POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT

Vol. 24, No. 6, June 2015, pp. 975–990
ISSN 1059-1478 | EISSN 1937-5956 | 15 | 2406 | 0975
DOI 10.1111/poms.12303

An Integrated Text Analytic Framework for Product Defect Discovery

Alan S. Abrahams
Business Information Technology Department, Virginia Tech, 1007 Pamplin Hall, Blacksburg, Virginia 24061, USA, abra@vt.edu

Weiguo Fan
Accounting and Information Systems Department, Virginia Tech, 3007 Pamplin Hall, Blacksburg, Virginia 24061, USA
School of Information Engineering and Management, Shanghai University of Finance and Economics, wfan@vt.edu

G. Alan Wang
Business Information Technology Department, Virginia Tech, 1007 Pamplin Hall, Blacksburg, Virginia 24061, USA, alanwang@vt.edu

Zhongju (John) Zhang
Operations and Information Management Department, School of Business, University of Connecticut, 2100 Hillside Road, Unit 1041, 379, Storrs, Connecticut 06269, USA, john.zhang@business.uconn.edu

Jian Jiao
Microsoft, Bellevue, Washington 98004, USA, Jian.Jiao@microsoft.com

The recent surge in the usage of social media has created an enormous amount of user-generated content (UGC). While there are streams of research that seek to mine UGC, these research studies seldom tackle analysis of this textual content from a quality management perspective. In this study, we synthesize existing research studies on text mining and propose an integrated text analytic framework for product defect discovery. The framework effectively leverages rich social media content and quantifies the text using various automatically extracted signal cues. These extracted signal cues can then be used as modeling inputs for product defect discovery. We showcase the usefulness of the framework by performing product defect discovery using UGC in both the automotive and the consumer electronics domains. We use principal component analysis and logistic regression to produce a multivariate explanatory analysis relating defects to quantitative measures derived from text. For our samples, we find that a selection of distinctive terms, product features, and semantic factors are strong indicators of defects, whereas stylistic, social, and sentiment features are not. For high sales volume products, we demonstrate that significant corporate value is derivable from a reduction in defect discovery time and consequently defective product units in circulation.

Key words: social media analytics; quality management
History: Received: August 2013; Accepted: September 2014 by Cheryl Gaimon, after 2 revisions.

1. Introduction

In recent years, the usage of social media such as Facebook, Twitter, Blogger, and various online forums has surged. With this consumer growth, companies have started to explore various ways of using social media. For instance, companies are creating social media campaigns that seek to boost their brand equity by connecting with their customers, and engaging in customer relationship management by addressing issues raised by users from different social media channels (Woolridge 2011). The surge in usage of social media has led to an explosion of growth in volume, velocity, and variety of data and the arrival of the Big Data age (Lohr 2012). Big data analytics that aim to process, analyze, and uncover hidden value from these data is quickly becoming one of the top priorities for many executives. While structured data are easier to manage and more understandable, most of the big data surge is caused by unstructured data (e.g., comments, texts, images, videos) that cannot be readily represented by formal data models in traditional databases (Lohr 2012). Moreover, it remains a daunting challenge to effectively and efficiently sift through the vast trove of unstructured social media content to glean knowledge as most common techniques (e.g., natural language processing, pattern recognition, machine learning) for dealing with

unstructured data involve inter alia, filtering, semantic content grouping, tagging, and information mining (Fan et al. 2006, Gopal et al. 2011, Lim et al. 2013). This has also created opportunities for a new wave of commercial software platforms for tracking social media, such as Google Trends, Nielsen BuzzMetrics, TweetDeck, Collective Intellect, Clarabridge, and others.

While customer relationship management, and specifically complaint management, in social media has attracted some attention, few studies have investigated defect management. Importantly, defects may be reported in factual tone, and in public conversations not directly addressed at the manufacturer. In this study, we propose an integrated text-based analytic framework called SMART (Social Media Analytic fRamework using Text) to quantify social media content and extract important features from such postings that can be subsequently used to discover and analyze product defects. We showcase the framework by analyzing user postings from both automotive and consumer electronics discussion forums to identify and predict product defects. Our empirical results demonstrate that the proposed framework can serve as a basis for quality management (QM) professionals, for analyzing unstructured user-generated content in social media.

This study makes the following contributions relative to the most closely related earlier work in defect detection from social media (Abrahams et al. 2012, 2013b). Firstly, we develop a generic Social Media Analytics framework that can detect product defects from social media postings in multiple domains. Prior work considers only the single narrow domain of vehicle defects. We test the new framework in two diverse domains—vehicles and consumer electronics. For the consumer electronic domain, we report, for the first time, smoke words for the iPod Classic portable media player, that are indicative of postings that describe potential defects with the device. We demonstrate the broad applicability and external validity of our new framework. Secondly, this study provides a comprehensive design of possible signal cues or features that can be extracted from the user postings. It incorporates methodological innovations in data acquisition and transformation as we consider seven categories of input cues (lexical, stylistic, social, sentiment, distinctive terms, product features, and semantic cues), while earlier work considers only two (style and sentiment cues). With the enhanced generality of the new framework, we are able to employ compound factor analyses using Principle Component Analysis, and are able to recognize and classify cues into context-independent vs. context-specific variables. Thirdly, this study performs a multivariate analysis that establishes the relative importance of the features extracted from social media in predicting product defects, which was overlooked in prior work. Prior work was restricted to univariate consideration of a restricted set of input variables. Fourth, we have added new baselines to show the relative advantages of our new framework over existing text analytical approaches. Finally, we provide a comprehensive discussion on the business value and organizational implementation of text analytics in product quality issue discovery, as well as limitations and biases, which was omitted from prior work.

The rest of the study is organized as follows. In section 2, we begin with a discussion of the business value of automated defect discovery from social media. After reviewing relevant literature on defect management and text classification in section 3, we present an integrated defect discovery framework that considers a comprehensive assortment of first- and second-order input features for text analytics in section 4. Our integrated framework establishes the connection between those input features and product defects. In section 5, we test our framework using case studies from two domains: automotives, and consumer electronics. In each case, we describe our research setting, data, methodology, and results. We review both third party discussion forums, and forums administered by the manufacturer. The automotive discussion forums, which are the subject of our vehicle case study, are run by third parties that are independent from the manufacturer. The Apple Support Community, which is the subject of our consumer electronics (iPod) case study is administered by the manufacturer (Apple). Section 6 describes limitations of our study, particularly statistical and human biases in the source data and defect management process, and section 7 concludes and discusses directions for future research.

2. The Business Value of Automated Defect Discovery

Product design quality encompasses both performance and conformance attributes of the product, and is an important aspect of product competitive advantage. Increased marketing–manufacturing integration is associated with greater product competitive advantage, and in turn, greater product success and higher return on investment (Swink and Song 2007).

Product defects can be costly to corporations for a number of reasons. In some industries, such as the motor vehicle industry, defect rectification for every affected product unit may be mandated. General Motors recently estimated the cost of repairing millions of recalled vehicles at $1.3B (Isidore 2014). Litigation may be initiated by some portion of consumers

as a result of the product defect, particularly if the product defect caused death, injury, or failure of the product to meet warranted performance criteria. Furthermore, regulatory agencies may initiate fines in the case of tardy corporate response, as in the recent $1.2B fine imposed by the US Department of Justice, and the $66M fine imposed by the NHTSA, on Toyota (Muller 2014). Finally, brand reputation may suffer, and dissatisfied customers are likely to have lower repurchase intention (Anderson and Sullivan 1993). In each of these cases, the cost of the defect is largely proportional to units sold. This is especially problematic when the number of units sold is very high. For example, in the automotive industry, 82.8 million units were sold in 2013 (Lebeau 2014), averaging more than 220,000 units sold per day. Early discovery of defects reduces the number of defective product units sold, as the production process can be altered to eliminate the defect prior to sale. Top-selling car and light-truck model such as the Toyota Camry, Ford F-Series, and Chevrolet Silverado each sells on average in excess of a thousand vehicles per calendar day (Binder 2013). Hence, a reduction in defect discovery time by as little as 10 days can, for a top-selling model, keep upwards of 10,000 defective units of that single model off the road, representing a significant saving to the manufacturer.

