Information Systems Frontiers
https://doi.org/10.1007/s10796-021-10122-y

Cross-Category Defect Discovery from Online Reviews: Supplementing Sentiment with Category-Specific Semantics

Nohel Zaman[1] · David M. Goldberg[2] · Richard J. Gruss[3] · Alan S. Abrahams[4] · Siriporn Srisawas[5,6] · Peter Ractham[5] · Michelle M.H. Şeref[4]

Accepted: 2 March 2021

Abstract
Online reviews contain many vital insights for quality management, but the volume of content makes identifying defect-related discussion difficult. This paper critically assesses multiple approaches for detecting defect-related discussion, ranging from out-of-the-box sentiment analyses to supervised and unsupervised machine-learned defect terms. We examine reviews from 25 product and service categories to assess each method's performance. We examine each approach across the broad cross-section of categories as well as when tailored to a singular category of study. Surprisingly, we found that negative sentiment was often a poor predictor of defect-related discussion. Terms generated with unsupervised topic modeling tended to correspond to generic product discussions rather than defect-related discussion. Supervised learning techniques outperformed the other text analytic techniques in our cross-category analysis, and they were especially effective when confined to a single category of study. Our work suggests a need for category-specific text analyses to take full advantage of consumer-driven quality intelligence.

Keywords Text analytics · Sentiment analysis · Quality management · Supervised learning · Unsupervised learning · Business intelligence

1 Introduction

The quality management (QM) literature extolls the value of firms' expedient responses and remediation in the event of defective products (Deming & Edwards, 1982; Hora et al., 2011). Firms committed to QM have observed improvements in their stock prices in the short-term (Hendricks & Singhal, 1997) and long-term (Hendricks & Singhal, 2001). Conversely, the effects of poor quality management practices on firm performance are well established, as firms that release defective products may experience a backlash from consumers in the short-term (Jarrell & Peltzman, 1985) as well as long-term damage to their reputation and loss of goodwill (Rhee & Haunschild, 2006). Despite the value of QM,

✉ David M. Goldberg
dgoldberg@sdsu.edu

Nohel Zaman
nohel.zaman@lmu.edu

Richard J. Gruss
rgruss@radford.edu

Alan S. Abrahams
abra@vt.edu

Siriporn Srisawas
srisawas.siri@gmail.com

Peter Ractham
peter@tbs.tu.ac.th

Michelle M.H. Şeref
mmhseref@vt.edu

[1] Department of Information Systems and Business Analytics, College of Business Administration, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045, USA

[2] Department of Management Information Systems, Fowler College of Business, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA

[3] Department of Management, College of Business and Economics, Radford University, P.O. Box 6954, Radford, VA 24142, USA

[4] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, 880 West Campus Drive, Pamplin Hall Suite 1007, Blacksburg, VA 24061, USA

[5] Department of Management Information Systems, Thammasat Business School, Thammasat University, 2 Prachan Road, Bangkok 10200, Thailand

[6] Centre of Excellence in Operations and Information Management, Thammasat Business School, 2 Prachan Road, Bangkok 10200, Thailand

Published online: 30 March 2021

detecting and predicting instances of poor product quality can be challenging. In recent times, online user-generated content (UGC) has presented firms with a new set of tools for monitoring discussions of product quality. Customers that post online about their experience with a product (or service) tend to be very motivated and, in many cases, provide useful feedback in great detail (He et al., 2016; Hu et al., 2012).

In recent times, online user-generated content (UGC) has presented firms with a new set of tools for monitoring discussions of product quality. Online content represents a remarkable opportunity for firms to source real-time consumer feedback on their products (Fan & Gordon, 2014; Zaman et al., 2020). First, online UGC is *voluminous* in that users post thousands of posts of product-related feedback every day (McAuley et al., 2015). This volume is advantageous because even for product lines that experience defects, not every individual product will necessarily be defective (Porter & Van der Linde, 1995), and not every consumer with a defective product will necessarily notice the defect. With a sufficiently large sample of consumer feedback, even rare defects are likely to be represented. Second, online UGC is *timely* in that it provides a continuous stream of consumers posting about their experiences with products (McAuley et al., 2015). On any given date, firms can source recent information pertaining to their products' quality. This facet can allow firms to source feedback rapidly on new product lines and to continuously monitor the status of existing product lines. Third, in many cases, online UGC provides a *targeted* assessment of a product's details and characteristics. Online product reviews and forum posts often consist of a customer's detailed feelings about a product, including their expectations and the product's strengths and weaknesses in relation (Hu et al., 2006). Customers that post online about their experience with a product tend to be very motivated and, in many cases, provide useful feedback in great detail (Hu et al., 2012).

Given the immense promise of online UGC as a source of product feedback, rapidly identifying product defects described in online reviews would allow firms to remediate quickly and likely improve firm performance as part of an effective QM strategy. Unfortunately, detecting discussions of defective products in online reviews and responding accordingly is a difficult process. The majority of online reviews describe positive experiences with products (Hu et al., 2009), and even most one-star reviews may not discuss product defects so much as generic negative sentiment relating to poor customer-product fit (Abrahams et al., 2015). Some researchers have suggested sentiment analyses of online reviews on the assumption that negative sentiment might follow a poor experience with a product of low quality (Hu et al. 2006, 2014; Yu et al., 2013). As many sentiment dictionaries are publicly available and widely used, ranking reviews by negative sentiment represents a rather simple solution. However, sentiment analysis has yielded mixed results in practice, as customers' negative emotive content may not align well with more factual accounts of defect existence (Abrahams et al., 2012; Gopal et al., 2011). A recent study of the automotive industry (Abrahams et al., 2012) found that sentiment was not positively correlated with defect existence, as most defect-related discussions included non-emotive mentions of product components.

Several researchers have attempted more advanced approaches, using machine learning to analyze text in online reviews. Many unsupervised approaches to this problem have involved topic modeling (Qiao et al., 2017; Shi et al., 2017), for which Blei et al. (2003)'s Latent Dirichlet Allocation (LDA) is immensely popular (Guo et al., 2017). The model uses Bayesian probability estimates to allocate terms into a user-specified number of topics, which comprise the corpus. Although the LDA methodology is well established, researchers have noted that it is quite resource-intensive and that the results may be very sensitive to the input parameters (e.g., number of topics) (Tirunillai & Tellis, 2014). As with any unsupervised technique, topic modeling does not guarantee that the output topics match those of the researcher's or practitioner's interests; instead, the algorithm provides the topics that best fit the data. Abrahams et al. (2012) describe a supervised learning approach for using online reviews for QM by mining them for indications of product defects using machine-learned "smoke terms" trained to identify language associated with quality issues. Some recent research has investigated the use of online media in defect discovery within a variety of product or service categories, such as toys, appliances, electronics, and automobiles (Abrahams et al., 2012, 2013, 2015; Adams, Gruss, & Abrahams, 2017; Goldberg & Abrahams, 2018; Law et al., 2017; Liu et al., 2018; Winkler et al., 2016), with recent studies also incorporating deep learning (Brahma et al., 2021). One study proposed an integrated text analytic framework for defect discovery across two product or service categories, the automotive and consumer electronics categories (Abrahams et al., 2015). The authors raise the question of the generalizability of machine-learned terms, as they are sometimes quite specific to the product or service category of study.

Although the literature has explored a variety of potential methods for defect detection in online reviews, researchers are not united as to the more effective techniques. This paper aims to develop and curate a large, cross-category sample of online reviews to assess both the prevalence of defect reports across categories and the efficacy of a variety of methods for identifying product defects in online reviews. We make three key contributions in this paper. First, our research quantifies the prevalence of defect-related discussion in online reviews across a broad cross-section of 25 product (or service) categories. We label a dataset of over 60,000 reviews, allowing us to understand the differences in defect rates across product categories as well as the nature of reviews that report defects (for

example, what portion of defect reports come from 1-star reviews?). We examine the properties of these product (or service) categories, which provides guidance for practitioners. Second, we compare the efficacy of a variety of techniques at predicting the existence of these defects in online reviews, as prior work has suggested techniques ranging from sentiment analysis (Hu et al. 2006, 2014; Yu et al., 2013) to unsupervised (Qiao et al., 2017; Shi et al., 2017) and supervised learning (Abrahams et al., 2015; Abrahams et al., 2013; Abrahams et al., 2012; Goldberg & Abrahams, 2018). We examine the performance of existing methodologies for defect surveillance, and we also apply other recent methods, such as deep learning (Brahma et al., 2021), which have not yet been applied for this task to the best of our knowledge. We compare this performance across the broad cross-section of categories to see how well each method performs in a variety of settings. Third, as researchers have questioned the potential for generalizing insights between product (or service) categories (Abrahams et al., 2015; Goldberg & Abrahams, 2018), we also analyze two different types of machine-learned terms for predicting product defects: terms generated based on a balanced sample of reviews across the 25 product (or service) categories and terms generated based on each category alone. In doing so, we provide guidance to the research community and to practitioners on best practices in detecting defect-related discussion in online reviews.

2 Literature Review

2.1 Quality Management

The quality management (QM) literature classifies defects as manufacturing problems, design errors, and insufficient instructions and warnings (Lyles et al., 2008). Previous research has investigated the relationship between financial performance and quality management. Fornell et al. (1996) found that customer satisfaction is more quality-driven than price-driven, and several studies have found an association between poor firm performance and poor quality in multiple product (or service) categories (Chen et al., 2009; Jarrell & Peltzman, 1985). Past research indicates that there is a positive relationship between investments in quality and market performance (Hora et al., 2011). In the past, QM tools have consisted of quality control tools such as cause and effect diagrams, Pareto charts, and control charts (Abrahams et al., 2015); however, none of these tools is particularly suited for the unstructured nature of textual datasets from the online reviews. In fact, despite the enormous potential value of using textual data for quality analytics, the QM literature has thus far been sparse in exploring these datasets (Abrahams et al., 2015).

Although prior works have suggested using online reviews as a data source to inform QM, the literature has also brought into question the reliability of online reviews as a data source given that consumers can post reviews with little verification (Hu et al., 2008). In recent years, online review platforms such as Amazon have implemented measures such as verifying that reviewers have actually purchased the products they are reviewing, which provides some basic level of validity. The literature has examined the extent to which online reviews provide a useful signal of product (or service) quality extensively. A brief survey of the literature is shown in Table 1.

Table 1 captures a broad cross-section of analyses across product and service categories, and in general the research has found that online reviews are a valuable signal of quality. However, it is important to emphasize that some prior work has found that both extremely positive and extremely negative experiences are highly represented (Hu et al., 2009). Yet, for defect surveillance, representation of negative experiences is especially important (Abrahams et al., 2015). Though online reviews may not be an unbiased sample of all consumers, the evidence of quality provides firms feedback to inform QM efforts. Each review may be taken as a single data point; however, a string of reviews referring to the same problem, such as multiple food reviews indicating digestive distress (Goldberg et al., 2021), may provide sufficient evidence to warrant investigation.

