.

Identifying Customer Needs from User-Generated Content

Abstract

Identifying customer needs is important to marketing strategy, product development, and marketing research. User-generated content (UGC) provides an opportunity to better identify customer needs for managerial impact. However, established methods are neither efficient nor effective for large UGC corpora because much content is non-informative and repetitive. We propose a machine-learning approach to select content for efficient review. We use a convolutional neural network to filter out non-informative content and cluster dense sentence embeddings to avoid sampling repetitive content. We further address two key questions: Are customer needs identified in UGC comparable to customer needs identified with standard methods? Do the machine-learning methods improve customer-need identification? These comparisons are enabled by a custom data set of customer needs for oral care products identified by professional analysts using industry-standard experiential interviews. The same professional analysts coded 12,000 UGC sentences to identify if each sentence contained one or more previously identified customer needs and/or new customer needs. Results: Customer needs identified from UGC are at least as valuable for product development, likely more-valuable, than those identified by conventional methods and (2) machine-learning methods improve efficiency (unique customer needs identified per unit of professional services cost).

Keywords: *Voice of the Customer; Machine Learning, User-generated Content; Customer Needs; Online Reviews; Market Research; Text Mining; Deep Learning; Natural Language Processing*

1. Introduction

In marketing strategy, customer needs help segment the market, identify strategic dimensions for differentiation, and make efficient channel management decisions. For example, Park, Jaworski, and MacInnis (1986) describe examples of strategic positioning based on fulfilling customer needs: "attire for the conservative professional" (Brooks Brothers) or "a world apart—let it express your world" (Lenox China). In product development, customer needs identify new product opportunities (Herrmann, Huber, and Braunstein 2000), improve the design of new products (Krishnan and Ulrich 2001; Sullivan 1986; Ulrich and Eppinger 2004), help manage product portfolios (Stone, et al. 2008), and improve existing products and services (Matzler and Hinterhuber 1998). In marketing research, customer needs help to identify the attributes used in conjoint analysis (Orme 2006).

New "excitement" customer needs are particularly important for product development (Kano, et al. 1984; Mikulić and Prebežac 2011). For example, consider the breakthrough laundry detergent, "Attack," developed by the Kao Group. Before Kao's innovation, firms such as Procter & Gamble thought they knew all (primary) customer needs: cleaning, safe and gentle, good for the environment, ready to wear after drying, easy to use, smell fresh and clean, and value. Because most products fulfilled these customer needs, perceived value played a major competitive role. Detergents were sold in large "high-value" boxes. Kao recognized that Japanese consumers did not have space for large boxes in their apartments, nor could they transport them easily by foot or bicycle from the store. To fulfill this newly-identified customer need, Kao launched a highly-concentrated detergent in an easy-to-store and easy-to-carry package. Despite a premium price, Attack quickly commanded almost 50% of the Japanese laundry market (Kao Group 2016). American firms soon introduced their own concentrated detergents.

A customer need is an abstract statement describing the benefits, in the customer's own words, that the customer seeks to obtain from a product or service (Brown and Eisenhardt 1995; Griffin, et al., 2009). For example, when describing their experience with mouthwashes, a customer might express the need "to know easily the amount of mouthwash to use." This customer need can be satisfied by various product attributes (solutions), including ticks on the cap and textual or visual descriptions on the bottle. Customer needs are described by precise sentences or phrases rather than "bags of words." For example, oral care products might be able to "maintain a natural shade of white that doesn't look phony no matter what products I use" or "easily get all particles, even the tiniest, out from between my teeth."

Because the identification of customer needs requires a deep understanding of a customer's experience, traditional methods rely on direct input from customers. Common methods to obtain direct input, such as experiential interviews and focus groups, are expensive and time-consuming, often resulting in delays in time to market. To avoid the expense and delays, some firms use heuristics, such as managerial judgment or the review of web-based product comparisons. However, heuristic methods often miss new "excitement needs" or new strategic positionings.

User-generated content (UGC), with its extensive rich textual content, is a promising source from which to identify customer needs more efficiently. UGC is available quickly and at low incremental cost to the firm. In many categories, UGC is extensive—for example, there are over 300,000 reviews on health and personal care products on Amazon alone. If UGC can be mined for customer needs, UGC has the potential to identify as many, or perhaps more, customer needs than direct customer interviews and to do so more quickly with lower cost. UGC provides additional advantages: (1) it is updated continuously enabling the firm to update its understanding of customer needs and (2) unlike customer interviews, firms can return to UGC at low cost to explore new insights further.

To use UGC, we must overcome three challenges. First, much UGC is repetitive and not relevant. Sentences such as "I highly recommend this product" do not express customer needs. We expect, and our analysis confirms, that most of UGC concentrates on a relatively few customer needs. Second, the very scale of UGC makes it difficult for human readers to process. Third, UGC data are unstructured and mostly text-based. To identify abstract customer needs, we must be able to identify and understand rich structure.

Our goals in this paper are two-fold. First, we examine whether a reasonable corpus of UGC provides sufficient content to identify customer needs. We construct and analyze a custom dataset in which we persuaded a professional marketing consulting firm to provide a complete coding of (a) experiential interview transcripts and (b) a sample of sentences from Amazon reviews in the oral-care category. Second, we develop and evaluate a machine-learning hybrid approach to identify customer needs from UGC. We use machine learning to identify relevant content and remove redundancy from a large UGC corpus, and then rely on human judgment to formulate customer needs from selected content. We draw on recent research in deep learning, in particular, convolutional neural networks (CNN; Collobert, et al. 2011; Kim 2014) and dense word and sentence embeddings (Mikolov, et al. 2013a; Socher, et al. 2013). The CNN filters out non-informative content from a large UGC corpus. Dense word and sentence embeddings embed semantic content in a real-valued vector space. We use sentence embeddings to sample a diverse set of non-redundant sentences for manual review. Both the CNN and word and sentence embeddings scale to large datasets. Manual review by professional analysts remains necessary in the last step because of the abstract nature of customer needs.

Our comparisons suggest that, if we limit costs to that required to review experiential interviews, then UGC provides a comparable set of customer needs to those obtained from experiential interviews. In industry, the dominant costs are the billing rates of experienced professionals, often called professional services costs. Even within cost constraints, UGC provides customer needs not identified from experiential interviews, although a few customer needs are missed. When we relax the professional services constraint for reviewing sentences, but maintain professional services costs to be less than would be required to interview and review, then UGC is a better source of customer needs. We further demonstrate that machine learning helps to eliminate irrelevant and redundant content and, hence, makes professional services investments more efficient.

2. Related Research

2.1. Traditional Methods to Identify Customer Needs (and Link Needs to Product Attributes)

Given a set of customer needs, product-development teams use a variety of methods, such as quality function deployment, to identify customer solutions or product attributes that address customer needs (Akao 2004; Hauser and Clausing 1988; Sullivan 1986). For example, Chan and Wu (2002) review 650 research articles that develop, refine, and apply QFD to map customer needs to solutions. Zahay, Griffin, and Fredericks (2004) review the use of customer needs in the "fuzzy front end," product design, product testing, and product launch. Customer needs can also be used to identify attributes for conjoint analysis (Green and Srinivasan, 1978; Orme, 2006). Kim, et al. (2017) propose a benefit-based conjoint-analysis model which maps product attributes to latent customer needs before estimation.

Researchers in marketing and engineering have developed and refined many methods to elicit customer needs directly from customers. The most common methods rely on focus groups, experiential interviews, or ethnography as input. Trained professional analysts then review the input, manually identify customer needs, remove redundancy, and structure the customer needs (Alam and Perry, 2002; Goffin, et al. 2012; Kaulio 1998). Some researchers augment interviews with structured methods such as repertory grids (Wu and Shich 2010).

Typically, customer-need identification begins with 20-30 qualitative experiential interviews. Multiple analysts review transcripts, highlight customer needs, and remove redundancy ("winnowing") to produce a basic set of approximately 100 abstract customer-need statements. Affinity groups or clustered customer-card sorts then produce a hierarchy of primary, secondary, and tertiary needs for product development and strategic positioning (Griffin and Hauser 1993; Jiao and Chen 2006). These methods are often called voice-of-the-customer (VOC) methods. Recently, researchers have sought to explore new sources of customer needs to supplement or replace common methods. For example, Schaffhausen and Kowlewski (2015; 2016) proposed using a web interface to ask customers to enter customer needs and stories directly. They then rely on human judgment to structure the customer needs and remove redundancy.