With the rapidly growing volume of Internet postings, keeping pace with consumer postings is increasingly difficult. For example, the three vehicle forums analyzed in this research, comprise over 1.5 million threads. A thread is defined as a series of postings in reply to an original posting. Since follow-up postings are replies to the original posting, all postings in the thread typically relate to the same topic, which is the question or issue raised at the start of the thread. With each thread in our sample averaging 500 words, at an average adult reading and comprehension speed from a computer monitor of 180 words per minute (Ziefle 1998), a total of over 9000 eight-hour person days would be required to exhaustively read all threads in those three forums alone, equating to one work year (227 work days) for a team of 40 full-time automotive experts. If an automated defect discovery technique is able to reliably identify threads that are more likely to describe defects, then the actual defects discovered per thousand threads reviewed can be increased. In turn, the time-to-discovery for each defect can be significantly reduced; fewer defective products will reach consumer's hands; and large savings can consequently accrue to manufacturers.

3. Background

Uncovering value from social media comments involves a systematic process of gathering social media comments, text preprocessing, feature extraction, and advanced text modeling and analytics. The entire process is commonly referred to as text analytics or text mining (Fan et al. 2006), or, if the data are primarily sourced from social media, the task is referred to as social media analytics (Fan and Gordon 2014). In this study, we focus on the latter part of the text mining and social media analytics process. In particular, we focus our attention on how to extract useful text-based features using techniques from natural language processing, linguistic analysis, and information retrieval, and subsequently, how to build explanatory and predictive models using those features to uncover product defects.

We begin with a review of relevant literature on defect management in the operations management field. We then turn to the computer science and information management literature, and review research on feature extraction from text, and on social media analytics.

3.1. Defect Management

The predominance of historic QM research has been done in Process Management and Business Results. Measurement, Analysis, and Knowledge Management is "almost devoid of research that connects (this) topic to the quality field" (Schroeder et al. 2005, p. 475). Collecting relevant information from all phases of an organization's operations, and using this knowledge to monitor and improve quality, is fundamental to TQM (Ahire et al. 1995, p. 283). Knowledge management tools for TQM have typically been restricted to structured information as input to Quality Control tools such as cause and effect diagrams, Pareto diagrams, control charts, check sheets, histograms, stratification, and scatter diagrams (Johannsen 2000). Analysis of unstructured content, specifically from the Internet, from a QM perspective, has received relatively little attention.

Having established unstructured knowledge management for quality control as an important and under-served field of QM research, we now look at mechanisms for text analysis in computer science and information systems.

3.2. Extracting Useful Features from Text

Features, also referred to as "input variables" or "cues", are quantitative metrics that describe an unstructured piece of qualitative (textual) data. Content analysis of text has long been a significant concern of researchers from sociology to business (Bryman and Burgess 2002, Duriau et al. 2007, Neuendorf 2001). More recently, interest in analyzing textual data from the Internet, and specifically social media, has burgeoned (Yang et al. 2002), and a vast array of feature variables for text have been proposed in the

literature. These can be broadly divided into the following major feature categories:

Lexical features. Lexical features such as unique words, phrases, noun phrases, or named entities are identified. Each document is then represented as a vector of occurrence or frequency of these unique terms (Abbasi and Chen 2008, Cao et al. 2011, Coussement and Van Den Poel 2008, Decker and Trusov 2010, Schumaker and Chen 2009, Zhang et al. 2009). These are generally referred to as "Bag-of-Words", "Bag-of-Terms", or "term-by-document matrix" representations.

Stylistic features. Measure number of total words, characters per word, number of unique words, words per sentence, sentences per paragraph, and text readability (Abbasi and Chen 2008, Abbasi et al. 2008, Coussement and Van Den Poel 2008).

Social features. count the number of messages, posts, or comments (Antweiler and Frank 2005, Duan et al. 2008, Zhu and Zhang 2010); number of readers or number of replies about a company, product, or user (Li and Wu 2010); credibility, expertise, or social influence of postings or users (Oh and Sheng 2011, Wang et al. 2011, 2013); or flag postings with date, time, or location (Oh and Sheng 2011, Spangler and Kreulen 2007).

Sentiment features. measure subjectivity, positivity, negativity, or overall user rating, for words, sentences, queries, or entire documents. Positive and negative words may be stored in a word list (Kelly and Stone 1975, Stone et al. 1966), with each word often viewed as context-independent and annotated with a universal sentiment strength in a range, such as +5 to −5 (Nielsen 2011). Domain-specific sentiment measures can be constructed, such as, in finance, the ratio of buy-to-sell messages per stock (Antweiler and Frank 2005). Sentence-level sentiment analysis techniques (Thelwall et al. 2010, 2012) account for contextual modifiers (e.g., "*not* bad", "*wish it was* good"). Some approaches determine polarity orientation toward a given query, topic, subject, or target (Santos 2012, Vechtomova 2010). Message-level opinion polarity computationally assigns a total opinion rating to a message (Abbasi et al. 2008, Loughran and Mcdonald 2011, Romano et al. 2003, Santos 2012) or extracts human user ratings—for example, product reviews on a 5-star scale—directly from each posting (Chevalier and Mayzlin 2006, Zhu and Zhang 2010).

Distinctive terms. Industry-specific dictionaries of distinctive terms for specific text categories have been created, by ranking words by highest relative prevalence between different types of messages. In *finance*, word lists have been created for negative ("bearish") words and positive ("bullish") words (Antweiler and Frank 2005, Das and Chen 2007, Loughran and Mcdonald 2011, Oh and Sheng 2011, Tetlock et al. 2008). In the *movie* industry, word/phrase classes such as "intent-to-see" words/phrases, have been created to gauge their relative frequency and ascertain correlation with box office sales. In the *automotive* domain, distinctive "smoke words" have been shown to occur significantly more frequently in performance or safety defects vs. other postings (Abrahams et al. 2012).

Product features. Any posting can be tagged with structured product feature data, such as a categorical value or numerical value. Product features are defined with tag types and tag values. Tagging can be performed by site users, using facilities provided by the discussion forum or by using third party collaborative tagging (social bookmarking) software such as deli.cio.us, or tagging can be performed retrospectively with automated software algorithms or by a human defect investigator. Each tag type is named and has one or more tag values. For example, the (categorical) tag type "Component" may have the tag values "Brakes", "Wheels", "Suspension", and so forth, or the tag type "Seller Type" may have the values "Original Equipment Manufacturer (OEM)" or "Aftermarket". Numeric tag types such as "Engine Size" may have values such as "50 cc", "150 cc", etc. Tag types correspond roughly to column names, and tag values correspond roughly to cell values in tabular databases. Tags are specifically suited to describing attributes of concepts referred to in textual data as these attributes are hyper-dimensional and sparse (many columns, with many missing values). To deal with hyper-dimensionality in text classification, prior research suggests the use of feature reduction techniques (Sebastiani 2002, Yang and Pedersen 1997) or algorithms specifically suited to hyper-dimensional data such as Support Vector Machines (Joachims 1999, Tong and Koller 2002).

Semantic features. count the frequency of occurrence of concept classes. Generic semantic classes (e.g., animals, humans, vehicles, economic words) can be identified using general purpose dictionaries, such as Harvard General Inquirer (Kelly and Stone 1975, Stone et al. 1966) to map words (or triggered rules) to semantic categories; for example, "car", "automobile", "train", and "plane" map to the class "vehicle").

3.3. Text Classification and Text Analytics for the Web and Social Media

The problem of classification of unstructured text has occupied computer scientists and computation linguistics for decades (Apté et al. 1994, Borko and Bernick 1963, Calvo et al. 2004, Joachims and Sebastiani 2002, Li and Jain 1998, Yang 1999). So-called "supervised" methods are provided with a training

set of tagged (categorized) documents, and attempt to match unseen documents to the correct categories. Popular computational methods for supervised text classification include Bayesian approaches, k-Nearest Neighbor, Decision Trees, Decision Rules, Logistic Regression, Neural Nets, and Support Vector Machines (Lewis et al. 2004, Ruiz and Srinivasan 2002, Sebastiani 2002). While studies have shown that Support Vector Machines and Neural Nets offer high accuracy (Sebastiani 2002), support still exists for more easily interpretable and explanatory approaches like Logistic Regression (Coussement and Van Den Poel 2008, Loughran and Mcdonald 2011, Ma et al. 2011).

The above text categorization methods have been applied to textual data from the Web in a number of ways. We highlight a selection of these applications in Appendix A1 in the Online Supplement. Appendix A1 provides a summary of previous research on text classification of various types of social media content (e.g., product reviews, Internet news articles, public regulatory filings, listservs, emails, newsgroups, blogs, and online discussion forums). The leftmost column of Appendix A1 indicates whether the prior study:

(i) relates to defect discovery, or
(ii) analyzes complaints and/or sentiment, but not specifically defects, or
(iii) directly discusses product defects.