2.2 Online Reviews

Consumers are increasingly influenced by online reviews in their purchasing decisions (Hu et al., 2014; Lee et al., 2008). Because online reviews reflect a personal experience with a product, they are perceived as more trustworthy than seller-controlled advertising (Lee et al., 2008). Online consumer reviews have altered word of mouth (WOM) so that it frequently occurs online in the form of electronic word of mouth (eWOM). The sales of a product can become reliant on the WOM that it produces, with consumers influenced by star ratings in particular (Hu et al., 2012). The volume of online opinion is rising at an astonishing rate (Hu et al., 2014), and it may be more effective for firms to focus on understanding a few specific categories of feedback, such as positive and negative responses. In eWOM communication studies, researchers refer to positivity or negativity as a review's "valence" (Lee et al., 2008; Stern, 1962). Previous research in the field of consumer behavior indicates that consumers focus more on negative information than positive information and consumers prioritize negative information more than positive information during decision-making (Park & Lee, 2009). Baumeister et al. (2001) established that the principle of "bad is stronger than good" is consistent across a broad array of phenomena and proposed that people have a psychological tendency to respond more strongly to negative events than positive ones.

Table 1 Survey of studies on online reviews as indicators of quality

Reference	Key finding
Duan, Gu, and Whinston (2008)	Volume of movie reviews is a significant indicator of box office sales
Hu et al. (2008)	Online Amazon review star ratings are predictive of product sales
Hu et al. (2009)	Consumers with extreme positive or negative experiences are most likely to post online reviews to "brag or moan"
Cui, Lui, and Guo (2012)	Both volume and star rating of electronics reviews are significant indicators of sales
Phillips, Zigan, Silva, and Schegg (2015)	Star ratings and assessments of room quality are predictive of hotel performance
Abrahams et al. (2015)	Products with high sales volumes and thus many reviews benefit from quality analytics, as reviews are more representative
Qi et al. (2016)	Topics discussed in online reviews and associated sentiment provide sufficient data to drive product improvement
Chong et al. (2018)	Existing reviews substantially impact consumer perceptions and purchase decisions

2.3 Sentiment Analysis

Many prior studies related to eWOM communication have performed sentiment analysis, also known as opinion mining (Cu et al., 2017; Ghiassi et al., 2016; Lee et al., 2010; Mostafa, 2013). Sentiment analysis applies methods from natural language processing to automatically detect the polarity of author's attitude toward a topic: positive, negative, or neutral. Sentiment analysis can be coarse-grained or fine-grained. Whereas course-grained sentiment analysis only detects polarity (positive or negative), fine-grained sentiment analysis detects both the polarity and strength. Sentiment strength may be dissimilar for words with the same polarity. For example, "good" and "excellent" are both classified as positive, but the latter states stronger sentiment than the former. Many approaches are dictionary-based. For example, positive and negative terms may be kept in a word list (Nielsen, 2011) such that each word is associated with sentiment strength within some range, such as +5 to −5.

Some text analytics approaches, such as SentiStrength, attempt to use entire sentences to access contextual modifiers that affect the meaning of an individual word (Thelwall et al., 2010). Sentiment analyses have been utilized for a wide range of applications. For instance, researchers have suggested sentiment analyses of online reviews on the assumption that negative sentiment might follow a poor experience with a product of low quality (Hu et al., 2006, 2014; Yu et al., 2013).

2.4 Unsupervised Text Analytics

Unsupervised text analytics involves using computer algorithms to make inferences about text without prior sorting of that text into desired classifications or categories. Thus, unsupervised text analytics do not necessitate the potentially expensive and/or time-consuming process of gathering training data. Perhaps the most popular dimension of unsupervised text analytics in recent literature has been topic modeling, in which a corpus of text is divided into clusters of words, or "topics," that are thematically similar and often used together. As an exploratory tool, topic modeling allows the user to quickly identify major emphases in a corpus.

Two major approaches in the literature are Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA has been an especially popular technique in recent literature. When analyzing a large number of documents, topic modeling such as LDA is often useful to condense the text into key themes or categories for further analysis (Moro et al., 2015). LDA has also been used for such applications as detecting areas of satisfaction and dissatisfaction in the tourism industry (Guo et al., 2017) and in human resources management (Jung & Suh, 2019). Qiao et al. (2017) suggest the use of LDA for a more targeted approach of mining product defects from online reviews, whereas Shi et al. (2017) apply a similar framework toward a safety-specific focus. Both papers suggest that LDA may be an appropriate fit for automated clustering of consumer complaints into categories of defects.

2.5 Supervised and Semi-Supervised Text Analytics

Unlike unsupervised approaches, supervised text analytics depend upon a training set of pre-labeled data. A computer algorithm learns from patterns in this initial training set so that it may predict classifications in an unseen holdout set. Although labeled data can be quite difficult to acquire, it can yield powerful results. Supervised text analytics have been used for a wide variety of applications, including financial projections (Eliashberg et al., 2014), fraud identification (Holton, 2009),

and plagiarism detection (Oberreuter & VeláSquez, 2013). Within this category, deep learning-based word embedding analyses have become a popular state-of-the-art tool (Brahma et al., 2021). These approaches utilize pretrained deep neural network architectures to capture the semantic qualities of text, and predictions tend to be highly accurate.

A recent stream of literature has also examined the use of supervised methods for classifying product defect-related discussions using machine-learned "smoke terms" (Abrahams et al., 2012, 2013, 2015; 2017; Goldberg & Abrahams, 2018; Law et al., 2017; Winkler et al., 2016). These techniques have been applied in numerous category-specific settings but have not yet been examined thoroughly in a cross-category analysis.

Some semi-supervised techniques have sought to include small amounts of labeled data in addition to some larger amount of unlabeled data. Das et al. (2017) describe the general bootstrapping procedure for these techniques. First, a small amount of training data is labeled, and it is then used to train a supervised classifier. Then, that classifier is used to classify unseen data, with some records classified more confidently than others. Finally, the most confidently classified new records are added to the training data, and the classifier is retrained. This method has been used for automating the classification of financial text (Das et al., 2017) as well as news articles (Zhang, 2008). However, semi-supervised methods in text analytics have also utilized combinations of supervised and unsupervised techniques. For example, Zhao et al. (2010) segmented aspect and opinion words by combining a supervised model unsupervised topic modeling. Lau et al. (2014) also use a semi-supervised method to combine expert opinion with topic modeling in social media.

3 Methodology

3.1 Dataset and Data Coding

Our dataset consists of 37 million online reviews of products and services across 25 different product (or service) categories (McAuley et al., 2015). For each category, we selected a random subset of reviews for manual tagging (data coding). Based on prior work (Abrahams et al., 2012, 2015; Goldberg & Abrahams, 2018), we developed the following protocol to classify each review as 1. "defect" or 2. "no defect."

1. **Defect:** We define a product or service to have a defect when it does not function to the manufacturer's specifications or to the consumer's reasonable expectations. Defects may be specified in emotive or non-emotive (factual) tone (e.g. "the picture separated from its backing just from handling it once"; "the plastic peice (sic) that holds the mic in place snapped in two") and are not limited to text with negative sentiment only. Our "defect" classification includes reviews from both "performance defects" and "safety defects" examined in prior work (Abrahams et al., 2012).
2. **No Defect**: We define a product or service to have no defect if the review does not reflect a specific product or service problem. The reviews include only positive feedback, recommendations, reviews with negative sentiment unrelated to product problems (e.g. "Miko will play until she is completely *exhausted* ... this has held up extremely well to all the *abuse* she can dish out"), or general comments.

3.2 Data Processing

Figure 1 illustrates the steps we employed in data processing.

We detail the steps that we employed for data processing as follows. Each numbered step in Fig. 1 is described in the corresponding numbered paragraphs below:

1. We randomly selected at least 2000 reviews from each of the 25 product (or service) categories. Prior research has suggested that the techniques discussed below that we will utilize for curating smoke terms are effective when at least 100 positive (defect) reviews are available in both the training set and the holdout set, or 200 total (Abrahams et al., 2013; Abrahams et al., 2012). As we did not initially know the portion of reviews tagged as defects, selecting at least 2000 reviews ensured that each product (or service) category was properly represented with enough reviews to generate meaningful smoke terms. In total, we selected 62,980 unique random reviews from the 25 different categories.
2. Using the protocol described in section 3.1, over 500 undergraduate research volunteers majoring in business were assigned to the manual tagging projects, working in small teams, with one team per product (or service) category, at a large public university in the United States. The reviews were randomly distributed to each tagger along with a detailed protocol describing the tagging procedure. Each individual tagger was asked to tag 100–200 reviews. To compute inter-rater reliability, a lead tagger was assigned to each product (or service) category. Lead taggers were randomly assigned shared reviews (overlapping cases) with other tagger in the group, and the tags of the lead tagger were compared to the tags of the other taggers by calculating Cohen's κ (1968). Per Landis and Koch (1977), κ in the range 0.41–0.60 represents moderate agreement, 0.61–0.80 represents substantial agreement, and $\kappa > 0.80$ represents almost perfect agreement. Per Fleiss et al. (2013), κ in the range 0.4–

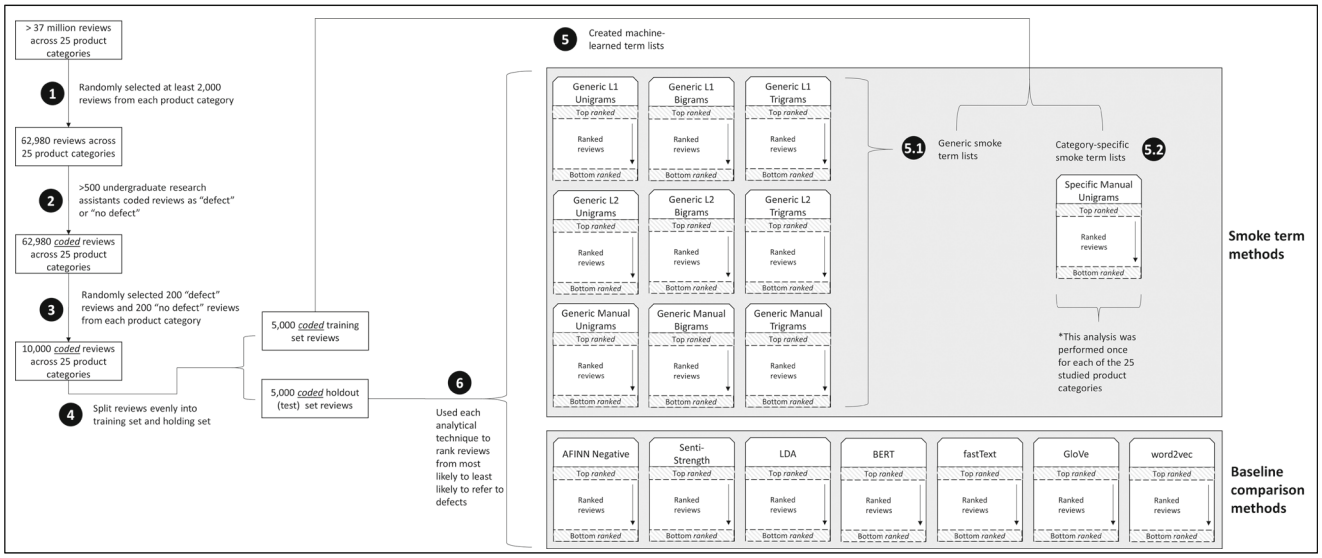

Fig. 1 Overview of methodology

0.75 represents fair to good agreement, and $\kappa > 0.75$ represents excellent agreement. Where conflicting codes were assigned by taggers, the majority opinion was accepted as correct, and in the event of a tie, we chose the most conservative decision (i.e., "defect" over "no defect") (Goldberg & Abrahams, 2018; Goldberg et al., 2021; Law et al., 2017; Mummalaneni et al., 2018). We observed at least "moderate" or "fair to good" agreement in every category of study, and in most categories, the agreement was at least "substantial" or "excellent." Table 2 details the product (or service) categories, data sources, number of reviews, and other details about the dataset. We also show Cohen's κ on defect existence, along with number of overlapping cases (in parentheses), for each product (or service) category, which confirms acceptable inter-rater reliability.