2.2. UGC Text Analysis in Marketing and Product Development

Researchers in marketing have developed a variety of methods to mine unstructured textual data to address managerial questions. See reviews in Büschken and Allenby (2016) and Fader and Winer (2012). The research closest to our goals uses word co-occurrences and variations of Latent Dirichlet Analysis (LDA) to characterize customer preferences or latent topics in product discussions (Archak, Ghose, and Panagiotis 2016; Büschken and Allenby 2006; Lee and Bradlow 2011; Tirunillai and Tellis 2014; Netzer, et al. 2012). LDA identifies “buckets of words,” that is, combinations of words that group together to form “topics.” LDA, and related methods, do not identify the deeper semantic structure of customer needs.

In engineering, the product attribute elicitation literature is closest to the goals of our paper, although the focus is on physical attributes rather than more-abstract customer needs. Jin, et al. (2015), Kuehl (2016), and Peng, Sun, and Revankar (2012) propose automated methods to identify engineering characteristics. The papers focus on particular parts of speech or manually identified word combinations and use clustering techniques or Latent Dirichlet Analysis to identify product attributes and levels to be considered in product development. Our methods augment the literatures in both marketing and engineering by focusing on the more-abstract deeper semantic structure of customer needs.

2.3. Deep Learning for Natural Language Processing

We draw on two literatures from natural language processing (NLP): convolutional neural networks (CNNs) and dense word and sentence representations. A CNN is a supervised prediction technique which is particularly suited to computer vision and natural language processing tasks. A CNN often contains multiple layers which transform numerical representations of sentences to create input for a final logit-based layer, which predicts the final outcome. CNNs demonstrate state-of-the-art performance with minimum tuning in such problems as relation extraction (Nguyen and Grishman 2015), named entity recognition (Chiu and Nichols 2015), and sentiment analysis (dos Santos and Gatti 2014).

Dense word and sentence embeddings are real-valued vector mappings (typically 20-300 dimensions), which are trained such that vectors for similar words (or sentences) are close in the vector space. The theory of dense embeddings is based on the Distributional Hypothesis, which states that words that appear in a similar context share semantic meaning (Harris 1954). High-quality word and sentence embeddings can be used as an input for downstream NLP applications and models (Lample, et al. 2016; Kim 2014). Somewhat unexpectedly, high-quality word embeddings capture not only semantic similarity, but also semantic relationships (Mikolov, et al. 2013b). Using the convention of bold type for vectors, then if $\boldsymbol{v}(\text{'word'})$ is the word embedding for 'word,' Mikolov et al. (2013b) demonstrate that word embeddings trained on the Google News Corpus have the following properties:

$$\boldsymbol{v}(\text{king}) - \boldsymbol{v}(\text{man}) + \boldsymbol{v}(\text{woman}) \approx \boldsymbol{v}(\text{queen})$$

$$\boldsymbol{v}(\text{walking}) - \boldsymbol{v}(\text{swimming}) + \boldsymbol{v}(\text{swam}) \approx \boldsymbol{v}(\text{walked})$$

$$\boldsymbol{v}(\text{Paris}) - \boldsymbol{v}(\text{France}) + \boldsymbol{v}(\text{Italy}) \approx \boldsymbol{v}(\text{Rome})$$

We train word embeddings using a large unlabeled corpus of online reviews. We then apply the trained word embeddings (1) to enhance the performance of the CNN and (2) to avoid repetitiveness among the sentences selected for manual review.

3. A Proposed Machine Learning Hybrid Method to Identify Customer Needs

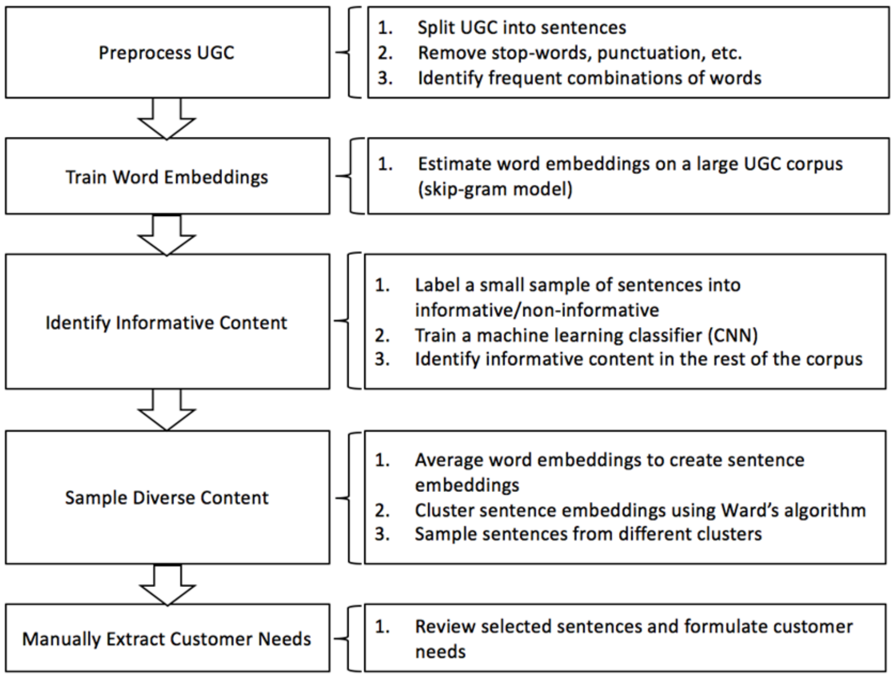
We propose a method that uses machine learning to screen UGC for sentences rich in a diverse set of customer needs. Identified sentences are then reviewed by professional analysts to formulate customer needs. Machine-human hybrids have proven effective in a broad set of applications. For example, Qian, et al. (2012) combine machine learning and human judgment to locate research when authors' names are ambiguous (e.g., there are 117 authors with the name Lei Zhang). Supervised learning identifies clusters of similar publications and human readers associate authors with the clusters. The resulting hybrid is more accurate than machine learning alone and more efficient than human classification. Colson (2016) describes Stitch Fix's machine-human hybrid in which machine learning helps create a short list of apparel from vast catalogues, then human curators make the final recommendations to consumers.

Figure 1 summarizes our approach. The proposed method consists of five stages:

1. **Preprocess UGC**. We harvest readily available UGC from either public sources or propriety company databases. We split UGC into sentences, eliminate stop-words, numbers, and punctuation, and concatenate frequent combinations of words.
2. **Train Word Embeddings**. We train word embeddings using a skip-gram model (§3.2) on preprocessed UGC sentences, and use word embeddings as an input in the following stages.
3. **Identify Informative Content**. We train and apply a CNN to filter out non-informative sentences so that the remaining corpus is rich in informative content.
4. **Sample Diverse Content**. We cluster sentence embeddings and sample sentences from different clusters to select a set of sentences likely to represent diverse customer needs.
5. **Manually Extract Customer Needs**. Professional analysts review the diverse, informative sentences to identify customer needs.

The proposed architecture achieves the same goals as voice-of-the-customer approaches in industry (§2.1). The preprocessed UGC replaces experiential interviews, the automated sampling of informative sentences is analogous to manual highlighting of sentences, and the clustering of word embeddings is analogous to manual winnowing to remove redundancy. Methods to identify a hierarchy of customer needs, if required, can be applied equally well to customer needs generated from UGC or from experiential interviews.

Figure 1 System Architecture for Identifying Customer Needs from UGC



3.1. Stage 1: Preprocessing Raw UGC

Prior experience in the manual review of UGC by professional analysts, suggests that sentences are most likely to contain customer needs and are a natural unit by which analysts process experiential interviews and UGC. We preprocess raw UGC to transform the UGC corpus into a set of sentences using an unsupervised sentence tokenizer from the natural language toolkit (Kiss and Strunk 2006). We automatically eliminate stop-words (e.g., 'the' and 'and') and non-alphanumeric symbols (e.g., question marks and apostrophes), and transform numbers into number signs and letters to lower case.

We join words that appear frequently together with the '_' character. For example, in oral care, the bigram 'Oral B' is treated as a combined word pair, 'oral_b.' We join words 'a' and 'b' into a single phrase if they appear together relatively often in the corpus. The specific criterion is:

$$\frac{count(a,b) - \delta}{count(a) \cdot count(b)} \cdot M > \tau$$

where M is the total vocabulary size. The tuning parameter, δ, prevents concatenating very infrequent words, and the tuning parameter, τ, is balanced so that the number of bigrams is not too few or too many for the corpus. Both parameters are set by judgment. For our initial test, we set $(\delta, \tau) = (5,10)$. We drop sentences that are too short (less than four words after preprocessing) or too long (more than fourteen words after preprocessing). The bounds are selected to drop approximately 10% of the shortest and 10% of the longest sentences.