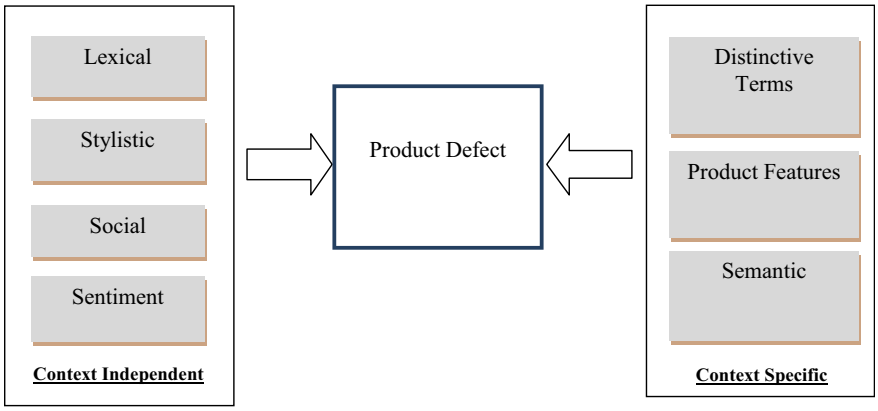
Appendix A1 shows that the vast majority of prior research in social media analytics focuses on analyses that do not relate to defects (item [i]), such as finding or predicting trending topics, detecting spam, ascertaining helpfulness, and so forth. The most closely related field to defect discovery analyzes complaints and sentiment (item [ii]), often with a view to predicting company or product performance (e.g., movement of the stock ticker, or sales of the movie box office release). However, recent research specifically on defects (item [iii]) demonstrates that complaints and sentiment correlate *poorly* with defects, for a number of reasons. Firstly, complaints and sentiment are expressed emotively. Defects, in comparison, are often expressed factually without any emotive content: "I noticed a bubble on my tire tread", "The actuator does not activate", "The airbag inflated while I was driving", "I hear a clicking sound". Where emotive content exists, its magnitude often does not correspond with defect severity: one customer may be more irate about music quality from their speakers (a performance issue) than another customer mumbling about a rattling mirror (a safety issue). Or community members may engage in highly emotive exchanges unrelated to the brand or its products: a recent discussion of the Breaking Bad television series among Honda enthusiasts at Honda-tech.com extended to 217 separate user postings and featured words and phrases such as "kinda stupid", "what the hell", "sucked", "F*** ya", "fantastic", and "awesome"—all completely unrelated to any Honda model. Secondly, complaints are directed: service requests come in at the dealership, complaint emails are sent to the manufacturer, and complaint reports are filed for specific vehicle models at the National Highway Traffic Safety Administration (NHTSA). In contrast, defect reports in social media may be haphazardly interspersed among other content, such as reviews, comparisons, questions, and social discussions, and debates. As many defect reports are not physically sent to the manufacturer, they are especially difficult to uncover in the "blizzard of buzz" (Woolridge 2011). Isolating the component under discussion may be helpful since some components are discussed primarily when defective. For instance, online enthusiasts frequently compare specs on extravagant engines and transmissions, with no mention of defects, whereas they seldom discuss mundane components like seat belts and air conditioning unless there is a specific defect of concern (Abrahams et al. 2013b).

4. SMART: An Integrated Social Media Analytic Framework for Product Defect Discovery

Synthesizing the prior literature in text mining, statistical language processing, and social media analytics, described above, we propose an integrated analytic framework (Figure 1) that links useful quantitative text features (signals) to the underlining existence (or absence) of a product defect, described in the social media content.

The feature categories (gray-shaded boxes) in Figure 1 consist of variables or signal cues of two broad types. *Context-independent* variables include variables that are applicable to all knowledge domains and discovery tasks, and mostly easily extractable or computable. *Context-specific* variables include variables that are specific to a particular knowledge domain and typically require extensive textual analysis (e.g., feature selection), categorization, and processing (e.g., dimension reduction through principal component analysis [PCA]) to extract. To the best of our knowledge, the integrative framework presented here is the first to provide structured guidance to researchers by suggesting a systematic mapping between product defect existence and emerging quantitative textual scores from the nascent literature on computational linguistics and social media analytics. Using the framework, practitioners and researchers can system-

Figure 1 SMART: An Integrated Text Analytic Framework for Product Defect Discovery from Social Media



atically assess which data sources and text scores provide the strongest and/or most robust relationship to product defect existence.

5. Evaluations and Findings

In this section, we evaluate the performance of the integrative defect discovery framework by showcasing two applications of it, in computerized defect detection in the automotive industry, and in the consumer electronics industry. We demonstrate the usefulness of each of the input features (derived from the textual data) in detecting product defects.

5.1. Case Study 1: Defect Discovery in the Automotive Industry

In our first case study, we reviewed user postings from online auto enthusiast discussion forums for three top-selling US automotive brands—Honda, Toyota, and Chevrolet—who together sold more than 4.5 million vehicles in the United States in 2012 (Binder 2013), giving a combined average of more than 12,000 units sold per calendar day.

5.1.1. Data. The full set of all discussion threads, as of June 2010 was crawled and extracted from Honda-Tech.com (1,316,881 threads), ToyotaNation.com (216,559 threads), and ChevroletForum.com (26,936 threads) respectively. In total, 1,560,379 discussion threads were extracted from the three forums. Threads containing less than 50 words or less than two postings were excluded from further analysis as they were found to contain too little information to describe and confirm a defect.

Exploratory analysis revealed an extremely low incidence of defects in the full thread dataset. We subsequently filtered the dataset to identify threads that are most likely to be related to defects. To achieve this, we took a snapshot of the US Department of Transportation, NHTSA, Office of Defect Investigations (ODI), vehicle complaint dataset, as of June 2010. We determined the top 200 component description keywords that were most frequently used in vehicle complaints by complaint filers to this federal agency. We then ascertained the number of hits for this shortlist of complaint keywords in each thread, and sorted the threads from highest use of such complaint keywords to lowest. Finally, we re-balanced the dataset by sampling an equal number of threads (the top 1500 threads containing the most complaint keywords) from each brand (4500 total discussion threads). In the construction of our sample, we used an over-sampling strategy that is appropriate for rare events (Shmueli et al. 2010). If one were to use a pure random sampling approach, the resulting sample would contain an extremely low percentage of vehicle defects. In this case, the best method would be to predict no defect for any discussion thread, and attempts to identify factors leading to defect would have no significance.

The sample dataset of 4500 threads contained 2471 verified defects and 2029 non-defects (roughly balanced)—the defect verification process and the procedure for determining the specific component(s) affected are described in detail in the next subsection. Threads discussed 62 unique model names, with the most common model names discussed being the well-known top sellers for each brand: Honda Accord (909 threads) and Civic (336 threads), Toyota Camry (535), Chevy Silverado (388), and Tahoe & Suburban (322). On average, threads contained 502 words (with standard deviation σ = 508), of which 151 were typically unique in each thread. Threads in the sample were generally heavily read, having a mean number of 492 views (σ = 1077). The high coefficients of variation in number of words and number of views in our sample are indicative of the typical long-tail distribution of these variables: many threads with a low number of words (similarly, views) and few threads with many words (views).

On average, five users (σ = 3.6) contributed to each thread. The threads in the sample discussed all 12 categories from the NHTSA's top-level classifications of vehicle components in filed safety complaints, with the most common components affected being Engine (1815 threads), Electrical (1625), Transmission (663), Structure & Body (319), and Braking (286). Less frequently discussed components were Air Bags (116), Lights (122), Steering (119), Wheels & Tires (98), Visibility (96), Suspension (68), and Seat Belts (61). We further identified two additional component categories—Acoustics (72) and Air Conditioning (261)—which are omitted from the NHTSA's component classification scheme, as the NHTSA is entirely safety-focused and is thus not concerned with component defects that pertain to performance issues. The remaining 321 threads could not be classified under any specific component label. A significant portion of discussion threads (1334 threads; 30% of threads) cited more than one troublesome component. The majority of these multicomponent interactions involved the electrical system (965 threads), and its interaction with the engine (536 threads).

