Decision Support Systems 54 (2012) 87–97

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Vehicle defect discovery from social media

Alan S. Abrahams [a,*], Jian Jiao [b], G. Alan Wang [a], Weiguo Fan [c]

[a] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, 1007 Pamplin Hall, Blacksburg, VA 24061, United States
[b] Department of Computer Science, Virginia Tech, 114 McBryde Hall, Blacksburg, VA 24061, United States
[c] Department of Accounting and Information Systems, Pamplin College of Business, Virginia Tech, 3007 Pamplin Hall, Blacksburg, VA 24061, United States

ARTICLE INFO

Article history:
Received 4 November 2011
Received in revised form 9 March 2012
Accepted 29 April 2012
Available online 8 May 2012

Keywords:
Quality management
Social media analytics
Business intelligence
Text mining

ABSTRACT

A pressing need of vehicle quality management professionals is decision support for the vehicle defect discovery and classification process. In this paper, we employ text mining on a popular social medium used by vehicle enthusiasts: online discussion forums. We find that sentiment analysis, a conventional technique for consumer complaint detection, is insufficient for finding, categorizing, and prioritizing vehicle defects discussed in online forums, and we describe and evaluate a new process and decision support system for automotive defect identification and prioritization. Our findings provide managerial insights into how social media analytics can improve automotive quality management.

1. Introduction

Vehicle safety and performance defects are of major concern to automotive manufacturers. In the last decade, safety defects reported by consumers to the federal agency responsible have resulted in enormous recalls in industries such as food, toys, medical devices, and automobiles [30]. In the motor vehicle industry, which is the concern of this paper, the National Highway Traffic Safety Administration (NHTSA) has issued over 90,000 recalls, resulting in billions of dollars of expenses to vehicle manufacturers, dealers, and consumers. Notably, recalls capture only one category of defect: safety defects. Performance defects, which are not reported to federal agencies, represent a huge additional concern for both quality management professionals and consumers.

Typically, vehicle manufacturers discover issues through their own vehicle tests, inspection procedures, or information gathering. Manufacturers may, for instance, review warranty claims or dealership service records, or consult consolidated insurance industry data, such as reports provided by the Insurance Institute for Highway Safety (IIHS). Manufacturers also regularly review consumer complaints filed with regulatory agencies, like the NHTSA. We believe, however, there are a lot of useful and hidden vehicle quality data embedded in social media that are largely untapped into by manufacturers. Recently, automotive companies like Chrysler have begun to employ "Twitter teams" to reply to whining tweets; but, detecting "whispers of useful information in a howling hurricane of noise" is a huge challenge and better filters are needed to extract meaning from the "blizzard of buzz" [64]. In this paper, we explore the use of decision support technologies geared towards social media data (hereafter called social media analytics) to mine the alternative vehicle defect data source: social media — specifically, online discussion forums. We create and evaluate a novel Vehicle Defect Discovery System (VDDS), which provides decision support to vehicle quality management professionals and consumers.

Consumers rely heavily on the Internet for information about automobile safety and reliability [35], consulting resources such as vehicle consumer surveys, insurance industry statistics, manufacturer websites, and complaints filed via regulatory agencies. However, Internet users go beyond *consumption* of vehicle safety and reliability information, and engage in *production* of such information using a variety of online social media tools (bulletin boards/forums, blogs, twitter, etc.), on a vast scale. Consumer feedback provides a valuable information source for improving product designs and marketing strategies [23]. Finding and analyzing consumer-produced knowledge of vehicle defects, buried among millions of consumer postings, is a difficult quality management challenge that has received very little attention in prior literature. Understanding and prioritizing the vast volume of automotive information produced by consumers, and sifting out the safety and performance issues from the discussion 'dregs', is the focus of this paper.