As is typical in machine learning systems, our model has multiple tuning parameters. We indicate which are set by judgment and which are set by cross-validation. When we set tuning parameters by judgment, we draw on the literature for suggestions and we choose parameters likely to work in many categories. When there is sufficient data, these parameters can also be set by cross-validation.

3.2. Stage 2: Training Word Embeddings with a Skip-Gram Model

Word embeddings are the mappings of words onto a numerical vector space, which incorporate contextual information about words and serve as an input to Stages 3 and 4 (Baroni, Dinu, and Kruszewski, 2014). To account for product-category and UGC-source-specific words, we train our word embeddings on the preprocessed UGC corpus using a skip-gram model (Mikolov, et al. 2013a). The skip-gram model is a predictive model which maximizes the average log-likelihood of words appearing together in a sequence of c words. Specifically, if I is the number of words in the corpus, V is the set of all feasible words in the vocabulary, and $\boldsymbol{v}_i$ are *d*-dimensional real-vector word embeddings, we select the $\boldsymbol{v}_i$ to maximize:

$$\frac{1}{I}\sum_{i=1}^{I}\sum_{\substack{-c \le j \le c \\ j \ne 0}} log\, p\big(word_{i+j} \big| word_i\big)$$

$$p\left(word_j \middle| word_i\right) = \frac{exp\left(\boldsymbol{v}_j \boldsymbol{v}'_i\right)}{\sum_{k=1}^{|V|} exp\left(\boldsymbol{v}_k \boldsymbol{v}'_i\right)}$$

To make calculations feasible, we use ten-word negative sampling to approximate the denominator in the conditional probability function. (See Mikolov, et al. 2013b for details on negative sampling.) For our application, we use $d = 20$ and $c = 5$.

The trained word embeddings in our application capture semantic meaning in oral care. For example, the three words closest to 'toothbrush' are 'pulsonic', 'sonicare' and 'tb', with the last being a commonly-used abbreviation for toothbrush. Similarly, variations in spelling such as 'recommend', 'would_recommend', 'highly_recommend', 'reccommend', and 'recommed' are close in the vector space.

3.3. Stage 3: Identifying Informative Sentences with a Convolutional Neural Network (CNN)

Depending on the corpus, UGC can contain substantial amounts of content that does not represent customer needs. Such content includes evaluations, complaints, and non-informative lists of features such as "This product can be found at CVS." Machine learning improves the efficiency of manual review by eliminating non-informative content. For example, suppose that only 40% of the sentences are informative in the corpus, but after machine learning screening, 80% are informative. Analysts can identify customer needs much more efficiently by focusing on a sample rich in informative sentences.

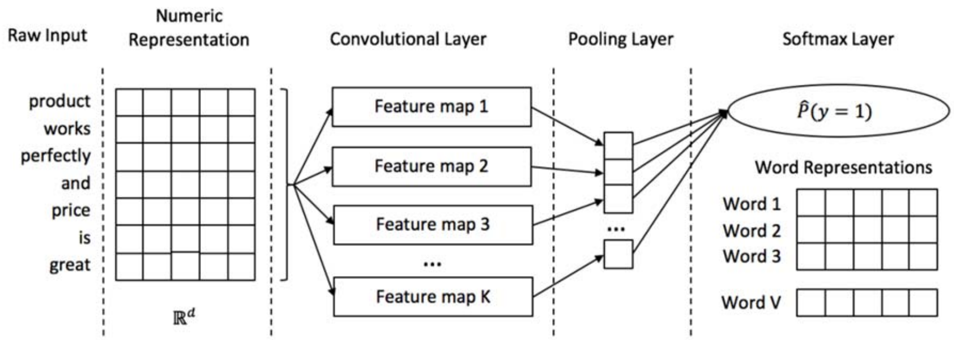
To train the machine learning classifier, some sentences must be labeled by professional analysts as informative ($y = 1$) or non-informative ($y = 0$). However, there are efficiency gains because such labeling requires substantially lower professional services costs than formulating customer needs from informative sentences. Moreover, in a small-sample study, we found that Amazon Mechanical Turk (AMT) has a potential to identify informative sentences for training data at a cost below that of using professional analysts. With further development to reduce costs and enhance accuracy, AMT might be a viable source of training data. (We address the size of the training sample empirically in §5.1.)

We use a convolutional neural network (CNN) to identify informative sentences. A major advantage of the CNN is that CNNs quantify raw input automatically and endogenously based on the training data. CNNs apply a combination of convolutional and pooling layers to word representations to generate “features,” which are then used to make a prediction. (“Features” in the CNN should not be confused with product features.) In contrast, traditional machine learning classification techniques, such as a support-vector machine or decision trees, depend critically on handcrafted features, which are the transformations of the raw data designed by researchers to improve prediction in a particular application. High-quality features require substantial human effort for each application. CNNs have been proven to provide comparable performance to traditional handcrafted-feature methods, but without substantial application-specific human effort (Kim 2014; Lei, et al. 2015).

A typical CNN consists of multiple layers. Each layer has hyperparameters, such as the number of filters and the size of the filters. We custom select these hyperparameters, and the number and type of layers, by cross-validation. Each layer also has numerical parameters, such as the parameters of the filters used in the convolutional layers. These parameters are calibrated during training. We train the CNN by selecting the parameter values that maximize the CNN’s ability to label sentences as informative vs. non-informative.

Figure 2 illustrates the architecture of the CNN in our application. We stack a convolutional layer, a pooling layer, and a softmax layer. This specification modifies Kim’s (2014) architecture for sentence classification task to account for the amount of training data available in customer-need applications.

Figure 2 Convolutional Neural Network Architecture for Sentence Classification



3.3.1. Numerical Representations of Words for Use in the CNN

For every word in the text corpus, the CNN stores a numerical representation of the word. Numerical representations of words are the real vector parameters of the model which are calibrated to improve prediction. To facilitate training of the CNN, we initialize representations with word embeddings from Stage 2. However, we allow the CNN to update the numerical representations to enhance predictive ability (Lample, et al. 2016). In our application, this flexibility enhances out-of-sample accuracy of prediction.

The CNN quantifies sentences by concatenating word embeddings. If $\boldsymbol{v}_i$ is the word embedding for the i^{th} word in the sentence, then the sentence is represented by a vector $\boldsymbol{v}$

$$\boldsymbol{v} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{R}^{d \times n}$$

where n is the number of words in the sentence and $d = 20$ is the dimensionality of the word embeddings.

3.3.2. Convolutional Layer

Convolutional layers create multiple feature maps by applying convolutional operations with varying filters to the sentence representation. A filter is a real-valued vector, $\boldsymbol{w}_t \in \mathbb{R}^{d \times h_t}$, where h_t is a size of the filter. Filters are applied to different parts of the vector $\boldsymbol{v}$ to create feature maps ($\boldsymbol{c}^t$):

$$\boldsymbol{c}^t = [c_1^t, \dots, c_{n-h_t+1}^t]$$

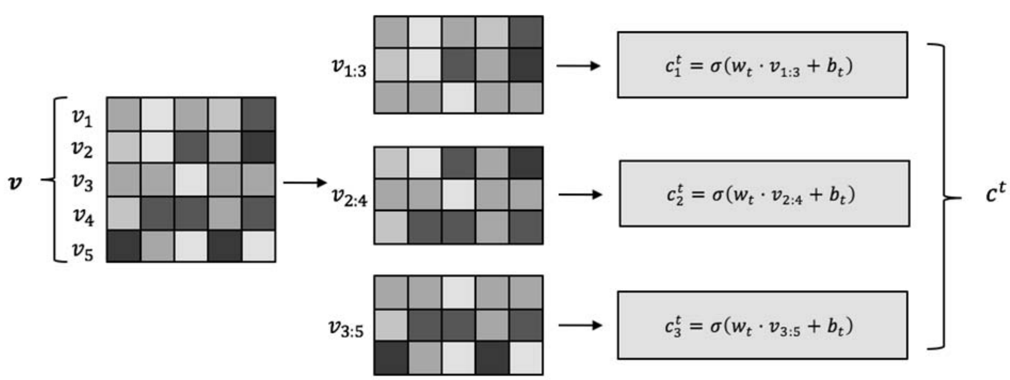
$$c_i^t = \sigma\left(\boldsymbol{w}_t \cdot \boldsymbol{v}_{i:i+h_t-1} + b_t\right)$$

where t indexes the feature maps, $\sigma(\cdot)$ is a non-linear activation function where $\sigma(x) = \max(0, x)$, $b_t \in \mathbb{R}$ is an intercept, and $\boldsymbol{v}_{i:i+h_t-1}$ is a concatenation of representations of words i to $i + h_t - 1$ in the sentence:

$$\boldsymbol{v}_{i:i+h_t-1} = [\boldsymbol{v}_i, \dots, \boldsymbol{v}_{i+h_t-1}]$$