5.1.2. Methodology. Our dependent variable is *Defect*, a Binary variable indicating whether a discussion thread relates to a perceived defect, as perceived by domain experts. Thread, rather than posting, was chosen as our unit of analysis as each posting within the thread can often only be understood within the context of the entire thread. As a posting is frequently unintelligible when removed from the context of a thread, and classifying every individual posting within a thread results in an order of magnitude increase in the number of items requiring review, the full thread (comprising all postings in response to the original posting, and typically dealing with a single topic or issue) represents a sensible and efficient unit of analysis for defect discovery.

We followed an established protocol for assessing whether a thread discussed a defect or not (Abrahams et al. 2012). The sample was tagged and labeled by three mechanical engineering domain experts. Threads were flagged as "defect" if the discussion could be characterized as pertaining to any of Safety Integrity Levels (SILs) 0 through 4, these being a set of standard motor industry controllability categories for classifying operability hazards (Jesty et al. 2000). "Defect" was restricted to cases where the issue occurred during the factory-warranted life of the particular component affected, and where the affected component was an OEM component. Discussions relating to driver error, owner abuse, routine service, non-OEM components, or irrelevant to operability hazards (e.g., "for sale" listings) were classified as non-defects.

Following defect classification by each expert rater, we assessed inter-rater agreement. Inter-rater agreement for defect existence was κ = 0.96, representing "almost perfect" agreement among raters (Landis and Koch 1977).

Our independent (input) variables—a representative and comprehensive, but not exhaustive, selection of cues from the SMART framework—were as follows:

Stylistic information was computed for each thread. We computed Total Words ("Word Count"), Readability ("FogIndex") (Flesch 1951), and Characters per Word ("Avg WordLength").

For *social cues*, we used regular expressions (Friedl 2006)—which are a basic information extraction mechanism (Chang et al. 2003, 2006, Cowie and Wilks 2000)—to extract from the discussion forum web pages: the number of views of a thread ("Num. of Views") and the number of participating users for a thread ("Num. of Users").

For *sentiment cues*, we used automated routines to compute the incidence of various types of positive and negative words, as described in (Abrahams et al. 2012). We then performed PCA on the large number of variables generated. Six major sentiment factors ("SentiFactor1-6") were found.

For *distinctive words*, we used a list of words most indicative of vehicle defects from prior automotive research (Abrahams et al. 2012)—these are so-called "smoke words" for defects. Using the 1000 most indicative smoke words from this list, we counted the incidence (hits per thousand words), in each thread, of any word in the smoke word list.

For *product features*, we focused on the component categories discussed in each thread. These were manually assessed by automotive experts, following the protocol defined in earlier work (Abrahams et al. 2013b). For each component category, we specified, using separate binary variables, whether that component category was discussed in the thread.

Finally, for *semantic cues*, we used the Harvard General Inquirer H4 + Laswell LVD lexicon (Kelly and Stone 1975, Stone et al. 1966), which comprises 11,788 lexical entries, to determine the semantic categories for each word in each thread, for each of 99 semantic categories. Each word was labeled with one or more General Inquirer tags (word categories): for example, "overstatement" words, "legal" words, "failure" words, "tool" words, "economic" words, "persistent" words, and other tags. Harvard General Inquirer also performs basic part-of-speech analysis (e.g., identifying interpretive verbs, descriptive verbs, stative verbs, adjectives, and pronouns). Each posting was scanned and the number of occurrence of these categorized words, per thousand words in the thread, was recorded, for each word category. Given the potential for multicollinearity and redundancy between the

large number of semantic variables in our dataset, we conducted PCA to identify the higher level factors that could be most influential, along with the variable loadings for these factors (Jolliffe 2002). Eight major semantic factors ("SemanFactor1-8") were identified.

Appendix A2 in the Online Supplement summarizes our implementation of the SMART framework on this dataset: showing the category and definition of each variable, and descriptive statistics for the sample set. We first conducted a pair-wise Pearson correlation analysis on all independent variables. All correlation values were below 0.6. We then performed the Variable Inflation Factor (VIF) test on all independent variables against the defect label and all VIF scores were below 2, suggesting that multicollinearity is not a major issue.

Finally, multivariate logistic regression (Agresti 2013) was run, to ascertain the nature of the association between our input variables and the existence (or absence) of defects. Models were run iteratively, first with the most basic set of context-independent variables, then iteratively adding each major variable category, to ascertain the relative improvement in model quality with each new information source. All analyses were completed using IBM SPSS V20.

5.1.3. Results and Discussion. The results of our automotive defect analyses are summarized in the left-hand columns of Table 1 (columns labeled "Automotive Case"). It can be seen that many of the input features derived from the textual data are important in predicting perceived defects. However, Model 1 (the context-independent set of variables) alone explained only about 5% of the variation in product defects with a classification accuracy of 59%. Model 2 (the context-specific set of variables) were able to explain about 27% of the variation with a classification accuracy of 70%. When combining both sets of variables in our analysis—Model 3—the model performance (both R^2 and classification accuracy) improved. The β coefficients for each variable show that smoke words more accurately identified a discussion thread related to a defect in the full model than in the context-specific model (β increased from 0.85–0.94). β coefficients for each variable in the full model (Model 3), show that smoke words as distinctive terms ($\beta = 0.94$) and certain product features (Air Conditioning: $\beta = 1.19$; SeatBelts: $\beta = 1.02$; Braking: $\beta = 0.96$) have the strongest positive association with the probability of the thread discussing a perceived defect. These features are moderately predictive of such defects (classification accuracy = 71.4%; $R^2 = 0.304$), for the sample dataset we studied. Stylistic, social, and sentiment features of the thread are only weakly proportional to or weakly inversely proportional to ($\beta = -0.22$ to 0.18) the probability of the thread discussing a perceived defect, and are poorly predictive of such defects (classification accuracy = 59%; $R^2 = 0.049$), for the sample dataset we studied.

Finally, we compared the classification accuracy using the second-order features in our framework, to traditional text classification baseline methods that use only first-order (unigram/"bag of words"/lexical) features. The first baseline method, Support Vector Machines (SVM), was executed using LibSVM (Chang and Lin 2011). The second baseline text classification method, Naïve Bayes (NB), was executed using Weka (Hall et al. 2009). Four unigram feature ranking methods for selecting significant unigrams were tested as follows: information gain (IG), chi-square (CS), document and relevance correlation (DRC), and Robertson's selection value (RSV) (Fan et al. 2005). For each method, different numbers of top ranked unigrams were tested, in increments of 100, from 100 to 2000, and results are shown only for the optimal number of top ranked unigrams as judged by overall accuracy. Appendix A3 in the Online Supplement shows the results of the 10-fold cross-validation run using the baseline unigram methods, and the comparison to the performance of multivariate logistic regression (logit) with second-order text metrics. F1-measures reported are defined in (Yang 1999). The AUC performance metric is the computation of area under the ROC curve (Bradley 1997). Both F1 [0, 1] and AUC [0, 1] are overall performance measures that take into account both false positives and false negatives. The higher these two values, the better the classification performance.

Appendix A3 indicates that the SMART framework delivers the highest precision, F1, and AUC with minimal loss of recall. Using a χ^2-statistic with 1 degree of freedom, precision on the holdout sample (450 records) is statistically significantly better than the next best method (SVM+CS, 400 unigrams), at the 99.9% confidence level ($p < 0.01$). Given the finite resources that companies have to verify suspected defects, and the exponential growth in discussion forum postings, models with high F1 and AUC are valuable, as these high-quality models allow substantially more valid defects to be manually verified by a human who is cross-checking the model's predicted defects, during a limited verification period.

To address the concern that our comparison of SMART using higher order features with the baselines using only lexical features may not be a fair comparison, we also performed additional robustness checks by feeding the same set of higher order features to the different text classification algorithms. The results indicate that the logistical regression still produced the best model, followed by the SVM and

NB. The logistic model offers the best performance results and also the most interpretable results.

5.2. Case Study 2: Defect Discovery in the Consumer Electronics Industry—the iPod Classic

In our second case study, we reviewed online discussions relating to the iPod Classic, a highly popular portable music player. The iPod Classic, which underwent six product generations, formed the foundation of the Apple iPod product line, that has together sold more than 275 million units (iPod + iTunes timeline 2014), averaging more than 70,000 units per calendar day, over 10 years.

5.2.1. Data. The data source is user postings from Apple's official Support Community forums for the iPod Classic portable music player (http://discussions.apple.com/). The final dataset, crawled as of November 2013, comprised a total of 37,104 threads, of which 2850 were randomly selected for our analysis.