3. There was a large imbalance in the number of defects in the random reviews sampled per product (or service) category for two reasons. First, as the taggers from different teams achieved differing tagging outputs, we had as few as 2000 reviews tagged in power tools and joint and muscle pain, but we had as many as 3746 reviews tagged in furniture. Second, the portion of defects also varied substantially by category, ranging from 10% for health and medical care to 73% for airports. Thus, using the unbalanced dataset for further stages of our analysis may have biased our results as some product (or service) categories would have been represented more than others, and given the variability in defect rates, classification would have been far easier in some product (or service) categories than others. To avoid these biases, we created a balanced sample of 200 random defects and 200 random non-defects from each product (or service) category. Thus, each product (or service) category was equally represented and had the same effective defect rate for the ensuing analysis. Additionally, creating a balanced dataset with equal numbers of positive (defect) and negative (non-defect) reviews ensures that our classification technique cannot take advantage of one class occurring more frequently than the other. For instance, if 10% of reviews were positives (defects) and 90% negatives (non-defects), then a technique could be guaranteed 90% accuracy by always choosing the "non-defect" classification. Balancing the dataset forces the modeling to rely upon the dataset's underlying features rather than the underlying positive/negative ratio.

4. The balanced sample was then further divided into a *training* set consisting of 5000 reviews (2500 defects and 2500 non-defects) and a *holdout* set consisting of 5000 reviews (2500 defects and 2500 non-defects). This division ensured that at least 100 positive (defect) reviews were available in both the training set and the holdout set, or 200 total, which prior research suggests for best performance (Abrahams et al., 2012, 2013).
5. We created sets of supervised smoke terms – unigrams (single words), bigrams (two-word sequences), and trigrams (three-word sequences) – or terms that were predictive of defect-tagged reviews based on a training set (Abrahams et al., 2012, 2015). In performing a cross-category analysis, we were interested in the ability of smoke terms to generalize across product (or service) categories. Thus, we first (5.1) generated smoke terms using the entire cross-industry training set. We refer to these smoke term lists as **generic smoke terms**, which were trained based on all 25 categories for the purposes of predicting defect existence in all 25 categories. As these

Table 2 Descriptive statistics for dataset of online reviews tagged across 25 product (or service) categories

								Percentage of defects in:				
Product (or service) category	Website	Total Reviews Available	Unique Entities Reviewed	Start - End Dates	Random Reviews Tagged	Cohen's κ (# cases)	Portion Defect	1-Star	2-Star	3-Star	4-Star	5-Star
Airline	Airline Quality	44,499	399	Nov 2003-Jan 2016	2054	0.66 (705)	65%	27	16	14	28	15
Airport	Airline Quality	15,925	774	Apr 2004-Jan 2016	3000	0.72 (312)	73%	45	30	11	10	4
Baby Products	Amazon	1,063,430	71,711	June 2000-Jul 2014	2223	0.73 (1471)	23%	28	24	28	14	6
Clothing	Amazon	5,352,922	1,038,411	Oct 1999-Jul 2014	2055	0.77 (959)	33%	24	13	24	29	10
Collectible & Fine Art	Amazon	7781	5339	Sep 2003-Jul 2014	2257	0.88 (743)	18%	37	15	23	17	8
Cosmetics	Amazon	301,457	48,573	Apr 2000-Jul 2014	2003	0.77 (565)	23%	41	13	18	19	9
Crafts & Sewing	Amazon	506,267	107,461	May 1998-Jul 2014	2652	0.82 (1471)	25%	23	17	29	27	4
Credit Cards	Amazon	9341	21	May 2012-Jul 2014	2279	0.75 (157)	42%	71	19	10	0	0
Dishwashers	Amazon	4620	559	Jul 2002-Jul 2014	2321	0.76 (354)	55%	67	14	6	3	10
Food	Amazon	1,288,309	160,783	Aug 2000-Jul 2014	2216	0.73 (794)	24%	23	17	28	17	15
Furniture	Amazon	295,970	34,729	Apr 2004-Jul 2014	3746	0.75 (768)	25%	33	17	23	22	5
Garden Tools	Amazon	3343	325	Jan 2003-Jul 2014	2066	0.75 (661)	16%	44	21	19	8	8
Health and Medical Care	Amazon	6687	566	Apr 2004-July 2014	3172	0.81 (411)	10%	32	20	17	9	22
Higher Education	Koofers	N.A.	N.A.	Sep 2008-Sep 2014	2619	0.73 (110)	49%	20	22	24	19	15
Hotels	Travel-ocity	31,433[a]	945	Feb 2003-Jan 2012	3000	0.69 (141)	43%	11	12	21	34	22
Household Products	Amazon	385,707	30,416	Oct 1999-Jul 2014	2140	0.82 (739)	25%	36	17	29	9	9
Industrial & Scientific	Amazon	260,398	43,407	Nov 1999-Jul 2014	2641	0.78 (1486)	25%	29	17	28	23	3
Joint & Muscle Pain	Amazon	32,501	1939	Mar 2004-Jul 2014	2000	0.80 (100)	28%	35	20	20	20	5
Musical Instruments	Amazon	435,151	64,655	Jan 2000-Jul 2014	2674	0.67 (775)	21%	29	17	30	19	5
Office Products	Amazon	1,225,988	124,807	Aug 1998-Jul 2014	2059	0.77 (700)	25%	41	23	17	16	3
Packaging	Amazon	3733	680	Feb 2005-Jul 2014	2655	0.48 (1575)	33%	12	13	23	26	26
Pet Products	Amazon	1,227,484	99,903	Oct 1998-Jul 2014	3230	0.82 (170)	23%	39	15	29	10	7
Power Tools	Amazon	140,823	10,846	Nov 1999-Jul 2014	2000	0.70 (486)	25%	22	14	15	29	20
Refrigerators	Amazon	3448	714	Oct 2005-Jul 2014	2918	0.78 (154)	53%	52	17	15	7	9
Toys	Amazon	2,234,519	315,974	Jul 1999-Jul 2014	3000	0.79 (758)	20%	37	18	19	18	8

[a] due to dataset volume, only data for top 10 hotels in 10 most-visited US cities was used

N.A. indicates "not available": data source vendor did not provide

smoke terms are trained across all studied product categories, they may include terms such as "broken," "defective," or "didn't work" that may be applicable across a wide variety of categories. After using the CC score algorithm (Fan et al., 2005) to generate initial candidate smoke terms, the lists were manually curated by the lead author and reviewed by a coauthor (for additional information on this process, please see Appendix A). Consistent with prior studies (Abrahams et al., 2012; Winkler et al., 2016), we removed common product or service-specific words (e.g., airline, airport, hotel, power tools) and common English words (e.g., a, in, for, but). We compared our methodology of manually curating *n*-grams with regularization (penalization) techniques, i.e., lasso (L1) and ridge (L2) logistic regression methods and other machine learning approaches, i.e., neural networks, naïve Bayes, and support vector machines (SVM). To do so, we constructed document-term matrices that related each review to each *n*-gram; these matrices were rather sparse as most reviews did not contain most candidate smoke terms. Each regularization or machine learning technique utilizes these matrices to select terms for retention or exclusion. As a second level of analysis (5.2), we also constructed **category-specific smoke terms**, which were trained based on only one category of interest for the purposes of predicting within that category. For example, we may observe terms such as "poisoning" or "made me hurl" in the food category; these terms may be especially effective within this category, but they may not be generalizable to other categories such as office products or toys. These smoke terms were also generated using the CC score algorithm (Fan et al., 2005) to generate initial candidate smoke terms, the lists were also manually curated by the lead author and reviewed by a coauthor. We repeated this analysis for each of the 25 product (or service) categories studied. For brevity, we focus on the results of unigram analysis only.

6. We assessed the holdout set using each of the methods discussed above. Each method was used to rank the holdout set from most likely to least likely to refer to a defect, and better performing methods yielded more true positives (defects) at lower ranks. As baselines, we also compared the performance of several other techniques referenced in the literature. The first sentiment analysis method, AFINN (Nielsen, 2011), is a lexicon-based tool that assigns each sentence a valence score between −5 (most negative) and +5 (most positive). For our purposes, only negative scores were of interest, because we are testing the conjecture that words with negative sentiments are indicative of defects. We took the absolute value of the negative sentiment score so that higher values would indicate a higher likelihood of defect existence. The second sentiment analysis method, SentiStrength (Thelwall et al., 2010), gives scores similarly to AFINN, assigning negative scores between −1 and −5, but additionally considers word context within a sentence. Again, we took the absolute value of the negative sentiment score. Using the 5000 reviews in the training set, we first used LDA (Blei et al., 2003), an unsupervised technique, to generate "topics" comprising our corpus. We ran the LDA for 1500 iterations to retrieve 10 topics of 15-word length each. Finally, word embeddings are a recent deep learning development that utilize large datasets to pre-train complex machine learning models that generate high-dimensional vectors capturing the semantic properties of passages. We deployed four recent word embedding models, including BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), fastText (Bojanowski et al., 2017), GloVe (Global Vectors) (Pennington et al., 2014), and word2vec (Mikolov et al., 2013). In each case, the pre-trained embeddings were utilized to build a Bidirectional Long Short-Term Memory (Bi-LSTM) model, which builds upon a traditional Recurrent Neural Network (RNN) and is commonly utilized in analogous tasks in the literature (Brahma et al., 2021).

3.3 Evaluation Metrics

We use smoke terms to rank reviews from most likely to least likely to refer to defects, and we assess the quality of this ranking (for more details on determining this ranking, see Appendix A). In the following, we establish a vernacular for assessing the quality of such a ranking. After ranking reviews from most likely to least likely to refer to defects, we can assess quality by determining a ranking cutoff, where reviews that fall within the cutoff are predicted to refer to defects, and reviews that fall outside of the ranking cutoffs are predicted to refer to non-defects. For instance, we could choose a cutoff of the top 200-ranked reviews, in which case the top 200-ranked reviews are predicted to refer to defects, and all other reviews are predicted to refer to non-defects. We refer to defects as the positive (or target) classification and to non-defects as the negative classification. Thus, within this top-ranking set of reviews, we refer to reviews that actually refer to defects as true positives. In turn, reviews that do not refer to defects are instead referred to as false positives.

It is desirable for as many of the reviews within the cutoff as possible to be true positives. For example, a technique that produces 150 true positives in the top 200-ranked reviews is performing better than a technique that produces only 100 true positives in the top 200-ranked reviews. In comparing the quality of predictions by alternate methods, we use the chi-square test of independence first to determine that the difference between one method's highly ranked reviews and its

lowly ranked reviews is statistically significant and second to determine whether the difference between two methods is statistically significant. This test makes few assumptions in that it is distribution-free and allows us to analyze categorical data in a contingency table. The chi-square test states a null hypothesis that two variables (in our case, the chosen method and performance) are independent and an alternative hypothesis that two variables are not independent. Using the chi-square distribution, if we find enough evidence to reject the null hypothesis, it suggests that there is a statistically significant relationship between the two variables. We first use this test to validate whether one method's highly ranked reviews differ from its lowly ranked reviews; if we reject the null hypothesis, then it implies that the proportion of defect-tagged reviews significantly differed, and the ranking scheme is working effectively. Second, we use this test to assess whether two methods differ from one another in performance. If we reject the null hypothesis, then it implies that one method significantly outperformed the other.