We consider filters of the size $h_t \in \{3, 4, 5\}$, and use three filters of each size. The number of filters and their size are selected to maximize prediction on the validation set. The numerical values for filters, $\boldsymbol{w}_t$, and intercepts, b_t, are calibrated when the CNN is trained. As an illustration, Figure 3 shows how a feature map is generated with a filter of size, $h_t = 3$. On the left is a sentence, $\boldsymbol{v}$, consisting of five words. Each word is a 20-dimenional vector (only 5 dimensions are shown). Sentence $\boldsymbol{v}$ is split into triplets of words as shown in the middle. Representations of word triplets are then transformed to the real-valued c_i^t's in the next column. The t^{th} feature map, $\boldsymbol{c}^t$, is the vector of these values. Processing sentences in this way allows the CNN to interpret words that are next to one another in a sentence together.

Figure 3 Example Feature Map, $\boldsymbol{c}^t$ Generated with a Filter, $\boldsymbol{w}_t$, of Size $\boldsymbol{h}_t = 3$.



3.3.3. Pooling Layer

The pooling layer transforms feature maps into shorter vectors. The role of the pooling layer is to reduce dimensionality of the output of the convolutional layer to be used in the next layer. Pooling to the k^{th} largest features or simply using the largest feature has proven effective in NLP applications (Collobert, et al. 2011). We selected $k = 1$ with cross-validation. The output of the pooling layer is a vector, $\boldsymbol{z}$, that summarizes the results of pooling operators applied to the feature maps:

$$z_t = max[c_1^t, \dots, c_{n-h_t+1}^t]$$

$$\boldsymbol{z} = [z_1, z_2, \dots, z_9]$$

The vector, $\boldsymbol{z} \in \mathbb{R}^9$, is now an efficient numerical representation of the sentence and can be used to classify the sentence as either informative or not informative.

3.3.4. Softmax Layer

The final layer of the CNN is called the softmax layer. The softmax layer transforms the output of the pooling layers, $\boldsymbol{z}$, into a probabilistic prediction of whether the sentence is informative or not informative. Marketing researchers will recognize the softmax layer as a binary logit model which uses the $\boldsymbol{z}$ vector as explanatory variables. The estimate of the probability that the sentence is informative, $P(y = 1|\boldsymbol{z})$, is given by:

$$\hat{P}(y = 1|\boldsymbol{z}) = \frac{1}{1 + e^{-\boldsymbol{\theta z}}}$$

The parameters of the logit model, $\boldsymbol{\theta}$, are determined when the CNN is trained. In our application, we declare a sentence to be informative if $P(y = 1|\boldsymbol{z}) > 0.5$, although other criteria could be used and tuned to a target tradeoff between false positives and false negatives.

3.3.5. Calibration of the Parameters of the CNN

For our application we calibrate the nine filters, $\boldsymbol{w}_t \in \mathbb{R}^{d \times h_t}$, and the nine intercepts, b_t, in the convolutional layer, and the vector $\boldsymbol{\theta}$ in the softmax layer. In addition, we fine tune the word embeddings, $\boldsymbol{v_i}$, to enhance the ability of the CNN's predictions (e.g., Kim 2014). We calibrate all parameters simultaneously by minimizing the cross-entropy error on the training set of professionally labeled sentences ($\boldsymbol{w}$ is a concatenation of the $\boldsymbol{w}_t$'s):

$$\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{b}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{v}} = argmax_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta},\boldsymbol{v}} L(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{v})$$

$$L(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{v}) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \, log \, \hat{y}_n + (1 - y_n) \, log(1 - \hat{y}_n)]$$

N is the size of the training set, y_n are the manually assigned labels, and $\hat{y}_n$ are the predictions of the CNN. We solved the optimization problem iteratively with the RMSProp optimizer on mini-batches of size 32 and a drop rate of 0.3. Optimization terminated when the cross-entropy error on the validation set did not decrease over five consecutive iterations. See Tieleman and Hinton (2012) for details and definitions of terms such as "drop rate."

3.3.6. Evaluating the Performance of the CNN

We evaluate the quality of the CNN classifier using two common criteria: precision and recall (Lee and Bradlow, 2011):

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

where TP, FP and FN correspond to a number of identified true positives (*TP*), false positives (*FP*), and false negatives (*FN*) in the holdout set. Precision is interpreted as the share of informative sentences among the sentences identified as informative and recall is the share of informative sentences correctly identified. In machine learning, it is common to combine these measures into an F_1 score as the primary measure of interest (Wilson et al., 2005).

$$F_1 = \frac{precision \cdot recall}{\frac{1}{2}(precision + recall)}$$

3.4. Stage 4: Clustering Sentence Embeddings and Sampling to Reduce Redundancy

UGC is repetitive and often focuses on a small set of customer needs. Manual review of repetitive content is inefficient. Moreover, repetitiveness makes the manual review onerous and boring for professional analysts, causing analysts to miss excitement customer needs that are mentioned rarely. If the analysts miss excitement customer needs, then the firm misses valuable new product opportunities and/or strategic positionings. To avoid repetitiveness, we seek to "span the set" of customer needs. We construct sentence embeddings which encode semantic relationships between sentences, and use sentence embeddings to reduce redundancy by sampling content for manual review from maximally different parts of the space of sentences.

Researchers often create sentence embeddings by taking a simple average of word embeddings corresponding to the words in the sentence (Iyyer et al., 2015), explicitly modeling semantic and syntactic structure of the sentences with neural methods (Tai, Socher and Manning 2015), or training sentence embeddings together with word embeddings (Le and Mikolov, 2014). Because averaging demonstrates similar performance to other methods and is both scalable and transferable (Iyyer et al., 2015), we use averaging in our application.

Being the average of word embeddings, sentence embeddings represent semantic similarity among sentences. Using this property, we group sentences into clusters. We choose Ward's hierarchical clustering method because it is commonly used in VOC studies (Griffin and Hauser 1993), and other areas of marketing research (Dolnicar 2003). To identify *Y* sentences for professional analysts to review, we sample one sentence randomly from each of *Y* clusters.

3.5. Stage 5: Manually Extracting Customer Needs

To achieve high precision in formulating abstract customer needs, the final extraction of customer needs is best done by trained analysts. We evaluate in §5 whether manual extraction becomes more efficient using informative, diverse sentences identified with the CNN and sentence-embedding clusters.

4. Evaluation of UGC's Potential in the Oral-Care Product Category

We use empirical data to examine two questions. (§4) Does UGC contain sufficient raw material from which to identify a broad set of customer needs? And (§5) Do each of the machine-learning steps enhance efficiency? We address both questions with a custom dataset in the oral-care category. We selected oral care because oral-care customer needs are sufficiently varied, but not so numerous as to overcomplicate comparisons. As a proof-of-concept test, our analyses establish a key example. We discuss applicability to other categories in §6.

4.1. Baseline Comparison: Experiential Interviews in Oral Care

We obtained a detailed set of customer needs from an oral care voice-of-the-customer (VOC) analysis that was undertaken by a professional market research consulting firm. The firm has almost thirty years of VOC experience spanning hundreds of successful product-development applications across a wide-variety of industries. The oral-care VOC provided valuable insight to the client and led to successful new products. The VOC was based on standard methods: experiential interviews, with sentences highlighted by experienced analysts aided by the firm's proprietary software. After winnowing, customer needs were clustered by a customer-based affinity group. The output is 86 customer needs aggregated into six primary and 22 secondary need groups. An appendix lists the primary and secondary need groups and provides an example of a tertiary need from each secondary-need group. Examples of customer needs include: "Oral care products that do not create any odd sensations in my mouth while using them (e.g. tingling, burning, etc.)" or "My teeth feel smooth when I glide my tongue over them." Such customer needs are more than their component words; they describe a desired outcome in the language that the customer uses to describe the desired outcome.