5.2.2. Methodology. Defect tagging, product feature tagging, stylistic, social, sentiment, general semantic cues, and subsequent PCA for factor derivation, were undertaken by following a similar process to that used in Case Study 1. For product features, which are highly domain-specific, we employed the

Table 1 Results of Multivariate Logistic Regressions (Case Studies)

		Automotive case			Consumer electronics case				
Feature category	Variable name	Model 1 (context-independent)	Model 2 (context-specific)	Model 3 (full model)	Model 1 (context-independent)	Model 2 (context-specific)	Model 3 (full model)	Variable name	
Stylistic	Word count	−0.12**		0.04	0.30*		0.69**	Word count	
	Fog index	−0.03		−0.11**	−0.08		−0.07	Fog index	
	Avg WordLength	0.01		−0.13**	−0.26**		−0.10	Avg WordLength	
Social	Number of views	0.17**		0.07	−0.87**		0.30*	Number of posts	
	Number of users	−0.22**		−0.16**	0.75**		−0.82**	Number of users	
Sentiment	SentiFactor1	−0.03		−0.14**	0.25**		−0.19**	SentiFactor1	
	SentiFactor2	−0.14**		0.03	−0.42**		−0.46**	SentiFactor2	
	SentiFactor3	0.18**		0.26**	−0.13**		−0.09	SentiFactor3	
	SentiFactor4	0.10**		0.05	0.03		0.05	SentiFactor4	
	SentiFactor5	−0.07*		−0.01					
	SentiFactor6	0.04		0.13**					
Distinctive	Smoke words		0.85**	0.94**		0.85**	1.05**	Smoke Words	
Product	Air conditioning		1.19**	1.10**		−0.39	−0.33	Display	Hardware
features	Airbag		−0.09	−0.15		0.77*	1.06**	Buttons & keys	
	Braking		0.90**	0.96**		−0.25	−0.21	Storage	
	Electrical system		0.48**	0.57**		−0.36	−0.54	Processor	
	Engine		0.41**	0.43**		−0.23	−0.20	Battery & power	
	Lights		0.17	0.26		0.03	0.06	Audio	
	Seat belts		0.87**	1.02**		−0.90**	−1.13**	AV output	
	Steering		0.95**	0.98**		−2.47**	−2.10**	Protective case	
	Structure & body		0.85**	0.90**		0.04	−0.04	Operating system	Software
	Transmission		0.22*	0.27*		−0.22	−0.27	Music	
	Visibility		0.75**	0.65*		0.14	−0.01	Photos	
	Wheels & tires		0.29	0.29		−0.92**	−0.78**	Video	
	Suspension		0.48	0.57*		−0.28	−0.39*	Connection	
	Acoustics		0.41	0.29		−0.45*	−0.44*	Applications	
	Other		−1.26**	−1.16**					
Semantic	SemanFactor1		0.05	−0.02		−0.29**	−0.26**	SemanFactor1	
	SemanFactor2		−0.15**	−0.15**		−0.20**	0.03	SemanFactor2	
	SemanFactor3		0.05	0.09*		0.44**	0.30**	SemanFactor3	
	SemanFactor4		−0.25**	−0.27		−0.06	−0.13*	SemanFactor4	
	SemanFactor5		0.10**	0.06		0.31**	0.21**	SemanFactor5	
	SemanFactor6		−0.02	0.03		0.04	−0.02	SemanFactor6	
	SemanFactor7		−0.04	−0.05		−0.29**	−0.35**	SemanFactor7	
	SemanFactor8		−0.08*	−0.11**		−0.19**	−0.18**	SemanFactor8	
						−0.37**	−0.30**	SemanFactor9	
						−0.28**	−0.18**	SemanFactor10	
	R^2	0.049	0.274	0.304	0.109	0.420	0.450	R^2	
	Recall (%)	80.2	74.7	76.4	51.8	72.0	73.6	Recall (%)	
	Classification accuracy (%)	58.8	70.2	71.4	62.7	76.4	77.3	Classification accuracy (%)	

Numeric values are β coefficients, except numeric values in the final three rows of the table, which are model overall statistics.
*Indicates significant at 95% confidence, **Indicates significant at 99% confidence.

following procedure. We obtained and thoroughly reviewed the official iPod Classic Technical Specifications document from Apple, and identified several prominent component categories, including both hardware and software components. We then provided the full technical specification and a tagging protocol explanation describing each preliminary component category to four MBA students, who coded the data and updated the component categories as necessary. In the end, we came up with eight hardware component categories (viz. Display, Buttons & Keys, Storage, Processor, Battery & Power, Audio, AV output, Protective Case) and six software categories (viz. Operating System, Music, Photos, Video, Connection, and Applications). We assembled a panel of 14 prior iPod owners and assigned the defect to the component category most frequently chosen by the panel. To verify reliability, an MBA student, also an iPod Classic owner, independently tagged all threads for component category. To verify inter-rater agreement between the panel and the independent expert, Cohen's kappa (κ) was computed ($\kappa = 0.76$), indicating substantial agreement (Landis and Koch 1977). Disagreements were adjudicated by a member of the research team who had previously owned an iPod Classic.

5.2.3. Results and Discussion. The results of our iPod defect analyses are summarized in the right-hand columns of Table 1 (columns labeled "Consumer Electronics Case"). As can be seen from Table 1, the results on iPod are very similar to the results of the automotive defect discovery. Smoke words are again very significant in detecting iPod defects. Some product components of the iPod are very effective in detecting iPod defects as well. Overall, the contextual variables (Model 2) are much more effective in detecting product defects than the context-independent variables (Model 1). The overall detection performance (Model 3) is a little bit higher than the automotive results. The results in Table 1 collectively confirm the usefulness of our SMART framework in identifying threads that have comments or statements about product defects.

A list of the thirty most significant iPod Classic "smoke words" identified by our procedure is shown in Appendix A4 in the Online Supplement. (Word roots are shown, because the smoke word includes plurals and other inflections of the word root form.) Appendix A4 also indicates whether each word appears in each of a number of conventional sentiment analysis dictionaries. Words shown in **bold** in column 1 of Appendix A4 are smoke words that are not found in any of the conventional sentiment analysis dictionaries listed. Reviewing the full list of 1000 smoke words that we uncovered in the iPod Classic discussion forum, we found that only a small minority of these—131 words (13%)—appear in any of the conventional sentiment detection dictionaries listed, re-affirming that smoke words are highly distinct from sentiment words.

6. Limitations

Our study is subject to the existence of possible biases in both defect information sharing and defect information utilization. We turn to these here.

In terms of *defect information sharing*, the behavioral research literature discusses a number of pertinent biases, such as self-selection bias (Heckman 1979), non-response bias (Armstrong and Overton 1977, Dellarocas and Wood 2008), negativity bias (Rozin and Royzman 2001), single source bias (Podsakoff and Organ 1986), and common method bias (Podsakoff et al. 2003). Specifically, in the online social media context, some of the possible biases include self-selection bias (users deciding whether or not to post), negativity bias and negativity contagion (worse experiences are more likely to be shared), selection and exclusion biases (the choice of forums, and postings, studied), and bias that is related to the popularity of the product (more popular products are more likely to be discussed).

Considering, for example, product popularity bias, we observed in our analysis that our vehicle dataset in Case Study 1 includes user discussions on 61 unique models. Six of these—Honda Accord and Civic; Chevy Silverado, Tahoe, and Suburban; and Toyota Camry—all top-sellers for their respective manufacturers in the United States, together account for 54% of the online discussions in the sample. Some normalization, to mitigate bias related to the product popularity, can be achieved by expanding the minimum number of threads gathered for each model—our case study sampled less than 1/3 of 1% of available threads—and by determining the percentage of defects for each model relative to the number of postings (rather than the raw absolute defects). One can then perform benchmarking analysis between these models, and assess relative proportion of defects by component category, for each model. Our framework is likely applicable and valuable only for products with relatively high sales volume, of thousands of units per day. Such products can generate sufficient online discussion data for analysis, and even a small reduction in defect discovery time would yield significant savings by substantially reducing defective units in circulation.