Our decision support system for vehicle defect discovery from social media – VDDS – focuses on a single social medium manifestation: discussion forums for current and prospective owners of a few major vehicle brands. We chose to study Honda, Toyota, and Chevrolet discussion forums due to the high usage of these forums by enthusiasts of these brands in the past decade. Automotive experts were employed to analyze and categorize thousands of postings from these forums. It is

* Corresponding author. Tel.: +1 540 231 5887; fax: +1 540 231 3752.
E-mail addresses: abra@vt.edu (A.S. Abrahams), jjiao@vt.edu (J. Jiao), alanwang@vt.edu (G.A. Wang), wfan@vt.edu (W. Fan).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2012.04.005

well known that consumer-produced content can rapidly overwhelm the firm's information processing capabilities [44]. We therefore designed a text analysis framework to distill the safety and reliability information into a digestible format for automakers, and implemented this framework in a novel VDDS. We evaluated the effectiveness of the VDDS across multiple vehicle brands.

In our analysis, we found that automotive consumers do indeed use social media for the production of information relevant to vehicle quality management. But conventional mechanisms, like sentiment analysis, for analyzing this social media content fall short. We therefore describe and evaluate an alternative defect discovery approach, based on a new class of linguistic marker words we found. We refer to this new class of words as 'automotive smoke words'.

The rest of this paper is structured as follows. First, we motivate the need for vehicle quality management research targeted specifically at defect discovery from textual online discussion forums. Next, we discuss and contrast related work. We describe our contributions and the research questions we aim to address. We lay out a workflow for vehicle quality management using social media analysis, and implement the workflow in a VDDS. We discuss the application of our VDDS to vehicle defect discovery and classification using a large sample data set. Finally, we draw some conclusions and propose future work.

2. Background and related work

In this section, we describe both the practical motivation behind our work, as well as the research motivation. We explore related work on social media, business/competitive intelligence, text mining, and sentiment analysis. We review the coverage and limitations of prior work, and the research questions raised.

2.1. Social media

Social computing platforms have recently received substantial attention [44]. We define social media as online services that provide for decentralized, user level content creation (including editing or tagging), social interaction, and open (public) membership. In our definition, public discussion forums, public listervs, public wikis, open online communities (social networks), public usenet groups, customer product reviews, public visitor comments, user-contributed news articles, and folksonomies would fall within the gamut of social media. However, internal corporate emails, private wikis, private discussion forums, and corporate news articles are excluded, as they are non-public and/or under centralized control. Social media are characterized by vast volumes of user-contributed content of variable quality. Navigation of this content is a significant research challenge [29,44] that may require filtering, semantic content, tagging, information mining, or other techniques.

2.2. Business intelligence

Competitive intelligence (a.k.a. business intelligence) technologies are valuable from multiple perspectives: understanding Customers, Competitors, Markets, Products, Environment, Technologies, Acquisitions, Alliances, and Suppliers [58]. These perspectives form the basic units of analysis that competitive intelligence tools typically focus on.

To gain a better understanding of the issues affecting their products, effective firms must gather product-relevant information both internally and externally [23,27]. It is widely accepted that consumer complaints are a valuable source of product intelligence [26,39,48,49]. The knowledge of outsiders or user communities is an important source of product-related business intelligence [4,16–18]. Historically, many companies have invested substantial effort in soliciting product usage stories from practitioners, for the purposes of diagnosing or understanding problems, or allocating issues to technicians able to solve them. Especially for firms selling a mechanical consumer product, these so-called 'communities of practice' are an important aspect of a firm's business intelligence repertoire, as they provide a repository of past usage experiences which can be drawn upon for operational issue resolution, product development, or other purposes [8,13,36,51,60,61].

The idea of using online information for competitive intelligence predates popular use of the Web [20]. The Web has become an important source of business intelligence information [9], and the use of web mining for business intelligence has been a popular area of research [33]. As unstructured, dynamic social computing content proliferates, an opportunity arises for firms to specialize as navigators of social computing content, using techniques like data mining, tagging, filtering and AI [44].