Professional analysts estimate that the professional-service time necessary to review, highlight, and winnow customer needs from experiential-interview transcripts is slightly more than the professional services required to review 8,000 UGC sentences to identify customer needs. Additionally, the professional services required to review, highlight, and winnow customer needs is about 40%-55% of the professional services required to schedule and interview customers. At this rate, professional analysts could review approximately 22,000 to 28,000 UGC sentences using the methods and professional services costs involved in a typical VOC study.

4.2. Fully-Coded UGC Data from the Oral-Care Category

To compare UGC to experiential interviews and evaluate a proposed machine learning method, we needed a fully-coded sample of a UGC corpus. In particular, we needed to know and classify every customer need in every sentence in the UGC sample. We received in-kind support from professional analysts to generate a custom data set to evaluate UGC and the machine-learning efficiencies. The in-kind support was approximately that which the firm would have allocated to a typical VOC study—approximately $50,000 of professional services costs.

From the 115,099 oral-care reviews on Amazon spanning the period from 1996 to 2014, we randomly sampled 12,000 sentences split into an initial set of 8,000 sentences and a second set of 4,000 sentences (McAuley, et. al. 2015). To maintain a common level of training and experience for reviewing UGC and experiential interview transcripts, the sentences were reviewed by three experienced analysts from the same firm that provided the interview-based VOC.

We chose 8,000 sentences for our primary evaluation because the professional services costs to review 8,000 sentences are comparable, albeit slightly less than, the effort to review a typical set of experiential-interview transcripts. For these sentences, the analysts fully coded every sentence to determine whether it contained a customer need and, if so, whether the customer need could be mapped to a customer need identified by the VOC, or whether the customer need was a newly identified customer need.

We were also able to persuade the analysts to examine an additional 4,000 sentences to focus on any customer needs that were identified by the traditional VOC, but not identified from the UGC. This second dataset enables us to address whether there exist customer needs that are not in UGC per se, or whether the customer needs are sufficiently rare that more than 8,000 sentences are required to identify them. This additional benchmark allows for more professional services, but well under those necessary to interview and process the transcripts. Fully-coding the 22-28,000 sentences, that would be the professional-services-equivalent to interviewing and processing transcripts, with highly-paid and sought-after analysts was not feasible.

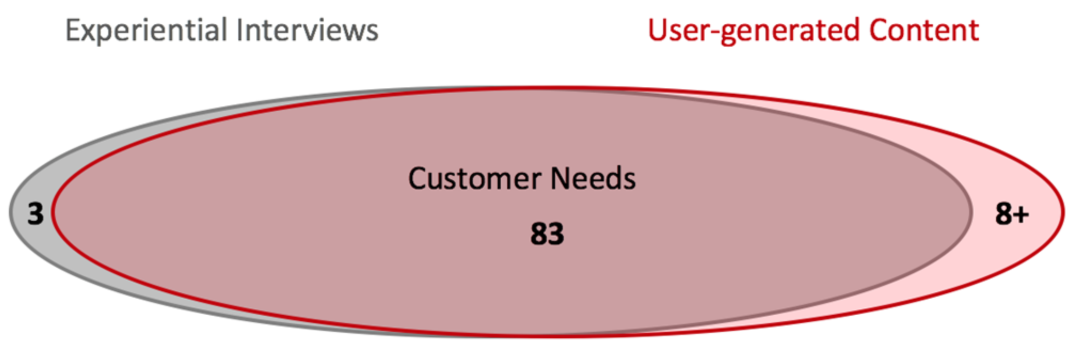
4.3. Descriptive Statistics and Comparisons

Using Amazon reviews, 52% of the 8,000 sentences contained at least one customer need and 9.2% of the sentences contained two or more customer needs. However, the corpus was highly repetitive; 10% of the most frequent customer needs were articulated in 54% of the informative sentences. On the other hand, 17 customer needs were articulated no more than 5 times in the corpus of 8,000 sentences.

We consider first the 8,000 sentences—in this scenario analysts allocate at most as much time coding UGC as they would have allocated to review experiential interview transcripts. This section addresses the potential of the UGC corpus, hence, for this section, we do not yet exploit machine-learning efficiencies. From the 8,000 sentences, analysts identified 74 of the 86 tertiary experiential-interview-based customer needs, but also identified an additional 8 needs.

We now consider the set of 4,000 sentences as a supplement to the fully-coded 8,000 sentences—in this scenario analysts still allocate substantially less time than they would to interview customers and review transcripts. From the second set of 4,000 sentences, the analysts were able to identify 9 of 12 missing customer needs. With 12,000 sentences, that brings the total to 83 of the 86 experiential-interview-based customer needs and 91 of the 94 total needs (97%). The analysts did not try to identify for customer needs other than the 12 missing needs. Had we had the resources to do so, we would likely have increased the number of UGC-based incremental customer needs. Overall, analysts identified (at least) 91 customer needs from UGC and 86 customer needs from experiential interviews. These results are summarized in Figure 4. At least in oral care, analyzing UGC has the potential to identify more customer needs at a lower overall cost of professional services, even without machine-learning efficiencies.

Figure 4. Comparison of Customer Needs Obtained from Experiential Interviews with Customer Needs Obtained from an Exhaustive Review of a UGC Sample



4.4. Prioritization of Customer Needs

To address whether the eight incremental UGC customer needs and/or the three incremental experiential-interview customer needs were important, we conducted a prioritization survey. We randomly selected 197 customers from a professional panel (PureSpectrum), screened for interest in oral care, and asked customers to rate the importance of each tertiary customer need on a 0-to-100 scale. Customers also rated whether they felt that their current oral-care products performed well on these customer needs on a 0-to-10 scale. Such measures are used commonly in VOC studies and have proven to provide valuable insights for product development. (Review citations in §2.1.)

Table 1 summarizes the survey results. On average, the customer needs identified in both the interviews and UGC are the most important customer needs. Those that are unique to UGC or unique to experiential interviews are of lower importance and performance. We gain further insight by categorizing the customer needs into quadrants via median splits. The vast majority of high-importance-low-performance customer needs are identified by both data sources. Such customer needs provide insight for product improvement.

Table 1. Importance and Performance Scores for Customer Needs Identified from UGC and from Experiential Interviews (Imp = Importance, Per = Performance)

				Quadrant (median splits)			
Source of Customer Need	**Count**	**Average Imp**	**Average Per**	**High Imp High Per**	**High Imp Low Per**	**Low Imp High Per**	**Low Imp Low Per**
Interviews ∩ 8,000 UGC [a]	74	65.5	7.85	29	11	11	23
Interviews ∩ 4,000 UGC [b]	9	63.9	7.97	6	0	0	3
UGC only	8	50.3	7.12	0	0	1	7
Interviews only	3	52.8	7.47	0	1	0	2

[a] Based on the first 8,000 UGC sentences that were fully-coded
[b] Based on the second 4,000 UGC sentences that were coded to test for interview-identified customer needs

Focusing on highly important customer needs is tempting, but we cannot ignore low-importance customer needs. In new product development, identifying hidden opportunities for innovation often leads to successful new products. Customers often evaluate needs below the medians on importance and performance when they anticipate that no current product fulfills those customer needs (e.g., Corrigan 2013). If the new product satisfies the customer need, customers reconsider its importance, and the innovator gains a valuable strategic advantage. Thus, we define low-importance–low-performance customer needs as hidden opportunities. By this criterion, the UGC-unique customer needs identify 20% of the hidden opportunities and the interview-unique needs identify 8% of the hidden opportunities. For example, two UGC-unique hidden opportunities are “An oral-care product that does not affect my sense of taste,” and “An oral care product that is quiet.” An interview-based hidden opportunity is “Oral care tools that can easily be used by left-handed people.”

In summary, UGC identifies the vast majority of customer needs (97%), the vast majority of opportunities for product improvement (92%), and the vast majority of hidden opportunities (92%).

UGC-unique needs identify at least seven hidden opportunities while interview-only needs identify two hidden opportunities.