Selection and exclusion bias, and single source bias, in our case studies can be reduced in commercial application by collecting data from additional online media outlets to increase the coverage of the user

voices that are relevant to the product of the study, such as user discussions at Edmunds.com, Kelly Blue Book, Consumer Reports, Twitter, Facebook, Blogger, Amazon, and other online product discussion sources. Furthermore, defect discovery from online consumer discussions should be relied on as only one arrow in a quiver of defect discovery approaches: traditional defect discovery procedures such as warranty service records, consumer surveys, and regulatory agency filings are complementary to social-media-based discovery techniques, and the different approaches can further mitigate single source bias.

We do not believe the biases discussed above posed a significant issue to our study since our focus here is not to exhaustively list all product defects related to different product models. Rather we are trying to show the viability of using social media signal cues for product defect discovery. A positive element of online discussion forum data, which is not available in traditional survey data, is the ability to weight the dataset based on user or posting popularity. For example, manufacturers may be more concerned with defects reported by "celebrity" users (e.g., with many followers) or postings that have been highly read, frequently replied to, and/or regularly searched for. Although our study did not re-weight data based on user or posting influence, such techniques may be helpful in practice.

Timeliness of defect discovery is limited by a number of significant lags, including the lag between when the consumer acquires the product and when they observe the defect, and the lag between observing the defect and reporting the defect to the organization or to the user community. In addition, there is a search lag on the part of the company, which may be the most important systems effect. Daily, rather than quarterly or annual, defect crawling through social media may be essential to reducing search lag.

Regarding utilization of defect information within the organization, we expect that defect predictions from social media analytics would be gathered and prioritized by the corporate quality unit and then reported to the VP for manufacturing (for forwarding to product engineers and plant operators), to the VP for marketing (for forwarding to product designers), as well as to the CEO for any additional assessment and action, such as forwarding to the legal department.

In terms of *defect information utilization*, the statistical biases discussed above omit a number of important issues and biases in responding to defect information, which the behavioral operations management and systems dynamics understanding literature shed further light on. The behavioral operations management and shared systems dynamics understanding literature suggests that, regardless of the technical sophistication of the defect discovery system, soft human factors may be of critical importance to overall project success. Cross-functional integration is not cost-free. It can have a detrimental influence on complexity, ease of production, and new product development timing, and also may complicate inter-functional relationships, increase organizational conflict, stress, confusion, and poor decision making (Swink and Song 2007). For example, features that have little product performance impact can have significant manufacturability impact. In relation to QM, the behavioral operations management literature (Bendoly et al. 2006), highlights gaps between model assumptions and actual human behavior. For defect discovery, some model assumptions are that it is sufficient to sample discussions randomly, computationally rank possible improvement efforts, and isolate defect remediation solutions. The behavioral OM literature highlights though that in actuality, humans are poor randomizers, improvement effort choice is affected by individual prejudices, and new procedures disrupt system dynamics (Ibid.). Furthermore, objective human verification of algorithmically predicted defects is hampered by the effect of human stress and fatigue on error rates. Behavioral moderation can also threaten quality improvement performance in other ways: for example, individuals have different tolerance for risk and ambiguity, meaning that the defect verification personnel who are tasked with reviewing the computationally predicted defects will naturally have varying predispositions to False Positive and False Negative defect prediction errors, which can have significant implications for the organization. The shared systems dynamics literature (Bendoly 2013) recognizes that a lack of appreciation of automotive organization dynamics by consumers reporting issues can mean that the most critical information is not shared at the most appropriate points in time with the most pertinent defect remediation team members. Consumers lack broad automotive organization expertise and may be less likely to understand the specific organizational relevance and value of new information. Consumers may not appreciate the breadth of information necessary to make full diagnosis viable (e.g. complete sensor and actuator logs from the vehicle on-board diagnostics computer may be needed by a variety of engineering specialists), and consumers may not appreciate the costs of organizational action. Hence, they may not be in a good position to supply useful quality information to the organizational responders who may benefit from the knowledge. Furthermore, consumer or employee perceptions of psychological safety (Edmondson 1999)—such as perceived risks of appearing foolish or different, or being ostracized—may drive the

frequency and volume of information sharing, and thus likely biases information sharing via discussion forum postings, exacerbating the non-response bias. Even well-informed individuals may perceive the risk of isolation if their views differ substantially from the majority in the group (Williams and O'reilly 1998), and action may also be inhibited by individual perceptions of difficulty (Bendoly et al. 2010), implying that, even internally, defect experts may sometimes be averse to sharing or acting on certain defect information. Assigning reward points to encourage expertise contribution is one mitigation strategy that has been employed in online forums (Wang et al. 2013) to reduce non-response bias, and "whistle-blower" rewards have been employed internally within organizations to encourage action, though such incentives may introduce their own biases.

In summary, defect information sharing and information utilization is affected by biases and issues that extend beyond the pure technical validity of the computerized defect discovery system. Recognition and mitigation of such biases is of critical importance to the overall success of the holistic organizational defect management project and process.

7. Discussions and Conclusions

Our analyses show that model accuracy improves as context-specific features (semantic factors, distinctive terms, and product feature information) are added to context-independent (stylistic, social, and sentiment) features. This is consistent with prior studies which show that general sentiment features have negligible value in predicting the existence of defects (Abrahams et al. 2012).

It should be noted that our findings on the strength and direction of associations between the input features and the probability of the thread discussing a perceived defect are applicable to the sample discussion forums we studied. These associations may or may not generalize to other industries, such as general retail, service, and hospitality. Future work is needed to study how the framework proposed in this study will play out in other industries, and to determine which associations generalize across industries, locations, and times. For those associations that do not generalize, it is important to document what alternative associations are evident and what models are appropriate.

7.1. Implications for Research

Our integrated framework in Figure 1 documents a fairly comprehensive list of features (signal cues) that can be automatically extracted from the textual data embedded in the vast amount of social media. These features can then be used as inputs for product defect discovery. Very few prior research studies have touched upon this topic. Future research studies engaged in text analytics for QM applications can greatly benefit from the framework.

When applying the framework in Figure 1, one should be advised that not all features mentioned in our framework have to be used. In our two case studies, we did not include lexical features in our study because we found that adding such features would greatly increase the difficulty of interpreting the results of our research findings, since the size of lexical features are roughly proportional to the number of unique words in our collection. Given that we are interested in understanding the relative importance of various text signals extracted from UGC, we did not feel it necessary to include lexical features. Instead, we used the *distinctive terms* feature ("smoke words") as a substitute for lexical features to reduce the dimensionality of the feature space.

While our study indicates that smoke words—which are identified for each domain using the methodology we describe—are more significant than sentiment words for the two industries covered in our case studies, it is possible that this finding may not generalize. Future research is necessary to identify product classifications and categories where users are neutral vs. emotive in their discussion of defects. Furthermore, given that our framework already measures emotive words in each thread, future behavioral research could be helpful to determine whether psychological safety bias is exacerbated in threads containing explosive rants (measured by negative emotive words), and to determine the effects that social media with lower levels of anonymity (e.g., person-centric sites such as Facebook, vs. product-centric discussion forums) have on psychological safety bias.

It needs to be noted that nomenclature differences do exist with regard to the naming of the signal cues in Figure 1. Different research studies in different disciplines often use different terminology to refer to similar concepts. For example, people in the Finance or Accounting domain often prefer to use *word counts* or *frequency counts* to capture the word distribution information from text (Li 2010, Loughran and Mcdonald 2011, Tetlock et al. 2008), which is referred to as *lexical cues* in our framework. While such nomenclature differences exist across disciplines, our proposed framework undoubtedly provides a coherent framework as the first step to streamline the terminology for various text analytics-based modeling studies.

In addition, some signal cues can be used as inputs to create other types of derived signal cues. This typically happens between context-independent features and context-specific features. For example, lexical approaches are often combined with semantic dictionary approaches, where a triggered term, rule, or frame maps to a semantic category (Riloff and

Lehnert 1994). The occurrence and frequency of each semantic category can then be computed for each document (Abrahams et al. 2012, Loughran and Mcdonald 2011). Extensions of this approach allow the synthesis of multiple distinct terms into single product attributes or concept groups—so-called product-feature/product-attribute extraction (Ferreira et al. 2008) or facet extraction (Vechtomova 2010). For example, "cheap" and "expensive" both imply that the "price" attribute of the product is being assessed (Decker and Trusov 2010). Similarly, lexical features can often be used as the first-order inputs to create the second-order distinctive terms feature for advanced text classification applications. For example, feature extraction techniques used in text mining can be used to perform the selection of more useful "distinctive terms" based on the distribution of lexical features (Abrahams et al. 2012, Fan et al. 2006).