In addition to the raw number of true positives (defects) on the top-ranking set of reviews, we can also calculate several more advanced metrics. Normalized discounted cumulative gain (nDCG) is an information retrieval metric that measures the quality of a ranking of items (Järvelin & Kekäläinen, 2002). Rather than simply counting the number of positive (defect) and negative (non-defect) reviews observed within a cutoff, nDCG uses a logarithmic function to measure the position of each record. The nDCG metric is weighted such that it credits higher-ranking records more than lower-ranking records. Values closer to 1 represent a closer to ideal ranking, and values closer to 0 represent the worst possible ranking. Next, precision, which refers to the proportion of items predicted to be positives that are true positives. For example, if 150 out of 200 predicted positives are true positives, then precision is 150/200 or 0.75. A further measure, known as recall, assesses the proportion of all positive items that have been identified by the classifier. For instance, if 300 total positives exist in the dataset, and 150 of them have been identified as such, then recall is 150/300 or 0.50. Generally, we observe an inverse relationship between precision and recall. That is, if we choose a smaller threshold, then we usually observe higher precision but lower recall as classifications are more confident, but fewer true positives are identified; at a large threshold, classifications are less confident, so we observe higher recall at the expense of lower precision. In this paper, we use both metrics to assess classification quality.

Each of the above evaluation metrics assumes a specific cutoff for delineating predicted positives from negatives, but an alternative evaluation metric, area under the curve (AUC), assesses performance graphically across a continuum of cutoffs. We use AUC to plot the relationship between chosen cutoffs and the number of true positives (defects) detected. The area under the resultant series is then geometrically calculated and scaled between 0 and 1. An AUC of 1 represents the best possible model; an AUC of 0 represents the worst possible model; and an AUC of 0.50 represents random chance. We use AUC as a complement to the above-mentioned metrics to assess the entire continuum of reviews in addition to the top-ranking portion.

Finally, as a last stage of our analysis, we construct predictive models that use a combination of several variables together to predict defect existence. To do so, we use logistic regression, which allows us to generate binary (i.e., defect or non-defect) predictions for each record. We can assess the statistical significance of each variable in the model as well as the model as a whole. To assess the statistical significance of each variable, we use Wald tests, which test the null hypothesis that the variable's coefficient equals zero versus the alternative hypothesis that the variable's coefficient does not equal zero. If we have sufficient evidence to reject the null hypothesis via a Wald test, then it suggests that the inclusion of the associated variable in the model meaningfully contributes to the model's predictive accuracy. On the model level, we use two statistical measures: the lack of fit test and the whole model test. The lack of fit test, sometimes referred to as the goodness of fit test, seeks to answer whether the model includes all relevant variables or if additional terms are clearly missing from the regression equation. This test analyzes the null hypothesis that the model contains all relevant variables against the alternative hypothesis that the model does not contain all relevant variables. In this test, failure to reject the null hypothesis suggests that the model is providing a high-quality prediction. Finally, we use the whole model test, which compares the prediction of the model including all regression parameters to the prediction of the model using only an intercept. This test compares the null hypothesis that the full model does not provide significantly better prediction than the intercept-only model to the alternative hypothesis that the full model provides significantly better prediction than the intercept-only model. For this test, rejection of the null hypothesis serves as evidence that the included variables as a whole provide meaningful prediction of the dependent variable.

4 Results and Discussion

4.1 Descriptive Statistics

Table 2 shows the number of defects found in the sample of 2000 random reviews tagged in each product (or service) category and, in parentheses, the percentage of tagged reviews containing defects. The product (or service) categories with the highest percentage of defects were airlines and airports, with 1457 defects (73% of tagged reviews in airline) and 1308 defects (65% of tagged reviews in airport) respectively. The product (or service) category with the lowest number of

defects per 2000 reviews was health and medical care (over-the-counter treatments and devices), with 208 defects per 2000 reviews (10%), indicating substantial variability in the proportions of defect reports across categories.

The right-most five columns of Table 2 show the percentage of defect reports in each star rating for each product (or service) category. Surprisingly, in 18 of 25 product (or service) categories, 1-star, 2-star, and 3-star reviews accounted for less than 80% of defect reports. Alternatively stated, in the vast majority of product (or service) categories, 20% or more of reported defects were in high star (4-star and 5-star) reviews.[1] We also observed considerable variability in the distribution across categories: while low star ratings seemed highly related to defect prevalence in credit cards, we observed the opposite relationship in the packaging category. This finding runs counter to the predominant conventional wisdom that defect reports are predominantly found in low-star reviews (Hu et al., 2014). With this new, broad, cross-product revelation that up to 56% of defects are reported in high star (4- and 5-star) reviews, companies can begin to exercise more effort in understanding defects being reported by product advocates, not just defects reported by product detractors who award low star ratings. Specifically, it is clear from this analysis that companies should not rely only on consumer sentiment (low star ratings) as their sole source of product defect insights. On average, across product (or service) categories, over one quarter (27%) of defects are mentioned in high (4-star or 5-star) star reviews.

4.2 Cross-Product (or Service) Category Analysis

We assessed the performance of a variety of text analytic methods for detecting defects in a cross-category setting. We began with out-of-the-box sentiment analysis techniques before incorporating more advanced techniques such as unsupervised and supervised learning. Our unsupervised learning technique, LDA, was run on the training set spanning the 25 product/service categories. As we sought to develop terms that predicted defects, we tried running LDA on only those reviews tagged as referring to defects. After carefully examining the words (see Table 3) retrieved from LDA from each topic, most of them pertain to descriptions of specific product or categories. For example, Topic 1 explains all the parts and a specific brand name (Bosch) of a dishwasher. Only Topic 10 explains a problem (product returns) with one product that was taken back after using it for months. However, none of the topics described any defect, even though LDA analyzed only the defect-tagged reviews. We also ran the LDA analysis for various other numbers of topics ranging from 5 to 25, and we observed similar results. These results present a stark contrast when compared to our supervised learning technique, in which *n*-grams pertaining to defect-related language were specifically prioritized. We also attempted a second LDA analysis on the entire training set of 5000 balanced defect and non-defect reviews to generate two topics (see Table 4), which we hoped would pertain to defects and non-defects. However, the LDA analysis again did not appear related to defects. As users cannot control the output topics generated by LDA, it did not appear to be a reliable tool for detecting defect-related discussion.

Table 5 summarizes the number of defects found in the top- and bottom-ranked reviews, as scored using each scoring method when analyzing the full holdout (test) set of 5000 reviews across all product (or service) categories (Step (5) of §3.2). **Bolded** text indicates the best-performing cross-product or service defect scoring method, or **smoke trigrams**, which detected 176 defects in the top 200-scoring reviews and also had the best nDCG values of 0.86 in the top 200-ranked reviews and 0.95 in the bottom 200-ranked reviews. The proportion of defects in the top-scoring reviews via each method were clearly differentiable from the proportion of defects in the bottom-scoring reviews, as chi-squared tests indicated significantly different proportions for each method at the 0.001 level. Overall, it appeared that the smoke term methods, and particularly the trigrams, outperformed the competing approaches.

Figure 2 compares each of these scoring methods to a random chance baseline (diagonal black line), where the baseline indicates the number of defects expected to be found in a random selection of that number of reviews. For each scoring method, Fig. 2[2] depicts the number of defects (y-axis) found in the top-ranked *N* reviews (x-axis) using that scoring method. As the holdout set was balanced (50% of reviews in the holdout set are defects) the baseline runs at a 45° diagonal.

We computed the area under curve (AUC) statistic to assess the performance of each method. Values of AUC approaching 1 indicate the best models, which seldom have false positives in the top-scored items and seldom have false negatives in the bottom-scored items. Additionally, we present an overview (see Fig. 3) of precision and recall for each of the best scoring methods. A chi-squared test comparing the

[1] The unexpectedly high proportion of defects in high-star reviews may, in part, be due to active moderation by online retailers of extremely negative reviews, which may reduce total defect reports, particularly in low-star reviews. For instance, Amazon does not post submitted reviews that "violate community guidelines". Manufacturers typically cannot access moderated reviews submitted on a retailer's website, since these have been suppressed; such reviews may have contained additional defect reports.

[2] Only the best three of the available scoring methods, as measured by AUC, were depicted in Figure 2 to enhance readability. A chi-squared test of the proportion of defects in the top 200 reviews for the two best sentiment methods (AFINN versus SentiStrength) indicates they did not significantly differ ($p = 0.18$); we chose AFINN as the best sentiment method due to its marginally better results. Similarly, smoke bigrams were marginally better performing than smoke unigrams but did not significantly differ from the smoke unigrams via a chi-squared test ($p = 0.14$).

Table 3 LDA-generated words from 2500 defect-tagged posts

Topic 1	dishwasher, water, dishes, clean, door, fridge, get, unit, dry, top, still, use, bosch, wash, even
Topic 2	class, hard, really, time, take, get, easy, test, lot, grade, book, need, read, every, work
Topic 3	room, hotel, desk, stay, night, front, us, one, nice, day, great, time, clean, get, check
Topic 4	get, card, amazon, pay, time, one, make, never, 34, credit, order, store, find, payment, quality
Topic 5	use, one, made, back, plastic, 34, size, little, great, fit, quality, top, side, put, together
Topic 6	flight, airport, time, food, security, service, seats, terminal, staff, minutes, one, check, passengers, seat, business
Topic 7	product, pack, color, box, buy, printer, got, better, bag, really, ordered, tea, one, picture, get
Topic 8	one, get, great, small, use, little, old, much, still, price, used, way, even, enough, put
Topic 9	product, quote, pain, work, take, hair, use, used, using, try, skin, works, help, really, much
Topic 10	one, product, years, new, buy, months, first, back, work, two, time, bought, problem, year, worked

proportion of defects in the top 200-ranked reviews for each of the top two scoring methods (smoke trigrams and AFINN negative) indicates the two methods significantly differed from one another ($p < 0.01$), as smoke trigrams identify more defects in the top 200-ranked reviews (176 versus 160).

Figure 2 shows graphically that the smoke term lists are more effective than sentiment analysis in finding defects over the whole of the distribution. Interestingly, while the trigram list was the most effective method in the top-scoring portion of reviews, *the bigram list was more effective over the entire distribution of reviews*, producing a superior AUC value. While the trigram list was very effective over the top 200-ranked reviews, its effectiveness tailed off thereafter, whereas the bigrams were most effective over the entire distribution of reviews. Relatedly, while each of the sentiment methods was fairly effective in the top-scoring portion of reviews, each smoke term method outperformed each sentiment method over the entire distribution of reviews as measured by the AUC scores. Thus, we will utilize the bigrams list as the most effective cross-category smoke term list in our ensuing analyses. Each practitioner must decide as to the balance between maximizing true positives (defects) in the very top-ranking reviews and considering the overall AUC offered by different methods. Practitioners with limited resources seeking to gain quick insights from online media may only have time to read the top-ranking set of reviews, in which case the improvement of 176 true positives (defects) for smoke trigrams versus 160 true positives (defects) in the top 200-ranked reviews represents a 10% improvement in efficiency (Table 5). For practitioners with more resources for surveillance, the higher AUC may make smoke bigrams more desirable.