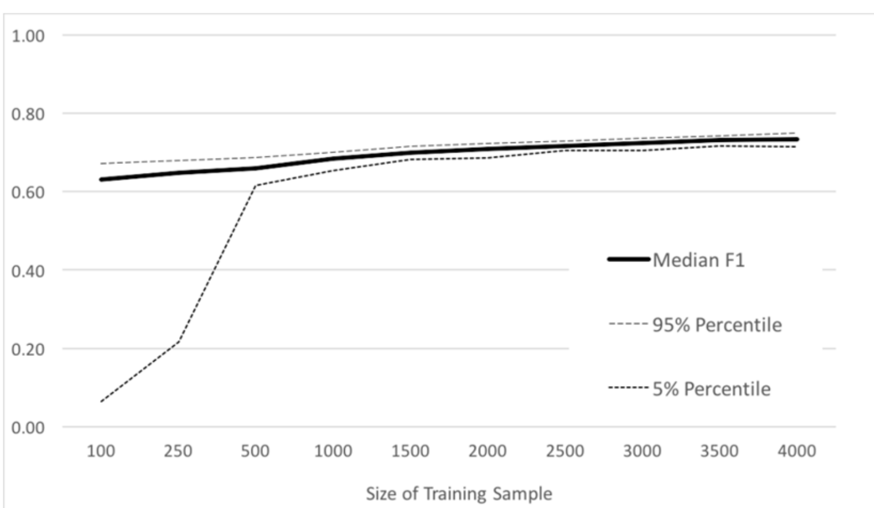
5. Oral Care: Evaluation of Machine-Human Hybrid Method

5.1. CNN to Eliminate Non-Informative Sentences

There is a tradeoff to be made when training a CNN. With a larger training sample, the CNN is better at identifying informative content, but there is an opportunity cost to using analysts to classify informative sentences. Fortunately, labeling sentences as informative or not is faster and easier than identifying abstract needs from sentences. The ratio of time spent for identifying informative sentences vs. formulating customer needs is approximately 20%. Furthermore, as described earlier, exploratory research suggests that Amazon Mechanical Turk might be developed as a lower-cost way to obtain a training sample.

Figure 5 plots the F_1-score of the CNN as a function of the size of the training sample. We conduct 100 iterations where we randomly draw a training set, train the CNN with the architecture described in §3.3, and measure performance on the test set. Figure 5 suggests that performance of the CNN stabilizes after 500 training sentences, with some slight improvement after 500 training sentences. We plot precision and recall as a function of the size of the training sample in Appendix A2.

Figure 5. $\boldsymbol{F_1}$ **score as a Function of the Size of the Training Sample**



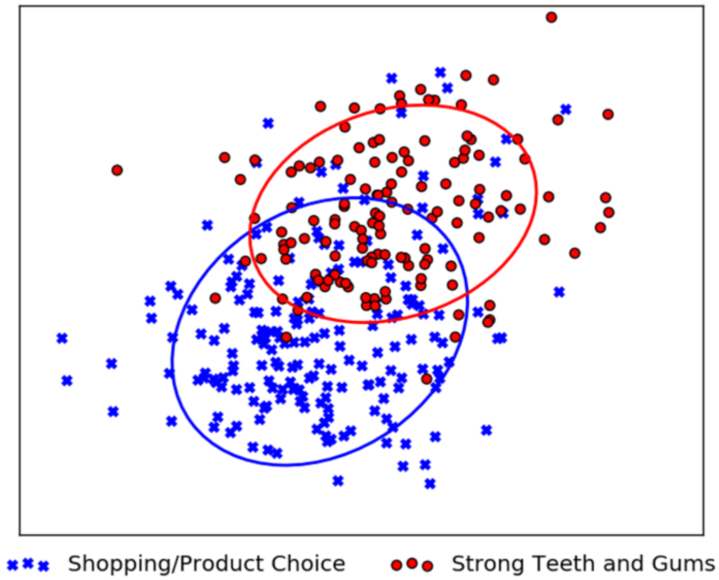
To be effective, the CNN should identify informative sentences that contain customer needs that were not in the sentences in the training set. In our case, the CNN identifies sentences that contain customer needs that were "new," that is, not in the sentences used to train the CNN. For example, the CNN identified "new" customer needs in the hold-out set such as: "using as few products as possible and still having an effective oral care routine." In the prioritization survey, this customer need was in the upper quartile of importance. There were many other examples.

5.2. Clustering Sentence Embeddings to Reduce Redundancy

To visualize whether or not sentence-embedding clusters reduce redundancy, we use principle components analysis to project the sentence embeddings onto two dimensions. Information is lost when we project from 20 dimensions to two dimensions, but the two-dimensional plot enables us to visualize whether or not customer needs were separated by the clusters. As an illustration, Figure 6 reports the two-dimensional projection for two primary needs. The axes correspond to the first two principal components.

The red dots are the projections of sentence embeddings that were coded (by analysts) as belonging to the primary customer need: "strong teeth and gums." The blue crosses are sentence embeddings that were coded as "shopping/product choice." (Review Table A1 in the appendix.) The ovals represent the smallest ellipses inscribing 90% of the corresponding set. Figure 6 suggests that, while not perfect, the clusters of sentence embeddings achieved separation among customer needs and, hence, are likely to reduce redundancy and enable analysts to identify a diverse set of customer needs. Figure 6 is typical; we get similar results for other customer needs.

Figure 6. Projections of 20-Dimensional Embeddings of Sentences onto Two Dimensions (PCA). Dots and Crosses Indicate Analyst-Coded Primary Customer Needs.



5.3. Gains in Efficiency Due to Machine Learning

We seek to determine whether the machine-learning methods improve efficiency relative to current practice. Efficiency is important because the reduced time and costs enable more firms to use advanced VOC methods to identify new product opportunities. In our approach, machine learning helps to identify content for review by professional analysts. Currently, the industry-standard practice is to review sentences selected randomly from UGC. Because we have a database of 8,000 fully-coded sentences (6,700 after preprocessing), we compute the expected number of customer needs that would have been identified by various methods for various numbers of sampled UGC sentences. In this evaluation, we want the CNN to be well-trained so that we evaluate the full potential of the method. We choose a training sample of 3,700 sentences to assure that F_1, recall, and precision have stabilized. At this level of training, the CNN achieves a precision of 71.6%, a recall of 70.8%, and an F_1 of 71.2%. After training, there are 3,000 sentences available for evaluation of content selection approaches.

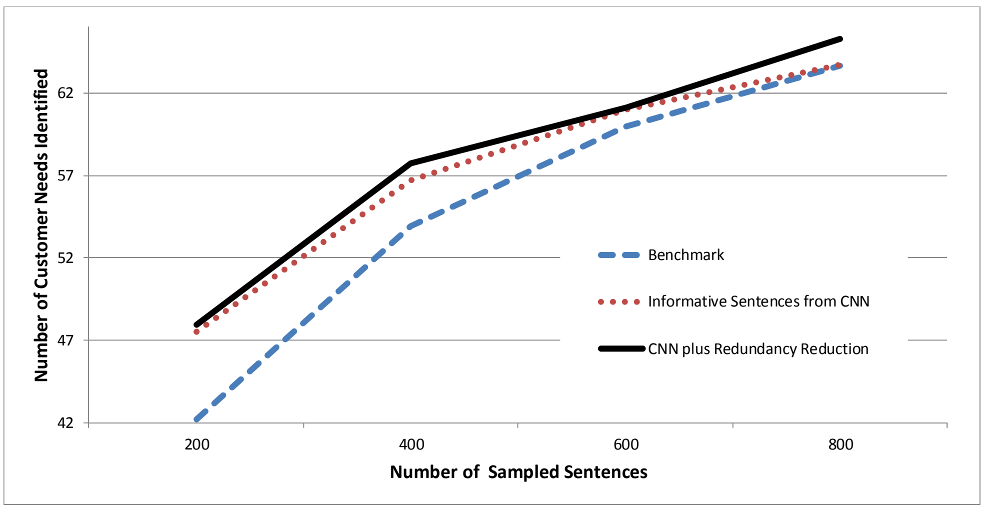
To evaluate machine-learning efficiencies, we consider scenarios where we hold professional services costs roughly equal to that required to review *Y* sentences to identify customer needs. We evaluate content selection approaches at multiple values of *Y*. The baseline method for selecting sentences for review is current practice—a random draw from the corpus. The second method uses the CNN to identify informative sentences, and then randomly samples informative sentences for review. The third method uses the sentence-clusters to reduce redundancy among sentences identified as informative by the CNN. For each method, and for each value of *Y*, we draw sufficient sentences from the corpus to provide *Y* sentences for review. We count the unique needs identified in the *Y* sentences and repeat the process 10,000 times. From 3,000 unclassified sentences, the largest possible value of *Y* is the 1,480 CNN-identified informative sentences.