One question that can be raised is whether there is really useful business intelligence available in online social media content? Prior research has demonstrated that the nature or volume of online reviews can, for instance, predict stock market volatility [5] or movie box office sales [18]. It has been found that online news postings have sufficient linguistic content to be predictive of a firm's earnings and stock returns [56]. For mechanical consumer products, the availability of significant product quality information within online forums has been demonstrated through exploratory studies involving the manual tagging of listserv emails or newsgroup postings about fly fishing gear [24] and power tool equipment [23]. However, manual tagging is arduous and too time-consuming to keep pace with the mountains of consumer-generated discussion content being generated daily, for hundreds or thousands of brands and products. In our study, we advance the business intelligence field by demonstrating the application of computational techniques for automated vehicle defect discovery and analysis from online automobile enthusiast discussions.

We turn next to past work in the text mining field aimed at extracting business intelligence from customer interactions, and we examine its applicability to the discovery and analysis of defects described by customers on online discussion forums.

2.3. Text mining

Text mining researchers have devoted substantial attention to the mining of emails, news articles, discussion forums, and customer reviews for useful knowledge. Text mining application areas include information extraction, topic tracking, summarization, categorization, concept linkage, clustering, information visualization, question answering, content filtering, and prediction [22,29]. Text mining has become a useful tool to analyze user generated texts, extract business intelligence, and support decision making.

Existing studies have applied text mining to various perspectives of competitive intelligence. For the purposes of customer complaint management, Coussement and van den Poel [14] classify large volumes of inbound emails to distinguish complaints from non-complaints. Spangler and Kreulen [54] propose a systematic approach for analyzing unstructured data generated by customers, in order to identify their concerns. In online sports-fan discussion forums, researchers have built methods for detecting and forecasting which sub-forums on a discussion forum currently contain, or are likely to contain, heated discussion [37]. From customer reviews, researchers have been able to project box-office popularity for film releases [18,47].

Abbasi and Chen [1] proposed a robust framework for textual content analysis of computer-mediated communication. They applied the framework to the understanding of pre-scandal and post-scandal communication patterns of Enron employees. From a market perspective, Schumaker and Chen [52] used different textual representations to analyze financial news articles and predict future stock price changes using a machine learning approach. Text mining has also occasionally been applied to the environment perspective of competitive intelligence. For example, Zhang et al. [66] monitor news articles in order to track the emergence of health epidemics.

Various authors have suggested that high quality domain-specific information may be difficult to locate. For instance, in the healthcare profession, reliable and up-to-date health-related data is highly distributed, of varying quality, and difficult to locate on the Web [10]. In the healthcare domain, special-purpose cancer "vertical search spiders", which use classification technology from the text mining literature, have been developed [11] to specifically identify high quality documents relating to cancer topics from among the thousands of low quality results returned by traditional search engines. In the automotive domain, sifting defect postings (especially safety defect postings) from the discussion 'dregs' is similarly challenging and important.

Few text mining studies particularly focus on the product quality perspective of competitive intelligence. Researchers have demonstrated the availability of product quality information in online forums [23,24]. However, their approach involved manual tagging that is often time-consuming and inefficient. Defect detection has nuances that distinguish it from typical text classification tasks. Unlike popular topics, defects may be rare (infrequently discussed). The criticality of the defect must be determined, and not just its popularity, since a defect likely to lead to injury or death is of much greater concern than defects that merely constitute a nuisance.

2.4. Sentiment analysis

Sentiment analysis is a popular text mining procedure, which allows the end user to rapidly discover postings containing highly emotive content [2,6,43,57]. For instance, in Harvard General Inquirer [34,55], each individual word in a piece of text is disambiguated and a lexicon of positive and negative word senses is consulted to determine the sentiment (if any) of each word sense.

Abbasi, Chen, and Salem [2] investigated the linguistic features of Web forum messages to determine the opinion polarity of each message. Their method can potentially be used to identify inferior features of products that cause customers' concerns. Other studies have determined the opinion polarity, either positive (praise) or negative (criticism), of a textual message in order to predict financial market volatility [5] or firms' earnings and stock returns [38,56].