Figure 3 shows visually both the precision and recall of the smoke terms versus the sentiment methods at arbitrary cutoffs. Precision measures the percentage of top-scoring reviews that are true positives (i.e., number of defects found at a threshold divided by that threshold), and recall measures the percentage of all true positives that are actually found at a certain threshold (i.e., number of defects found at a threshold divided by the total number of defects in the holdout set). For instance, the bigram scoring approach identified true positives 76% of the time (precision) and captured 15% of the true positives (recall) in the top 500-ranked reviews. As the bigram smoke term method had the highest performance of the text analytic methods, we then compared this specific smoke scoring method with L1 (see Fig. 4) as well as L2 (Fig. 5) regression-curated terms.[3] Overall, both figures show that the manually curated bigram scoring technique was more effective than the L1 *n*-grams and the L2 *n*-grams in detecting defects.

Table 4 LDA-generated words from 2500 defect-tagged and 2500 non-defect-tagged posts

Topic 1	one, get, product, use, time, work, really, great, little, bought, back, much
Topic 2	room, time, service, get, card, flight, hotel, one, airport, check, food, us

[3] As benchmarks, we also compared these techniques to more general machine learning techniques, namely neural networks, naïve Bayes, and support vector machines (SVM). We implemented neural networks in JMP Pro, which uses a penalized Gaussian (least squares) maximum likelihood function. We initially used a single hidden layer, and we later found that that adding an additional hidden layer did not improve results. We implemented SVM using the scikit-learn Python library, and we used the default settings of a 1.0 penalty parameter using a radial basis function (RBF) kernel with a polynomial kernel function. We found that neural networks, naïve Bayes, and SVM yielded 156, 124, and 154 true positives (defects) in the top 200-ranked reviews of the holdout set, and we observed AUC values of 0.58, 0.54, and 0.58 respectively. As such, these techniques did not outperform the other methods that we attempted. However, smoke terms are advantageous in that they are more easily interpretable and explainable, whereas these other methods may be "black boxes" for which it is difficult to articulate clear reasoning as to each prediction.

Table 5 Number of defect-tagged reviews and normalized discounted cumulative gains (nDCG) in top 200 and bottom 200-ranked reviews

	Scoring method	Defect	No defect	nDCG
Sentiment	*AFINN Negative*			
	Bottom 200	43	157	0.71
	Top 200	160	40	0.80
	SentiStrength			
	Bottom 200	41	159	0.65
	Top 200	152	48	0.80
Deep learning	*BERT*			
	Bottom 200	32	168	0.82
	Top 200	140	60	0.68
	fastText			
	Bottom 200	43	157	0.70
	Top 200	121	79	0.59
	GloVe			
	Bottom 200	38	162	0.75
	Top 200	136	64	0.63
	word2vec			
	Bottom 200	39	161	0.72
	Top 200	128	72	0.61
Generic Smoke List	*Unigrams*			
	Bottom 200	21	179	0.94
	Top 200	140	60	0.69
	Bigrams			
	Bottom 200	21	179	0.94
	Top 200	149	51	0.74
	Trigrams			
	Bottom 200	**21**	**179**	**0.95**
	Top 200	**176**	**24**	**0.86**

4.3 Product (or Service) Category-Specific Analysis

To investigate differences in sentiment across product (or service) categories, we scored the reviews in each product (or service) category using the top-performing sentiment method, AFINN. Figure 6a, b show the negative and positive sentiment AFINN words per 100 words for defects versus non-defects in each product (or service) category. Across all product (or service) categories, there were more negative sentiment words per 100 words for defects than for non-defects. However, the proportions of negative sentiment words in non-defects in some product (or service) categories (e.g., health and medical care) were higher than the proportion of negative sentiment words in defects in other product (or service) categories (e.g., baby products), and vice versa. For example, defects in household products had a lower proportion of negative sentiment words than non-defects in joint and muscle pain relief. This finding implies that there is no single cross-product (or service) category negative sentiment threshold that seems to reliably identify defects.

Finally, we ran logistic regression analyses to establish the nature of the relationships between each of the scoring methods (independent variables) and defect existence (dependent variable) for each of the product (or service) categories. In particular, this assesses the value of *category-specific* smoke terms relative to *generic* smoke terms. We ran this analysis on our holdout set to mitigate potential overfitting. We used four input variables (inverse star ratings, standardized AFINN negative sentiment score, standardized generic bigram score, and standardized category-specific smoke

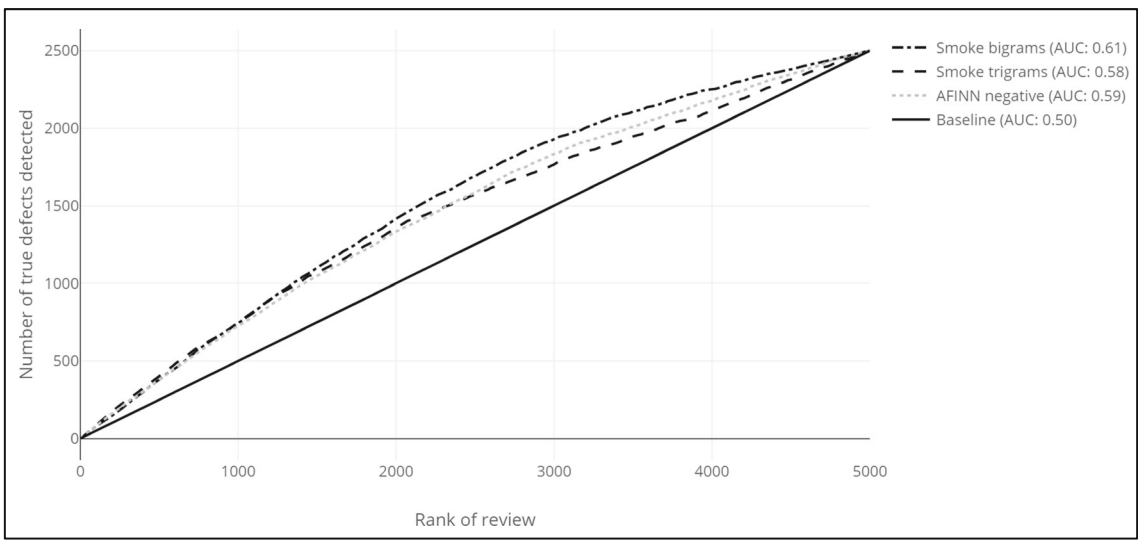

Fig. 2 Defect-discovery performance of generic smoke bigrams, generic smoke trigrams, and AFINN negative

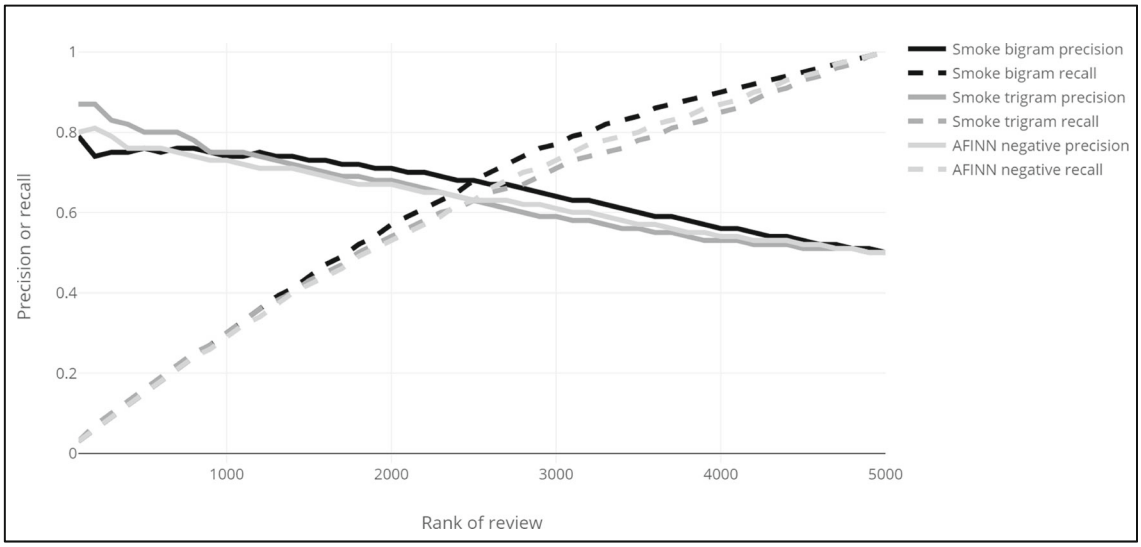

Fig. 3 Precision and recall of generic smoke bigrams, generic smoke trigrams, and AFINN negative for defect classification in top *N*-ranked reviews

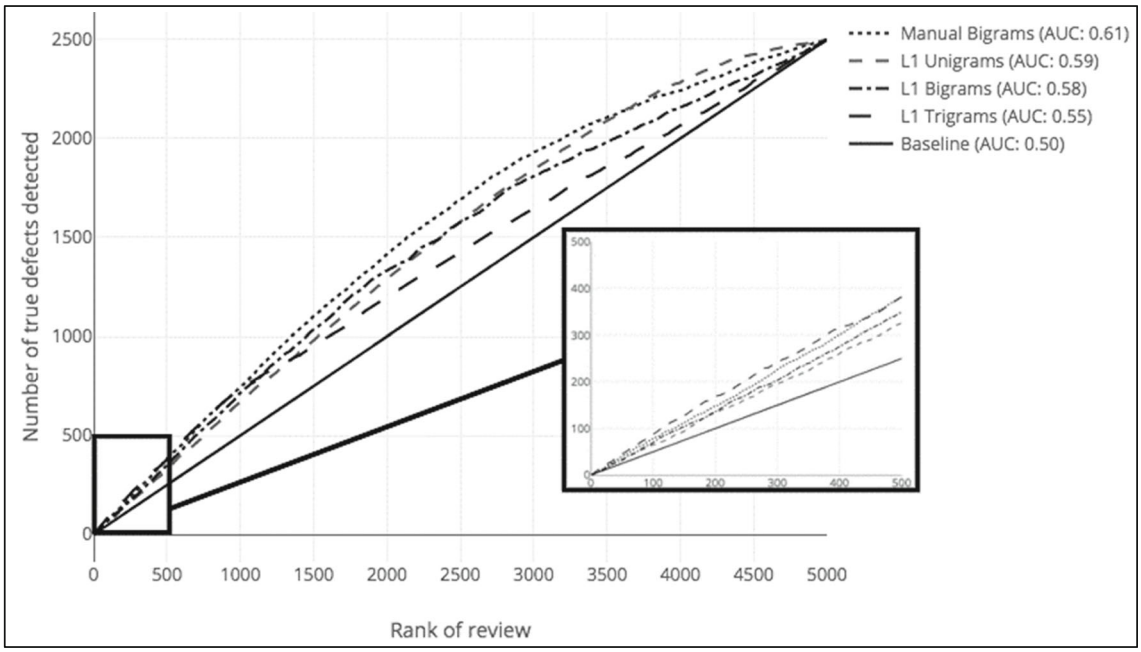

Fig. 4 Defect discovery performance of generic smoke bigrams (manual curation) versus L1 regression scoring methods

method).[4] To enhance interpretability, star ratings are inverted such that positive (5-star) reviews receive low scores and negative (1-star) reviews receive high scores. The results are reported in Table 6. *p*-values less than 0.05 indicate that there is significant evidence supporting the association between scores for each review using the scoring method and defect existence in the scored reviews. Associations with greater than 95% confidence are **bolded**. Information regarding model fit (i.e., AUC, lack of fit, and whole model test) is provided in the right-hand columns of Table 6.