While it is tempting to consider *Y* in the range from 0 to 1,480, it is not informative to test the potential of word-embedding clusters at $Y = 1{,}480$. At $Y = 1{,}480$, there would be 1,480 clusters—the same number as if we sampled all available sentences. To minimize this saturation effect on the oral-care corpus, we consider $Y = \{200, 400, 600, 800\}$ for the plot in Figure 7.

The blue dashed line in Figure 7 reports benchmark performance. The CNN improves efficiency as indicated by the red dotted line. Using the CNN and clustering sentence embeddings increases efficiency further as indicated by the solid black line. Over the range of *Y*, there are gains due to using the CNN to eliminate non-informative sentences and additional gains due to using sentence embeddings to reduce redundancy within the corpus. At $Y = 800$ we see the expected saturation effects of using the CNN. This is exacerbated by the long-tail nature of customer needs—many customer needs have relatively few mentions in the data. Had the marketing research firm donated additional in-kind professional services to code more sentences and/or had we allocated fewer sentences for training, saturation would have occurred at a higher *Y*. Thus, efficiency gains are best interpreted for $Y < 800$.

We also interpret Figure 7 horizontally. The benchmark requires, on average, 832 sentences to identify 64.2 customer needs. If we prescreen with machine learning to select non-redundant informative sentences, analysts can identify the same number of customer needs from approximately 700 sentences—84% of the sentences required by the baseline. The efficiencies are even greater at 200 sentences (71%) and 400 sentences (78%). At professional billing rates across many categories, this represents substantial time and cost savings and could expand the use of VOC methods in product development. VOC customer-need identification methods has been optimized over almost thirty years of continuous improvement; we expect the machine-learning methods, themselves, to be subject to continuous improvement as they are applied in the field.

Figure 7. Efficiencies among Various Methods to Select UGC Sentences for Review



6. Discussion, Summary, and Future Research

We addressed two questions: (1) Can UGC be used to identify abstract customer needs? And (2) can machine learning enhance the process? The answer to both questions is yes. UGC is at least a comparable source of customer needs to experiential interviews—likely a better source. The proposed machine-learning architecture successfully eliminates non-informative content and reduces redundancy. In our initial test, gains are 16-29%, but such gains are likely to increase with more research.

Answering these questions is significant. Every year thousands of firms rely on voice-of-the-customer analyses to identify new opportunities for product development, to develop strategic positioning strategies, and to select attributes for conjoint analysis. Typically, VOC studies, while valuable, are expensive and time-consuming. Time-to-market savings, such as those made possible with machine learning applied to UGC, are extremely important to product development. In addition, UGC seems to contain customer needs not identified in experiential interviews. New customer needs mean new opportunities for product development and/or new strategic positioning.

While we are enthusiastic about UGC, we recognize that UGC is not a panacea. UGC is readily available for oral care, but UGC might not be available for every product category. For example, consider specialized medical devices or specialized equipment for oil exploration. The number of customers for such products is small and such customers may not post reviews, blog, or tweets. On the other hand, UGC is extensive for complex products such as automobiles or cellular phones. Machine-learning efficiencies in such categories may be necessary to make the review of UGC feasible.

Although our research focuses on developing and testing new methods, we are beginning to affect industry. The market research firm is currently testing our approach for small appliances. We are also aware of other firms experimenting with our methods—one application to a complex electronic product and another application in frozen-foods.

Further research will enhance our ability to identify abstract customer needs with UGC. For example,

- Deep neural networks and sentence embeddings are active areas of research in the NLP community. We expect the performance of the proposed architecture to improve significantly with new developments in machine learning.
- UGC is updated continuously. Firms might develop procedures to monitor UGC continuously. Sentence embeddings can be particularly valuable. For example, firms might concentrate on customer needs that are distant from established needs in the 20-dimenional vector space.
- Future developments might automate the final step, or at least enhance the ability of analysts to

abstract customer needs from informative, non-redundant content.

- Other forms of UGC, such as blogs and Twitter feeds may be examined for customer needs. We expect blogs and Twitter feeds to contain more non-informative content, which makes machine learning filtering even more valuable.
- Field experiments might assess whether, and to what degree, abstract customer needs provide more insights for product development than insights obtained from lists of words.
- Amazon Mechanical Turk is a promising means to replace analysts for labeling training sentences, but further research is warranted.

References

Akao Y (2004) *Quality Function Deployment (QFD): Integrating customer requirements into product design*, (New York, NY: Productivity Press).

Archak N, Ghose A, Ipeirotis PG (2016) Deriving the pricing power of product features by mining consumer reviews, *Management Science*. 57(8): 1485-1509.

Alam I., Perry C. (2002) A customer-oriented new service development process. *Journal of services Marketing*. 16(6):515-534.

Baroni M, Dinu G, Kruszewski G (2014) Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD. 238-247.

Brown SL, Eisenhardt KM (1995) Product development: Past research, present findings, and future directions. *The Academy of Management Review*. 20(2):343-378.

Büschken, J, Allenby GM (2016) Sentence-based text analysis for consumer reviews. *Marketing Science*. 35(6):953-975.

Chan L-K, Wu M-L (2002) Quality Function Deployment: A literature Review. *European Journal of Operational Research*. 143:463-497.

Chiu JP, Nichols E (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4:357–370.

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Pavel K (2011) Natural language processing (almost) from scratch. *Journal of Machine Learning Research*. 12:2493-2537.

Colson E (2016) Human machine algorithms: Interview with Eric Colson. http://blog. fastforwardlabs.com/2016/05/25/human-machine-algorithms-interview-with-eric.html.

Corrigan KD (2013) Wise choice: The six most common product development pitfalls and how to avoid them. *Marketing News*. (September) 39-44.

Dolnicar S (2003) Using cluster analysis for market segmentation – typical misconceptions, established methodological weaknesses and some recommendation for improvement. *Australasian Journal of Market Research*. 11(2):5-12.

dos Santos CN, Gatti M (2014) Deep convolutional neural networks for sentiment analysis of short texts. *Proceedings the 25th International Conference on Computational Linguistics: Technical Papers.* Dublin, Ireland, 69–78,

Fader PS, Winer RS (2012) Introduction to the special issues on the emergence can impact of user-generated content. *Marketing Science*. 31(3):369-371.

Goffin K, Varnes CJ, van der Hoven C, Koners U (2012) Beyond the voice of the customer: Ethnographic market research. *Research Technology Management*. 55(4):45-53.

Green PE, Srinivasan V (1978) Conjoint analysis in consumer research: issues and outlook. *Journal of Consumer Research* 5(2):103-123.

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

Griffin A, Price RL, Maloney MM, Vojak BA, Sim EW (2009) Voices from the field: how exceptional electronic industrial innovators innovate. *Journal of Product Innovation Management*. 26:222-240.

Hauser JR, Clausing D (1988) The house of quality. *Harvard Business Review*. 66(3):63-73.

Herrmann A, Huber F, Braunstein C (2000) Market-driven product and service design: Bridging the gap between customer needs, quality management, and customer satisfaction. *International Journal of Production Economics*. 66(1):77-96.

Iyyer M, Manjunatha V, Boyd-Graber J, Daumé III H. (2015) Deep unordered composition rivals syntactic methods for text classification. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

Processing, Beijing, China. 1:1681-1691.

Jiao J, Chen CH (2006) Customer requirement management in product development: a review of research issues. *Concurrent Engineering: Research and Applications*. 14(3):173-185.

Jin J, Hi P, Liu Y, and Lim SCJ (2015) Translating online customer opinions into engineering characteristics in QFD: A probabilistic language analysis approach. *Engineering Applications of Artificial Intelligence.* 41:115-127.

Kano N, Seraku N, Takahashi F, Tsuji S (1984) Attractive quality and must-be quality. *The Japanese Society for Quality Control* 14(2):39-48.

Kao Group (2016). http://www.company-histories.com/Kao-Corporation-Company-History.html.

Kaulio MA (1998) Customer, consumer and user involvement in product development: A framework and a review of selected methods. *Total Quality Management.* 9(1):141-149.

Kim Y (2014) Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kim DS, Bailey RA, Hardt N, Allenby A (2017) Benefit-based conjoint analysis. *Marketing Science*, 36(1):54-69.

Kiss T, Strunk J (2006) Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485-525.

Krishnan V, Ulrich KT (2001) Product development decisions: A review of the literature. *Management Science.* 47(1):1-21.

Kuehl N (2016) Needmining: Towards analytical support for service design. *International Exploring Services Science*. 247:187-200.

Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C (2016) Neural architectures for named entity recognition. *Proceedings of 2016 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA:260-270.

Le QV, Mikolov T (2014) Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 32, 1188-1196.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *Journal of Marketing Research.* 48(5), 881-894.

Lei T, Barzilay R, Jaakkola T (2015) Molding CNNs for text: non-linear, non-consecutive convolutions. *Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal. 1565–1575.

Matzler K, Hinterhuber HH (1998) How to make product development projects more successful by integrating Kano's model of customer satisfaction into quality function deployment. *Technovation*. 18(1):25-38.

McAuley J, Pandey R, Leskovec J (2015) Inferring networks of substitutable and complementary products. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 785-794.

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. arXiv:1301.3781v3 [cs.CL]m Sept 7,1301.3781.

Mikolov T, Sutskever I, Chen K., Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. 26, 3111–3119.

Mikulić J, Prebežac D (2011). A critical review of techniques for classifying quality attributes in the Kano model. *Managing Service Quality.* 21(1):46-66.

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

Nguyen TH, Grishman R (2015) Relation extraction: Perspective from convolutional neural networks.

Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, CO. 39-48.

Orme BK (2006) *Getting started with conjoint analysis: Strategies for product design and pricing research, 2E.* (Madison WI: Research Publishers LLC).

Park CW, Jaworski BJ, MacInnis DJ (1986) Strategic brand concept-image management. *Journal of Marketing.* 50:135-145.

Peng W, Sun T, Revankar S (2012). Mining the `voice of the customer' for business prioritization. *ACM Transactions on Intelligent Systems and Technology.* 3 (2), 38:1-38-17.

Qian Y-N, Hu Y, Cui J, Nie Z (2001) Combining machine learning and human judgment in author disambiguation. *Proceedings of the 20th ACM Conference on Information and Knowledge Management*. Glasgow, United Kingdom.

Schaffhausen CR, Kowalewski TM (2015). Large-scale needfinding methods of increasing user-generated needs from large populations. *Journal of Mechanical Design.* 137(7): 071403.

Schaffhausen CR, Kowalewski TM (2016) Assessing quality of unmet user needs: effects of need statement characteristics. *Design Studies.* 44:1-27.

Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg PA. 1631-1642.

Stone RB, Kurtadikar R, Villanueva N, Arnold CB (2008) A customer needs motivated conceptual design methodology for product portfolio planning. *Journal of Engineering Design.*19(6):489-514.

Sullivan LP (1986) Quality function deployment. *Quality Progress.* 19(6), 39-50.

Tai KS, Socher R, Manning CD (2015) Improved semantic representations from tree-structured long short-term memory networks. *Proceedings of the 53st Annual Meeting on Association for*

Computational Linguistics. Stroudsburg, PA. 1556-1566.

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using Latent Dirichlet Allocation. *Journal of Marketing Research*. 51:463-479.

Tieleman T, Hinton G (2012) Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4.

Ulrich KT, Eppinger SD (2016) *Product design and development*, 6E. (New York, NY: McGraw-Hill).

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference On Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver BC. 347-354.

Wu H-H, Shich JI (2010) Applying repertory grids technique for knowledge elicitation in Quality Function Deployment. *Quality and Quantity.* 44:1139-1149.

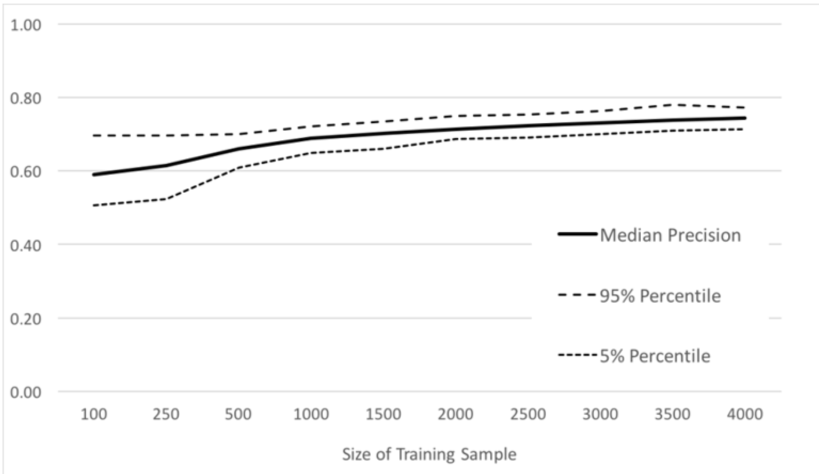
Zahay D, Griffin A, Fredericks E (2004) Sources, uses, and forms of data in the new product development process. *Industrial Marketing Management*. 33:657-666.

Appendix

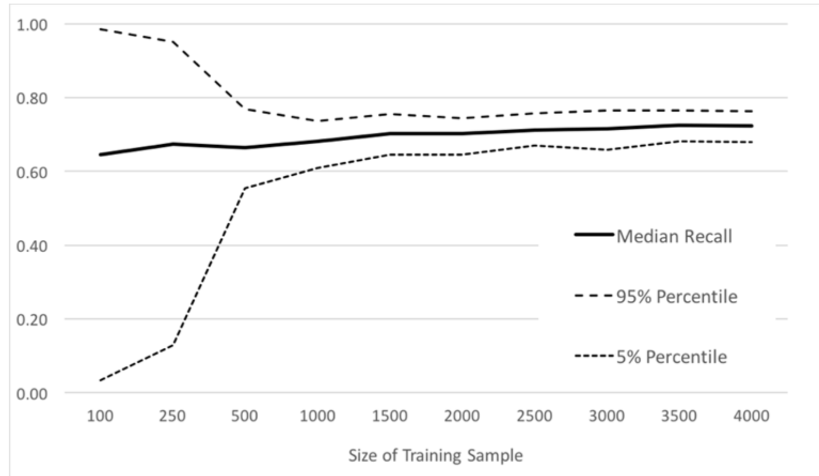
Table A1. **Voice of the Customer for Oral Care as Obtained from Experiential Interviews (22 examples of the 86 tertiary customer needs are shown—one for each secondary group. A full list of tertiary customer needs is available from the authors.)**

Primary Group	Secondary Group	#Needs	Examples of Tertiary Customer Needs (22 of 86 shown)
Feel Clean And Fresh (Sensory)	Clean Feeling in My Mouth	4	My mouth feels clean
	Fresh Breath All Day Long	4	I wake up without feeling like I have morning breath
	Pleasant Taste and Texture	3	Oral care liquids, gels, pastes, etc. are smooth (not gritty or chalky)
Strong Teeth And Gums	Prevent Gingivitis	5	Oral care products and procedures that minimize gum bleeding
	Able to Protect My Teeth	5	Oral care products and procedures that prevent cavities
	Whiter Teeth	4	Can avoid discoloration of my teeth
Product Efficacy	Effectively Clean Hard to Reach Areas	3	Able to easily get all particles, even the tiniest, out from between my teeth
	Gentle Oral Care Products	4	Oral care items are gentle and don't hurt my mouth
	Oral Care Products that Last	3	It's clear when I need to replace an oral care product (e.g. toothbrush, floss)
	Tools are Easy to Maneuver and Manipulate	6	Easy to grasp any oral care tool—it won't slip out of my hand
Knowledge And Confidence	Knowledge of Proper Techniques	5	I know the right amount of time to spend on each step of my oral care routine
	Long Term Oral Care Health	4	I am aware of the best oral care routine for me
	Motivation for Good Check-Ups	4	I want to be motivated to be more involved with my oral care
	Able to Differentiate Products	3	I know which products to use for any oral care issue I'm trying to address
Convenience	Efficient Oral Care Routine (Effective, Hassle-Free and Quick)	7	Oral care tasks do not require much physical effort
	Oral Care "Away From the Bathroom"	5	The oral care items I carry around are easy to keep clean
Shopping / Product Choice	Faith in the Products	5	Brands of oral care products that are well known and reliable
	Provides a Good Deal	2	I know I'm getting the lowest price for the products I'm buying
	Effective Storage	1	Easy to keep extra products on hand (e.g. packaged securely, doesn't spoil)
	Environmentally Friendly Products	1	Environmentally friendly products and packaging
	Easy to Shop for Oral Care Items	3	Oral care items I want are available at the store where I shop
	Product Aesthetics	5	Products that have a "cool" or interesting look

Figure A2. Precision and Recall as a Function of the Size of the Training Sample



(a) Precision



(b) Recall