The operationalization of the listed features in our framework can be quite flexible. For instance, the social features mentioned in the framework can use any derived features from user social networks to gauge useful properties about a user's social network. One can easily perform social network analysis on users' social networks and compute various features related to the users' social network: centrality, in-degree, out-degree, eigenvalue, clustering coefficient, etc. (Wang et al. 2011). Similarly, the product feature tags can use operationalizations based on the study context. In this study, we used product component category as the product feature descriptor. For other consumer products and services, one can use the corresponding product/service components and their attribute values as product feature descriptors to help with the discovery tasks. As an example, for camera-related products, one can use lens, shutter speed, image resolution, picture quality, etc. as product feature tags.

The list of features in the framework is by no means exhaustive. Future researchers can always enrich and extend the set of features to suit their defect discovery needs. In addition, one can also complement the text-based features with other context relevant quantitative features or variables for advanced modeling.

Finally, further research is needed into how companies would incorporate automated defect discovery from social media into their QM practices. Unresolved questions include, for instance, how often it should be done, what unit of the company should do it, how they should report it, how to tie it back into their FMEA, and how it can be tied into their quality function deployment or other product development tools.

7.2. Implications for Practice

Our proposed framework has several implications for business practice. As firms continue to tap into social media for defect discovery, our study offers them a systematic framework to engage in such efforts. The framework gives firms' concrete ideas on how to relate various textual signals, extracted from social media, with quality issues. Since our framework requires people with expertise in text analytics and data mining, firms should hire people that possess the technical skill sets in performing large scale data gathering, indexing, text preprocessing, feature extraction, and advanced modeling.

The defect discovery tool described here can be used in practice to improve quality, and reduce defect-associated costs. We have shown that automated defect discovery can be achieved with reasonable accuracy. This can reduce the number of postings that need to be manually reviewed by human experts to identify defects, and consequently lower the average time to defect discovery. The consequent reduction in defective product units in the hands of consumers would substantially lower defect-associated costs to the manufacturer. In addition, the data can be used to get a normalized evaluation of defects to compare product models from one manufacturer to other models in the same or different product class from the same or other manufacturers, to compare a new product generation to earlier generations of the same product model, or to identify problematic components that are in use within multiple product models or generations. This information can be used to allocate engineering resources to improvement efforts, and can be used to identify areas for preventive maintenance.

As demonstrated in both case studies, automated defect detection is imperfect, with classification accuracy currently slightly higher than 70%. This implies that firms must continue to employ manual review by experts, to verify suspected defects. The automated tool provides benefit in reducing the number of postings that must be reviewed per defect verified, and consequently reduces defect discovery time, cost, and reviewer fatigue.

Depending upon the goals of their social media mining efforts, firms should also be selective when choosing various modeling techniques. If the goal of the data mining task is to understand and explain, then they should use techniques that offer more interpretable results like our usage of logistic regression. On the other hand, if the goal of the project is for prediction, then they can use other advanced machine learning techniques such as support vector machines or neural networks.

In our review of limitations (section 6), we discussed a number of issues and biases in both defect information sharing and defect information utilization that can potentially affect performance of the holistic defect management system, which incorpo-

rates both the "hard" technical and "soft" human elements. In practical application, various mitigation strategies, such as those suggested in section 6, ought to be pursued to address such biases and issues.

When applying our framework, firms should be creative since the framework is comprehensive, but not exhaustive, and is extensible. While we have shown certain variables to be of most significance in defect discovery for our sample selection of brands in a certain snap-shot of time, firms should be careful to build dynamic models for their particular product set as specific results will vary by time, geography, and product. The framework and defect discovery exemplar cases demonstrated here can be used as a springboard to guide and expedite such efforts.

Acknowledgments

The authors are grateful to Dr. Mehdi Ahmadian, Director of the Center for Vehicle Systems and Safety, Virginia Tech, for his helpful feedback and assistance with the vehicle defect case study.

References

Abbasi, A., H. Chen. 2008. Cybergate: A design framework and system for text analysis of computer-mediated communication. *MIS Q.* **32**(4): 811–837.

Abbasi, A., H. Chen, A. Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.* **26**(3): 12.

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

Abrahams, A. S., E. Coupey, E. X. Zhong, R. Barkhi, P. S. Manasantivongs. 2013a. Audience targeting by b-to-b advertisement classification: A neural network approach. *Expert Syst. Appl.* **40**(8): 2777–2791.

Abrahams, A. S., J. Jiao, W. Fan, G. A. Wang, Z. Zhang. 2013b. What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decis. Support Syst.* **55**(4): 871–882.

Agresti, A. 2013. *Categorical Data Analysis*, 3rd ed, Vol. 359. Wiley-Interscience, Hoboken, NJ.

Ahire, S. L., R. Landeros, D. Y. Golhar. 1995. Total quality management: A literature review and an agenda for future research. *Prod. Oper. Manag.* **4**(3): 277–306.

Anderson, E. W., M. W. Sullivan. 1993. The antecedents and consequences of customer satisfaction for firms. *Mark. Sci.* **12**(2): 125–143.

Antweiler, W., M. Z. Frank. 2005. Is all that talk just noise? The information content of internet stock message boards. *J. Finance* **59**(3): 1259–1294.

Apté, C., F. Damerau, S. M. Weiss. 1994. Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst.* **12**(3): 233–251.

Armstrong, J. S., T. S. Overton. 1977. Estimating nonresponse bias in mail surveys. *J. Mark. Res.* **14**(3): 396–402.

Bendoly, E. 2013. System dynamics understanding in projects: Information sharing, psychological safety, and performance effects. *Prod. Oper. Manag.* **23**(8): 1352–1369.

Bendoly, E., K. Donohue, K. L. Schultz. 2006. Behavior in operations management: Assessing recent findings and revisiting old assumptions. *J. Oper. Manag.* **24**(6): 737–752.

Bendoly, E., J. E. Perry-Smith, D. G. Bachrach. 2010. The perception of difficulty in project-work planning and its impact on resource sharing. *J. Oper. Manag.* **28**(5): 385–397.

Binder, A. K. 2013. *Wards Automotive Yearbook 2013*, 75th edn. Penton Media Inc, Southfield, MI.

Borko, H., M. Bernick. 1963. Automatic document classification. *J. ACM* **10**(2): 151–162.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**(7): 1145–1159.

Bryman, A., R. G. Burgess. 2002. *Analyzing Qualitative Data*. Routledge, London.

Calvo, R. A., J.-M. Lee, X. Li. 2004. Managing content with automatic document classification. *J. Digital Inf.* **5**(2).

Cao, Q., W. Duan, Q. Gan. 2011. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decis. Support Syst.* **50**(2): 511–521.

Chang, C.-C., C.-J. Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**(3): 27.

Chang, C.-H., C.-N. Hsu, S.-C. Lui. 2003. Automatic information extraction from semi-structured web pages by pattern discovery. *Decis. Support Syst.* **35**(1): 129–147.

Chang, C., M. Kayed, M. R. Girgis, K. F. Shaalan. 2006. A survey of web information extraction systems. *IEEE Trans. Knowl. Data Eng.* **18**(10): 1411–1428.

Chevalier, J. A., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Mark. Res.* **43**(8): 345–354.

Coussement, K., D. Van Den Poel. 2008. Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decis. Support Syst.* **44**(4): 870–882.

Cowie, J., Y. Wilks. 2000. Information extraction. R. Dale, H. Moisl, H. L. Somers, eds. *Handbook of Natural Language Processing*. Marcel Dekker, New York, NY, 241–260.

Das, S. R., M. Y. Chen. 2007. Yahoo! For Amazon: Sentiment extraction from small talk on the web. *Manage. Sci.* **53**(9): 1375–1388.

Decker, R., M. Trusov. 2010. Estimating aggregate consumer preferences from online product reviews. *Int. J. Res. Mark.* **27**(4): 293–307.

Dellarocas, C., C. A. Wood. 2008. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Manage. Sci.* **54**(3): 460–476.

Duan, W., B. Gu, A. B. Whinston. 2008. Do online reviews matter?—an empirical investigation of panel data. *Decis. Support Syst.* **45**(4): 1007–1016.

Duriau, V. J., R. K. Reger, M. D. Pfarrer. 2007. A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organ. Res. Methods* **10**(1): 5–34.

Edmondson, A. 1999. Psychological safety and learning behavior in work teams. *Adm. Sci. Q.* **44**(2): 350–383.