While sentiment analysis was broadly effective in top-ranking reviews of in the cross-category sample considered above, it is possible that these top-ranking reviews represent "easier" classifications. Therefore, as we consider more category-specific use cases with more difficult classifications, sentiment is not a very effective technique, and in some cases our sentiment method was associated with a negative coefficient. Generic cross-product or service bigram smoke terms are significantly predictive of defects across 3 product (or service) categories. Category-specific smoke terms were far

[4] For interpretability, all variables were scaled from 0 to 1 where 0 indicates lower defect likelihood and 1 indicated higher defect likelihood.

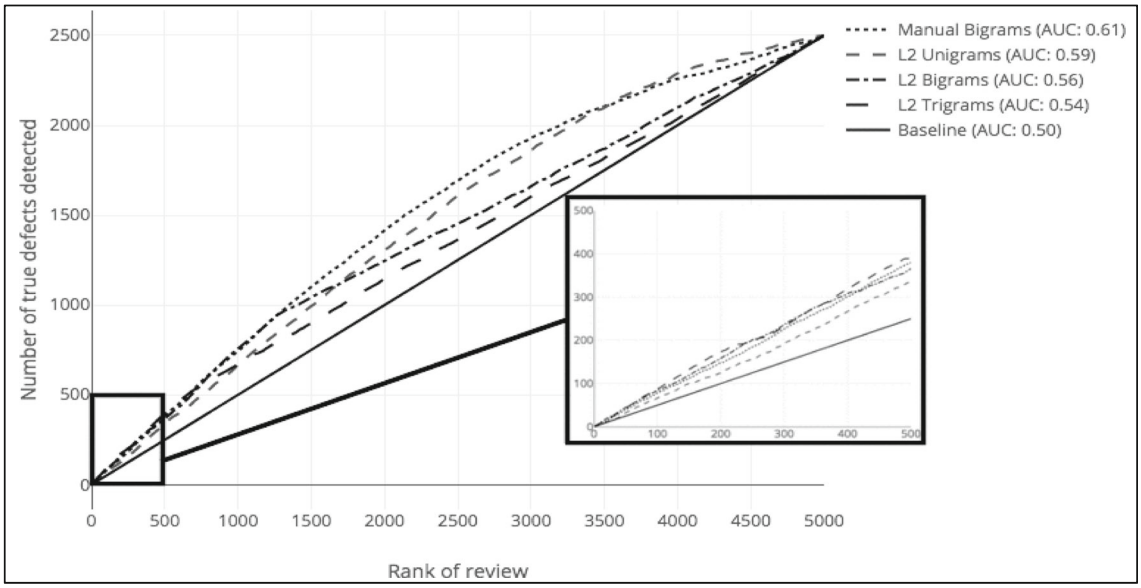

Fig. 5 Defect-discovery performance of generic smoke bigrams (manual curation) versus L2 regression scoring methods

more effective and were statistically significant in 20 out of 25 categories.

In most cases, the models indicated that the category-specific smoke terms had greater positive coefficients than other variables considered, indicating that they were most predictive of true defect-tagged reviews. Our results suggest that smoke terms are in fact specific to product (or service) categories. While a universal smoke term list is elusive, product (or service) category-specific smoke word lists have proven more effective than star ratings in multiple product (or

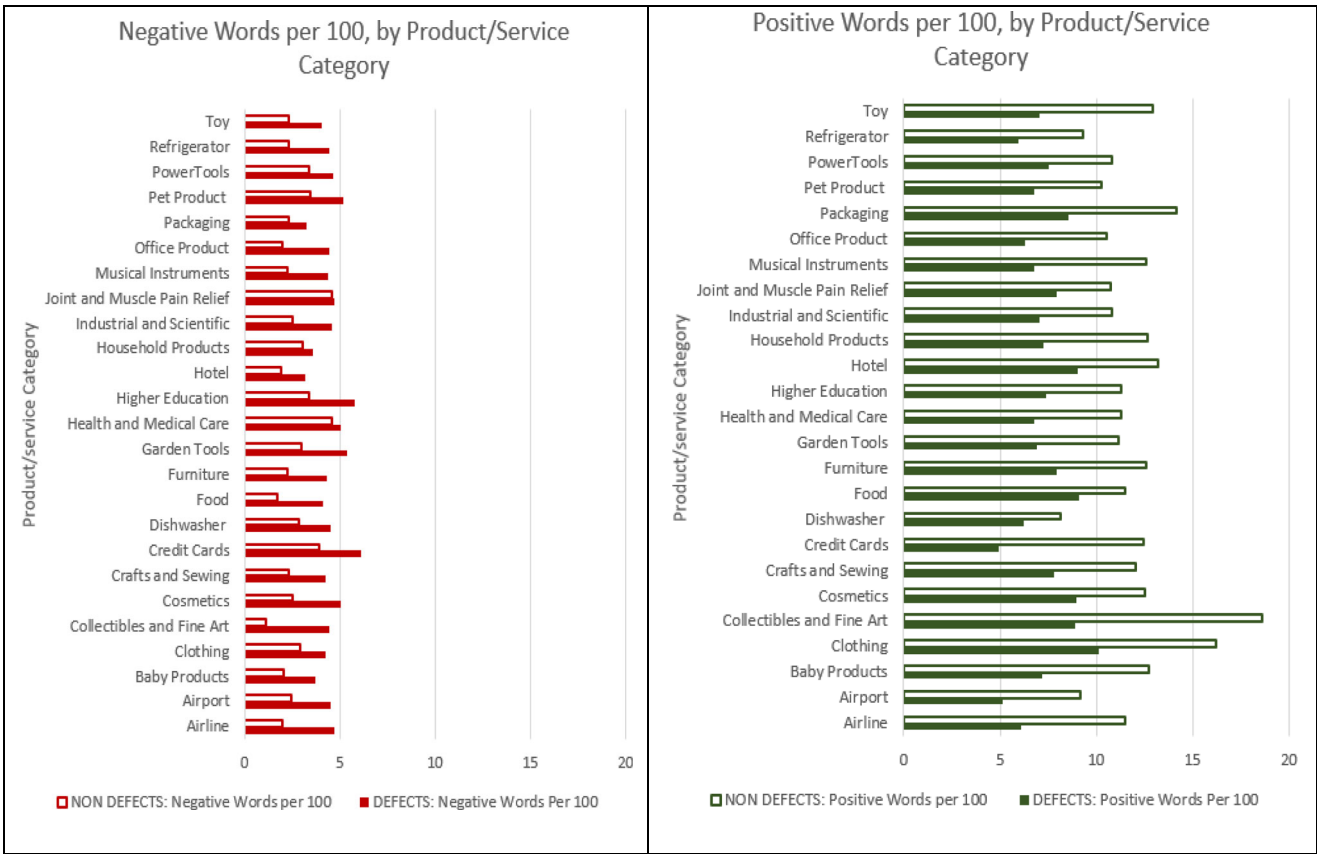

Fig. 6 **a** Negative sentiment words per 100 words for defects versus non-defects by product (or service) category. **b** Positive sentiment words per 100 words for defects versus non-defects by product (or service) category

Table 6 Logistic regression analyses predicting defect existence by product category

Product (or service) category	Parameter estimates					Model metrics		
	Intercept	Inverse star ratings	Best sentiment method (Standardized AFINN negative)	Best generic smoke term method (Standardized bigrams)	Category- specific smoke method (Standardized unigrams)	Area under curve AUC	Lack of fit	Whole model test
Airline	**−3.23****	**4.08****	**16.37***	**−21.32****	**21.01****	0.90	0.88	**
Airport	**−4.51****	**5.01****	5.83	−10.04	**38.98****	0.89	0.94	**
Baby Products	**−4.57****	**9.68****	0.08	13.84	**28.36****	0.95	1.00	**
Clothing	**−4.00****	**9.07****	**−17.86***	−3.56	**42.57****	0.92	0.99	**
Collectibles and Fine Art	**−7.15****	**15.34****	19.77	4.43	**101.76****	0.95	1.00	**
Cosmetics	**−3.92****	**7.76****	−0.01	−14.78	**33.65****	0.93	0.98	**
Craft and Sewing	**−8.21****	**21.48****	**13.91***	−24.92	**57.48****	0.98	1.00	**
Credit Cards	**−7.07****	**12.90****	−6.55	−13.08	4.93	0.98	1.00	**
Dishwashers	**−3.73****	**6.72****	−2.76	**−20.50****	**22.22****	0.93	1.00	**
Food	**−3.85****	**8.23****	2.46	−17.11	**28.43****	0.91	0.94	**
Furniture	**−3.97****	**8.85****	−0.97	−1.28	13.46	0.93	0.99	**
Garden Tools	**−4.51****	**10.22****	−5.40	−4.37	13.27	0.96	1.00	**
Health and Medical Care	**−3.14****	**6.59****	−8.52	−6.54	**32.31****	0.91	0.91	**
Higher Education	**−3.70****	**6.98****	−2.12	−2.70	**29.10****	0.90	0.99	**
Hotels	**−3.28****	**7.15****	−1.09	−11.59	**19.36****	0.88	0.70	**
Household Products	**−6.06****	**13.26****	−12.69	**−36.01***	**85.22****	0.98	1.00	**
Industrial and Scientific	**−7.64****	**19.01****	11.93	−14.38	36.93	0.97	1.00	**
Joint and Muscle Pain Relief	**−3.78****	**6.73****	−10.00	−12.06	**55.50****	0.93	0.99	**
Musical Instrumen-ts	**−5.95****	**12.90****	5.96	−9.22	**30.19****	0.97	1.00	**
Office Products	**−4.85****	**8.87****	1.55	0.80	**23.36***	0.96	1.00	**
Packaging	**−3.01****	**8.55****	−9.13	−25.48	**40.29****	0.87	0.77	**
Pet Products	**−5.04****	**10.09****	−10.93	14.97	**23.33****	0.95	1.00	**
Power Tools	**−3.44****	**7.45****	−9.87	−1.65	**38.68****	0.91	0.89	**
Refrigerators	**−3.95****	**8.00****	5.42	9.75	−7.98	0.93	1.00	**
Toys	**−4.02****	**8.52****	−1.25	−14.26	**31.16****	0.92	1.00	**

*Indicates significance at 95% confidence, **Indicates significance at 99% confidence

service) categories (Abrahams et al., 2013, 2015; ; Adams et al., 2017; Goldberg & Abrahams, 2018; Law et al., 2017; Winkler et al., 2016). We also found that sentiment is not especially effective in finding defect-related discussion. The clear weakness of universal cross-product or service analytics suggests that product or service-focused smoke term list discovery should be a high priority for ongoing defect discovery research.

4.4 Defect Source Analysis

Our work, as discussed in prior sections, has focused on the existence of defect mentions in online reviews. In this subsection, we perform a finer-grained analysis of the causes of the defects identified in these reviews. As the reviews cover a wide range of product (or service) categories, we expect this analysis to demonstrate a broad cross-section reflecting the diversity of categories studied. From our dataset of online reviews, we randomly sampled a segment for further tagging analysis. Tagging for the source of each defect was "emergent" in that we did not begin with a defined tagging protocol, but rather created new categories as they became identifiable. If a review fit into an existing category, then it was tagged as such, and if it did not, then a new category was created. In total, we were able to successfully categorize the source of the defect for 12,653 reviews; within these reviews, we identified 65 unique types of consumer complaints, 63 of which occurred more than once.

Tagging was performed in a tiered structure. In an initial round of tagging, the 12,653 reviews were analyzed and categorized; then, one coauthor verified the coherence of the categories identified and consolidated similar categories; finally, a different coauthor blindly tagged an overlapping set of 200 reviews based on the consolidated categories to ensure reliability. A Cohen's κ value of 0.84 was observed, reflecting "excellent" agreement (Landis & Koch, 1977). In Table 7, we show the top 10 most frequent defect sources identified in our dataset. The top 10 most frequent defect sources account for nearly 86% of defects examined. Products that did not function as intended and safety complains were by far the most frequently occurring categories, after which we observed a steep drop in frequency. Interestingly, of the 10 most frequent defect sources, only one, shipping complaints (the tenth most frequent), is somewhat outside of the firm's control; the remaining categories are all within the firm's control. In addition, we note that customer service complaints were only identified as defects for service categories, such as airlines or higher education.

Table 7 Most frequently observed defect categories

Defect category	Count	Percentage
Did not function as intended	5457	43.1%
Safety complaint	3316	26.2%
Multiple complaints	707	5.6%
Material quality complaint	287	2.3%
Defective component	216	1.7%
Customer service complaint	195	1.5%
Product durability complaint	185	1.5%
Inaccurate product description	182	1.4%
Design flaw	167	1.3%
Shipping complaint/product damaged on arrival	151	1.2%
Other	1790	14.1%

5 Discussion

The enormous volume of online consumer-generated text has created opportunities and challenges for quality management analytics. Sentiment analysis is one of the most active research areas for mining online reviews for insights. This study examined the extent to which sentiment analysis, unsupervised techniques, and supervised techniques can be effectively deployed to identify product defects from online reviews. We found the unsupervised techniques we implemented to be very ineffective at locating defects. We discovered that negative sentiment and generic supervised defect terms are mildly predictive of product defects. Finally, category-specific supervised defect terms were highly effective and seemed to pinpoint defects quite reliably.

5.1 Implications

Our findings have several implications for both research and practice. First, our cross-category analysis provides context as to the general prevalence of defects across 25 product (or service) categories. We found substantial differences in the prevalence of defects across categories, suggesting no one-size-fits-all approach. Additionally, star ratings are limited as predictors, as much of the defect-related discussion was actually found in high-star reviews. These findings suggest the need for nuanced text analytics to pinpoint these discussions. Furthermore, our findings assist managers in each of these categories in understanding the underlying rate of defects in their category, which provides a benchmark for comparing their products with the industry. Second, text analytic research often applies sentiment analysis, but our findings suggest that, at least in the domain of defect detection, sentiment analysis in not as effective as more nuanced supervised learning techniques. Prior studies (Abrahams et al., 2012, 2013, 2015; Adams et al., 2017; Goldberg & Abrahams, 2018; Law et al., 2017; Winkler et al., 2016) have shown that product or service-category-specific smoke word lists are effective in finding defects, but a cross-category analysis provides empirical verification that these category-specific terms are indeed higher performing than the alternative. We found that these terms perform quite well when concentrated to a specific category of focus. Further research in this area should focus on product (or service) category-specific studies for additional product (or specific) categories to develop additional category-specific smoke term lists.

Our results have valuable implications for practitioners. We found that online reviews are a substantial source of

defect-related feedback across a broad cross-section of categories. However, we also found substantial variability across categories in the proportion of reviews that pertain to defects and to the manner that defect-related discussion manifests. One way to examine variability is to consider the difference between search goods and experience goods. In general, we found that search goods had relatively lower prevalence of defects mentioned in reviews, whereas experience goods, such as airlines, airports, higher education, or hotels, had far higher proportions of defect-related discussion. This is a novel empirical finding, and future research could further explore the potential psychological underpinnings of the phenomenon. Interestingly, we did not observe any meaningful difference in the performance of the text analytics techniques for identifying defects in these experience goods (see Table 7). Thus, it seems that although the defects reports are more common in experience goods, the linguistic properties of these reports are such that they are not meaningfully more difficult to predict or identify.

In addition, the proportion of defects found in some experience goods emphasizes the need for our defect surveillance tools. Consider, for instance, the health and medical care product category, where 10% of reviews refer to defects. Examining these reviews is high-value for firms as they seek to innovate and improve their products, but it is also difficult because the reviews are relatively uncommon. In addition, 1-star and 2-star reviews contain a similar number of defects to 3-star, 4-star, and 5-star reviews, so sorting by star rating only provides limited utility. Utilizing text analytics approaches as suggested in this work is immensely promising for such as use case, as reviews can be sorted to expedite these analyses and target remediation efforts.

While sentiment analysis continues to be a valuable tool for gauging the customer's emotive opinion about products or services, it is not the most efficient way of identifying potential defects across multiple product (or service) categories. The inferior performance of the broad list of prevalent terms in defect discovery across product (or service) categories suggests that practitioners should not use broad cross-product (or service) category prevalent term lists but should rather consider developing or deploying product or service-category specific quality defect terms. Firms should continue to hire analysts who can codify reviews in their product (or service) category into defects versus non-defects and build context-specific term lists. Category-specific term lists can be employed to reduce defect discovery time, cost, and reviewer fatigue.

5.2 Limitations and Future Work

We noticed that in some product (or service) categories, customers' reviews were predominantly cons or predominantly pros, whereas in other product (or service) categories (e.g., airlines), customers sometimes reported pros in addition to cons within a single online review. For these product (or service) categories, we noticed the anomalous occurrence of positive words and unexpected phrases appearing among the most prevalent terms for defects. This is most likely because reviews were tagged for defect existence at the full post level (rather than at the sentence- or phrase-level), and a positive word or phrase may have coincidentally appeared in a review that was marked as containing a defect. This issue can be mitigated in future work by tagging only the text snippet specifically describing the defect from each review rather than analyzing the full review text, and subsequently running analyses using the defect-reporting snippets only. A downside of using snippets only is that items genuinely concurrent with the reporting of negative issues will not be as easily identified.

To avoid biasing our results with respect to certain categories, we balanced the sample size and partition of defect versus non-defect reviews for each category. This step ensured that no single category or classification had an exaggerated effect on the results due to a larger sample size. However, doing so limited the sample size available for our ensuing analyses, and dividing the sample into a training set and a holdout set induced a further constriction. In future work, we aim to experiment with instance weighting to improve the workable sample size for each category and classification. This would involve resampling from categories from which few defect reports were observed to improve the effective sample size. As this technique may improve the performance of the smoke terms, we expect that it would serve to reinforce the dominance of smoke terms over sentiment.

This study utilizes online reviews as a means of monitoring consumer feedback on a wide range of product (or service) categories; in doing so, we assume that online reviews are of sufficient quality that they indicate consumers' experience with the relevant products or services. Prior work has argued that online reviews represent a skewed sample of consumers due to a bias toward extreme experiences; that is, consumers with particularly positive or particularly negative experiences are most likely to write online reviews, while consumers with more middling experiences often do not have sufficient motivation to put in the effort (Hu et al., 2009). For our purposes, however, we argue that it is more important for online reviews to be indicative than fully representative (Abrahams et al., 2015). That is, while the reviews may not offer a totally unbiased sample of all consumers, they may serve as evidence of experiences that consumers have had with product (or service) defects that are highly relevant to firms and other interested parties. These reports should serve as signals to firms that they may need to investigate a product (or service) offering, and we suggest that they do so using a multi-faceted approach, which may also include warranty reports, consumer surveys, and other internal testing (Abrahams et al., 2015).

Table 8 Contingency table for CC score algorithm (adapted from Fan et al. (2005))

	Document is relevant	Document is non-relevant	Row total
Document contains term	A	B	$A+B$
Document does not contain term	C	D	$C+D$
Column total	$A+C$	$B+D$	N

A further limitation of our work is that, while we seek to identify defects, we have only begun to categorize them in our final round of emergent tagging. Due to sample size limitations, we were unable to construct machine learning models to subclassify by defect source; however, this would be a very useful capability in practice, and we suggest it as a research area for future work.

Finally, as the ultimate objective of defect discovery from online reviews is improving the efficiency and efficacy of the QM process, future research will need to develop structured procedures to *triage* (quantify and prioritize) the identified risks. Once systematic and product or service-specific text analytic procedures are in place to identify defects, rigorous triage mechanisms for identified issues can become the focus of attention for QM researchers and practitioners. We expect this will continue to lead QM research and practice beyond a sentiment-driven response to defect discovery and into more objective and structured mechanisms for listening and responding to online posts relating to product quality.

Appendix

To curate smoke terms, we utilize the CC score algorithm (Fan et al., 2005), an information retrieval technique that prior works have shown to be quite effective in smoke term curation (Abrahams et al., 2012, 2013, 2015; Goldberg & Abrahams, 2018). As the chi-squared distribution is used to test for the independence between two variables in statistics, information retrieval has utilized this principle to examine relationships between documents and words that they contain. Ng et al. (1997) first suggested a means of using a one-sided chi-squared test to select words or phrases associated with a relevant classification of documents; Fan et al. (2005) later expanded upon this technique. The CC score algorithm generates a relevance score for each term (word or phrase) in a corpus, where higher scores indicate more relevant terms that may be predictive of the target classification. We first distinguish between relevant documents (in our study, reviews) from the target classification (in our study, defect-related reviews) and non-relevant documents not from the target classification (in our study, non-defect-related reviews). Consider Table 8, which defines the relationships between document relevance and inclusion/exclusion of terms.

Given this contingency table, terms are given higher scores when they are especially frequent in documents that are relevant and especially infrequent in documents that are irrelevant. The CC score algorithm defines this relevance as follows for each term in the corpus:

$$Relevance = \frac{\sqrt{N} \times (AD - CB)}{\sqrt{(A+B) \times (C+D)}} \quad (1)$$

The CC score algorithm generates a relevance score for each term in the corpus such that scores with high relevance scores occur frequently in relevant documents and infrequently in irrelevant documents. Thus, we may use high-scoring terms as predictors of relevance in unseen documents. After using the CC score algorithm to generate relevance scores for each smoke term, the lead author analyzed the top-ranking terms to remove any stop words (common English words like "a," "an," and "the"), common brands names, and/or common product (or service) categories (Abrahams et al., 2012, 2013, 2015). A coauthor further reviewed and reverified these decisions to ensure accuracy. The retained terms are referred to as smoke terms, and each set of smoke terms is referred to as a smoke term list.