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

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

Fan, W., L. Wallace, S. Rich, Z. Zhang. 2006. Tapping the power of text mining. *Commun. ACM* **49**(9): 76–82.

Ferreira, L., N. Jakob, I. Gurevych. 2008. A comparative study of feature extraction algorithms in customer reviews. Proceed-

ings of 2008 IEEE International Conference on Semantic Computing, Santa Clara, CA, pp. 144–151.

Flesch, R. F. 1951. *How to Test Readability*. Harper, New York, NY.

Friedl, J. E. F. 2006. *Mastering Regular Expressions*. O'Reilly, Sebastopol, CA.

Gopal, R., J. R. Marsden, J. Vanthienen. 2011. Information mining—reflections on recent advancements and the road ahead in data, text, and media mining. *Decis. Support Syst.* **51**(4): 727–731.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* **11**(1): 10–18.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica* **47**(1): 153–161.

Isidore, C. 2014. GM's $1.3 billion recall cost wipes out profit. *CNN Money*. Available at http://money.cnn.com/2014/04/24/news/companies/gm-earnings-recall/ (accessed date September 8, 2014).

Jesty, P. H., K. M. Hobley, R. Evans, I. Kendall. 2000. Safety analysis of vehicle-based systems. F. Redmill, T. Anderson, eds. *Lessons in System Safety: Proceedings of the 8th Safety-Critical Systems Symposium*. Springer, New York, NY, 90–110.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, pp. 200–209.

Joachims, T., F. Sebastiani. 2002. Guest editors' introduction to the special issue on automated text categorization. *J. Intell. Inf. Syst.* **18**(2): 103–105.

Johannsen, C. G. 2000. Total quality management in a knowledge management perspective. *J. Documentation* **56**(1): 42–54.

Jolliffe, I. 2002. *Principle Component Analysis*, 2nd edn. Springer, New York.

Kelly, E. F., P. J. Stone. 1975. *Computer Recognition of English Word Senses*. North-Holland, Amsterdam.

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

Lebeau, P. 2014. Global auto sales hit record high of 82.8 million. *CNBC*. Available at http://www.cnbc.com/id/101321938 (accessed date 8 September 2014).

Lewis, D. D., Y. Yang, T. G. Rose, F. Li. 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* **5**: 361–397.

Li, F. 2010. Textual analysis of corporate disclosures: A survey of the literature. *J. Account. Lit.* **29**: 143–165.

Li, Y. H., A. K. Jain. 1998. Classification of text documents. *Compu. J.* **41**(8): 537–546.

Li, N., D. D. Wu. 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decis. Support Syst.* **48**(2): 354–368.

Lim, E.-P., H. Chen, G. Chen. 2013. Business intelligence and analytics: Research directions. *ACM Trans. Manag. Inf. Syst.* **3**(4): 1–10.

Lohr, S. 2012. The age of big data. *New York Times*, February 12, 2011.

Loughran, T., B. Mcdonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *J. Finance* **66**(1): 35–65.

Ma, Z., G. Pant, O. R. Sheng. 2011. Mining competitor relationships from online news: A network-based approach. *Electron. Commer. Res. Appl.* **10**(4): 418–427.

Muller, J. 2014. Toyota admits misleading customers; agrees to $1.2 billion criminal fine, *Forbes*. Available at http://www.forbes.com/sites/joannmuller/2014/03/19/toyota-admits-misleading-customers-agrees-to-1-2-billion-criminal-fine/ (accessed date September 8, 2014).

Neuendorf, K. A. 2001. *The Content Analysis Guidebook*. Sage Publications, Incorporated, Thousand Oaks, CA.

Nielsen, F. Å. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the 1st Workshop on Making Sense of Microposts, Heraklion, Crete, pp. 93–98.

Oh, C., O. R. L. Sheng. 2011. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. Proceedings of the 32nd International Conference on Information Systems, Shanghai, China, Paper 17, pp. 1–19.

iPod + iTunes timeline. 2014. Available at http://www.apple.com/pr/products/ipodhistory/ (accessed date June 10, 2014).

Podsakoff, P. M., D. W. Organ. 1986. Self-reports in organizational research: Problems and prospects. *J. Manag.* **12**(4): 531–544.

Podsakoff, P. M., S. B. Mackenzie, J.-Y. Lee, N. P. Podsakoff. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *J. Appl. Psychol.* **88**(5): 879.

Riloff, E., W. Lehnert. 1994. Information extraction as a basis for high-precision text classification. *ACM Trans. Inf. Syst.* **12**(3): 296–333.

Romano, N. C., C. Donovan, H. Chen, J. F. Nunamaker. 2003. A methodology for analyzing web-based qualitative data. *J. Manag. Inf. Syst.* **19**(4): 213–246.

Rozin, P., E. B. Royzman. 2001. Negativity bias, negativity dominance, and contagion. *Pers. Soc. Psychol. Rev.* **5**(4): 296–320.

Ruiz, M. E., P. Srinivasan. 2002. Hierarchical text categorization using neural networks. *Inf. Retrieval* **5**(1): 87–118.

Santos, R. L. 2012. Information retrieval on the blogosphere. *Found. Trends Inf. Retrieval* **6**(1): 1–125.

Schroeder, R. G., K. Linderman, D. Zhang. 2005. Evolution of quality: First fifty issues of production and operations management. *Prod. Oper. Manag.* **14**(4): 468–481.

Schumaker, R. P., H. Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.* **27**(2): 12.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.* **34**(1): 1–47.

Shmueli, G., N. R. Patel, P. C. Bruce. 2010. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, 2nd edn. John Wiley & Sons Inc, Hoboken, NJ.

Spangler, S., J. Kreulen. 2007. *Mining the Talk: Unlocking the Business Value in Unstructured Information*. IBM Press, Boston, MA.

Stone, P. J., D. C. Dunphy, M. S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge, MA.

Swink, M., M. Song. 2007. Effects of marketing-manufacturing integration on new product development time and competitive advantage. *J. Oper. Manag.* **25**(1): 203–217.

Tetlock, P. C., M. Saar-Tsechansky, S. Macskassy. 2008. More than words: Quantifying language to measure firms' fundamentals. *J. Finance* **63**(3): 1437–1467.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, A. Kappas. 2010. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **61**(12): 2544–2558.

Thelwall, M., K. Buckley, G. Paltoglou. 2012. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.* **63**(1): 163–173.

Tong, S., D. Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2**: 45–66.

Vechtomova, O. 2010. Facet-based opinion retrieval from blogs. *Inf. Process. Manag.* **46**(1): 71–88.

Wang, G.A., X. Liu, W. Fan. 2011. A knowledge adoption model based framework for finding helpful user-generated contents in online communities. Proceedings of 30th Second Interna-

tional Conference on Information Systems, Shanghai, China, Paper 15.

Wang, G. A., J. Jiao, A. S. Abrahams, W. Fan, Z. Zhang. 2013. Expertrank: A topic-aware expert finding algorithm for online knowledge communities. *Decis. Support Syst.* **54**(3): 1442–1451.

Williams, K. Y., C. A. O'reilly. 1998. Demography and diversity in organizations: A review of 40 years of research. *Res. Organ. Behav.* **20**: 77–140.

Woolridge, A. 2011. Too much buzz: Social media provides huge opportunities, but will bring huge problems. *Economist* **401**(8765): 50.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Inf. Retrieval* **1**(1): 69–90.

Yang, Y., J. O. Pedersen. 1997. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, pp. 412–420.

Yang, Y., S. Slattery, R. Ghani. 2002. A study of approaches to hypertext categorization. *J. Intell. Inf. Syst.* **18**(2): 219–241.

Zhang, Y., Y. Dang, H. Chen, M. Thurmond, C. Larson. 2009. Automatic online news monitoring and classification for syndromic surveillance. *Decis. Support Syst.* **47**(4): 508–517.

Zhu, F., X. Zhang. 2010. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Mark.* **74**(2): 133–148.

Ziefle, M. 1998. Effects of display resolution on visual performance. *Hum. Factors* **40**(4): 554–568.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix A1: Summary of Text Analytics of Social Media Content.

Appendix A2: Variables from SMART framework, Definitions, & Descriptive Statistics for Threads (Automotive Case Study).

Appendix A3: Performance Comparison of Our Framework with Other Baseline Text Classification Methods (Automotive Case Study).

Appendix A4: Most Significant "smoke words" for Case Study 2 (iPod Classic), with comparison to conventional sentiment dictionaries.