When analyzing unseen reviews (e.g., our holdout set), we use the appropriate smoke term list to generate "smoke scores" for each review. For a given review, we determine this smoke score by searching for any occurrences of the smoke terms within that review. Each time we observe an occurrence of a smoke term, we increment that review's smoke score by that smoke term's relevance score as determined by the CC score algorithm. Finally, using these smoke scores, we can prioritize the reviews believed to refer to defects. We can sort the reviews from the highest smoke score to the lowest smoke score, where the highest smoke scores are the most likely to refer to defects.

Acknowledgments Alan S. Abrahams and Peter Ractham gratefully acknowledge support for this work from Thammasat University in the form of the Bualuang ASEAN Fellowship.

References

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems, 54*(1), 87–97.

Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., & Zhang, Z. (2013). What's buzzing in the blizzard of buzz? Automotive component

isolation in social media postings. *Decision Support Systems, 55*(4), 871–882.

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management, 24*(6), 975–990.

Adams, D. Z., Gruss, R., & Abrahams, A. S. (2017). Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International Journal of Medical Informatics, 100*, 108–120.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Brahma, A., Goldberg, D. M., Zaman, N., & Aloiso, M. (2021). Automated mortgage origination delay detection from textual conversations. *Decision Support Systems, 140*, 113433.

Chen, Y., Ganesan, S., & Liu, Y. (2009). Does a firm's product-recall strategy affect its financial value? An examination of strategic alternatives during product-harm crises. *Journal of Marketing, 73*(6), 214–226.

Chong, A. Y. L., Khong, K. W., Ma, T., McCabe, S., & Wang, Y. (2018). Analyzing key influences of tourists' acceptance of online reviews in travel decisions. *Internet Research, 28*, 564–586.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.

Cu, T., Schneider, H., & Van Scotter, J. (2017). How does sentiment content of product reviews make diffusion different? *Journal of Computer Information Systems*, 1–9.

Cui, G., Lui, H.-K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce, 17*(1), 39–58.

Das, A. S., Mehta, S., & Subramaniam, L. V. (2017). AnnoFin–A hybrid algorithm to annotate financial text. *Expert Systems with Applications, 88*, 270–275.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Deming, W. E., & Edwards, D. W. (1982). *Quality, productivity, and competitive position* (Vol. 183). Cambridge, MA: Massachusetts Institute of Technology, Center for advanced engineering study.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*.

Duan, W., Gu, B., & Whinston, A. (2008). Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems, 45*(4), 1007–1016.

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2014). Assessing box office performance using movie scripts: A kernel-based approach. *IEEE Transactions on Knowledge and Data Engineering, 26*(11), 2639–2648.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM, 57*(6), 74–81.

Fan, W., Gordon, M. D., & Pathak, P. (2005). Effective profiling of consumer information retrieval needs: A unified framework and empirical comparison. *Decision Support Systems, 40*(2), 213–233.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. Hoboken: Wiley.

Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American customer satisfaction index: Nature, purpose, and findings. *The Journal of Marketing, 60*, 7–18.

Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. *Journal of Management Information Systems, 33*(4), 1034–1058.

Goldberg, D. M., & Abrahams, A. S. (2018). A Tabu search heuristic for smoke term curation in safety defect discovery. *Decision Support Systems, 105*, 52–65.

Goldberg, D. M., Khan, S., Zaman, N., Gruss, R. J., & Abrahams, A. S. (2021). Text mining approaches for postmarket food safety surveillance using online media. *Risk Analysis.*

Gopal, R., Marsden, J. R., & Vanthienen, J. (2011). Information mining—Reflections on recent advancements and the road ahead in data, text, and media mining. In: Elsevier.

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

He, W., Tian, X., Chen, Y., & Chong, D. (2016). Actionable social media competitive analytics for understanding customer experiences. *Journal of Computer Information Systems, 56*(2), 145–155.

Hendricks, K. B., & Singhal, V. R. (1997). Does implementing an effective TQM program actually improve operating performance? Empirical evidence from firms that have won quality awards. *Management Science, 43*(9), 1258–1274.

Hendricks, K. B., & Singhal, V. R. (2001). The long-run stock price performance of firms with effective TQM programs. *Management Science, 47*(3), 359–368.

Holton, C. (2009). Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems, 46*(4), 853–864.

Hora, M., Bapuji, H., & Roth, A. V. (2011). Safety hazard and time to recall: The role of recall strategy, product defect type, and supply chain player in the US toy industry. *Journal of Operations Management, 29*(7–8), 766–777.

Hu, N., Pavlou, P. A., & Zhang, J. (2006). *Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication.* Paper presented at the proceedings of the 7th ACM Conference on Electronic Commerce.

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

Hu, N., Pavlou, P. A., & Zhang, J. J. (2009). Why do online product reviews have a J-shaped distribution? Overcoming biases in online word-of-mouth communication. *Communications of the ACM, 52*(10), 144–147.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems, 52*(3), 674–684.

Hu, N., Koh, N. S., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision Support Systems, 57*, 42–53.

Jarrell, G., & Peltzman, S. (1985). The impact of product recalls on the wealth of sellers. *Journal of Political Economy, 93*(3), 512–536.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446.

Jung, Y., & Suh, Y. (2019). Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. *Decision Support Systems., 123*, 113074.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lau, R. Y., Li, C., & Liao, S. S. (2014). Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems, 65*, 80–94.

Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications, 67*, 84–94.

Lee, J., Park, D.-H., & Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic Commerce Research and Applications, 7*(3), 341–352.

Lee, S., Song, J., & Kim, Y. (2010). An empirical comparison of four text mining methods. *Journal of Computer Information Systems, 51*(1), 1–10.

Liu, Y., Jiang, C., & Zhao, H. (2018). Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums. *Decision Support Systems, 105*, 1–12.

Lyles, M. A., Flynn, B. B., & Frohlich, M. T. (2008). All supply chains don't flow through: Understanding supply chain issues in product recalls. *Management and Organization Review, 4*(2), 167–182.

McAuley, J., Pandey, R., & Leskovec, J. (2015). *Inferring networks of substitutable and complementary products.* Paper presented at the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*.

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications, 42*(3), 1314–1324.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241–4251.

Mummalaneni, V., Gruss, R., Goldberg, D. M., Ehsani, J. P., & Abrahams, A. S. (2018). Social media analytics for quality surveillance and safety hazard detection in baby cribs. *Safety Science, 104*, 260–268.

Ng, H. T., Goh, W. B., & Low, K. L. (1997). *Feature selection, perceptron learning, and a usability case study for text categorization.* Paper presented at the 20th annual international ACM SIGIR conference on Research and Development in information retrieval.

Nielsen, F. Å. (2011). *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. Paper presented at the 1st Workshop on Making Sense of Microposts.

Oberreuter, G., & VeláSquez, J. D. (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications, 40*(9), 3756–3763.

Park, C., & Lee, T. M. (2009). Information direction, website reputation and eWOM effect: A moderating role of product type. *Journal of Business Research, 62*(1), 61–67.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Paper presented at the Conference on Empirical Methods in Natural Language Processing.

Phillips, P., Zigan, K., Silva, M. M. S., & Schegg, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism Management, 50*, 130–141.

Porter, M. E., & Van der Linde, C. (1995). Toward a new conception of the environment-competitiveness relationship. *Journal of Economic Perspectives, 9*(4), 97–118.

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

Qiao, Z., Zhang, X., Zhou, M., Wang, G. A., & Fan, W. (2017). *A domain oriented LDA model for mining product defects from online customer reviews.* Paper presented at the 50th Hawaii International Conference on System Sciences.

Rhee, M., & Haunschild, P. R. (2006). The liability of good reputation: A study of product recalls in the US automobile industry. *Organization Science, 17*(1), 101–117.

Shi, D., Guan, J., Zurada, J., & Manikas, A. (2017). A data-mining approach to identification of risk factors in safety management systems. *Journal of Management Information Systems, 34*(4), 1054–1081.

Stern, H. (1962). The significance of impulse buying today. *The Journal of Marketing, 26*, 59–62.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544–2558.

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

Winkler, M., Abrahams, A. S., Gruss, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision Support Systems, 90*, 23–32.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919–926.

Zaman, N., Goldberg, D. M., Abrahams, A. S., & Essig, R. A. (2020). Facebook hospital reviews: Automated service quality detection and relationships with patient satisfaction. *Decision Sciences.*

Zhang, Z. (2008). Mining relational data from text: From strictly supervised to weakly supervised learning. *Information Systems, 33*(3), 300–314.

Zhao, W. X., Jiang, J., Yan, H., & Li, X. (2010). *Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid.* Paper presented at the 2010 Conference on Empirical Methods in Natural Language Processing.

Nohel Zaman is an Assistant Professor of Information Systems and Business Analytics in the College of Business Administration at Loyola Marymount University. His research interests fall under the areas of business analytics, text mining, safety concerns of consumer products, and quality services. He received his PhD in Business Information Technology from Virginia Tech. Additionally, he earned a BSc degree in Business Administration along with a MSc degree in Economics from the University of Texas at Dallas and a MSc degree in Computational Science and Engineering from the North Carolina A&T State University. His work has recently been published in *Decision Sciences*, *Decision Support Systems*, and *Risk Analysis*.

David M. Goldberg is an Assistant Professor of Management Information Systems in the Fowler College of Business at San Diego State University. He received his doctoral and bachelor's degrees from Virginia Tech. His current research interests are in the areas of text mining, machine learning, decision support systems, and expert systems. He has published in *Decision Support Systems*, *Decision Sciences*, *Risk Analysis*, *Information Technology & Management*, *Communications of the Association for Information Systems*, and others.

Richard J. Gruss is an Assistant Professor in the College of Business and Economics at Radford University. His research interests include social media and text analytics, and his work has appeared in *Decision Support Systems*, *Journal of the Association for Information Science and Technology*, *Expert Systems with Applications*, and *Journal of Hospitality and Tourism Research*.

Alan S. Abrahams is an Associate Professor in the Department of Business Information Technology at Virginia Tech and a member of the Affiliated Faculty at the Center for Injury Research and Policy at the Johns Hopkins Bloomberg School of Public Health. His research on quality analytics from text is published in *Production and Operations Management*, *Decision Support Systems*, and other high-impact journals. He holds a PhD in Computer Science from the University of Cambridge, and a Bachelor of Business Science degree in Information Systems from the University of Cape Town.

Siriporn Srisawas received her Master's degree in Management Information Systems from Thammasat University. Currently, her research interests are social media, knowledge management and e-learning.

Peter Ractham is an Associate Professor in the Department of Management Information Systems at Thammasat University. His research focuses on ICT-enabled innovation, social media analytics and e-business. He has published in *Journal of the Association for Information Systems*, *Information Systems Journal*, and *Computers in Human Behavior* amongst other information systems and business journals. He holds a PhD in Information Systems & Technology from Claremont Graduate University.

Michelle M.H. Seref received a Ph.D. in Operations Management in 2009 from the University of Florida, where she also earned Master's and Bachelor's degrees in Industrial and Systems Engineering. As part of her Master's research, she co-authored the textbook *Developing Decision Support Systems* using Excel and VBA. Her Ph.D. dissertation focuses on Supply Chain Management Decisions under various marketing strategies, including advance sales inventory management and new product development pricing and timing decisions. She has published related research articles in top journals such as the *European Journal of Operations Research* and *Computational Management*.