In the OpinionFinder system [46,62,63], sentiment analysis is at the phrase-level, rather than at the word-sense level or message-level. Multiple subjective expressions (phrases) in each message are identified and tagged with a contextual polarity. Lexical entries (e.g. "ugly") are assigned an initial prior polarity (e.g. negative polarity). However, the final polarity (contextual polarity) of the phrase is also influenced by other features, such as negation, syntax, and modality. For instance, while the word "ugly" has a negative prior polarity, the phrases "She is *not* ugly" and "She is *never* ugly" have positive contextual polarity (because the words "not" and "never" influence "ugly", making the phrases as a whole positive, even though the phrases contain the negative word "ugly").

In the sentiment analysis literature, it is presumed that heavily negative postings (complaints) will be indicative of product defects. However, whether this presumption – that negative sentiment predicts defect existence – is true for automotive defects has not been tested in prior research. There is some evidence that generic sentiment analysis fails when applied across domains. Loughran and McDonald [38] found that sentiment-indicative words differ across domains: specifically, in the field of finance, sentiment indicators were different from sentiment marker words previously thought to be generally applicable to all fields. O'Leary [42] found that generic positive and negative dictionaries had some limitations in describing negative behavior in the stock market, and suggested that domain specific terms be accounted for to improve the quality of the analysis.

In the vehicle domain, therefore, generic sentiment polarity analysis may be insufficient. A thread poster may be more aggrieved by a malfunctioning air conditioner than with a sticky accelerator pedal, yet the latter is almost certainly a more serious defect. Further, sentiment alone is not enough, in the motor vehicle industry. For instance, to enable proper investigation, the defect must be associated with the troublesome *component*, so hazard analysis can be performed [32,59]. Defects must be prioritized so that those which threaten safety can be sieved from those that are merely a nuisance.

2.5. Summary

Table 1 summarizes previous research on the organizational use of text analysis of traditional Internet media and social media, for competitive intelligence, in various application domains. For each study, Table 1 shows the medium studied, domain studied, competitive intelligence perspective, and method of analysis, for the study. We classify competitive intelligence perspectives using Vedder et al. [58], with the addition of the Employee perspective that was omitted from Vedder et al.'s original scheme.

Table 1 highlights the research gap which we aim to address in this paper: the application of automated text mining to vehicle defect analysis in online customer discussion forums.

3. Research questions and contributions

In this paper, we tackle two major research questions. Firstly, do auto enthusiast discussion forums contain substantial content related to motor vehicle defect existence and criticality? Secondly, analyzing the content of the discussion threads, can conventional sentiment analysis be used to distinguish defects from non-defects and safety from performance defects? If not, are there other characteristics that differentiate defects from non-defects, and safety issues from other postings?

We make three major contributions in this paper. This is the first large scale case study, to our knowledge, that confirms the usefulness of social media for vehicle quality management. Secondly, we demonstrate that conventional sentiment analysis – though successfully applied previously to complaint detection in retail, finance, film, and other industries – must be adapted for defect detection and prioritization for the automotive industry. Thirdly, we define a new class of vehicle 'smoke' words that are valuable to the auto industry for this task, and we describe a new VDDS that provides robust vehicle defect discovery from social media postings, across multiple automotive brands.

4. A text mining framework for vehicle quality management

Fig. 1 shows a process model for vehicle quality management from social media. Vehicle quality management includes defect identification and prioritization (items 1.–7.), and remediation (items 8.–10.). The Vehicle Defect Discovery System (VDDS) we described provides decision support for the defect identification and prioritization process (items 1.–7.). Vehicle defect remediation (items 8.–10.) is outside of the scope of this work.

The quality management process in Fig. 1 begins with the crawling (1.) of discussion forums to gather customer chatter from social media sources. In this paper, we confine the analysis to discussion boards, though other social media sources like Twitter and Facebook are equally pertinent. Next, data extraction (2.) is performed: the username, date and time of posting, and discussion text are extracted. Third, linguistic analysis (3.) is performed. Off-the-shelf linguistic analysis tools perform word-sense disambiguation, word categorization, and stemming. For example, feeding the sentence "my tires are bubbling" to the Harvard General Inquirer [34,55] we obtain:

```
MY (self)
TIRE#1 (object/tool)
ARE#2 (state verb)
BUBBLE#2 (natural process/interpretive verb).
```

Table 1
Comparison of text analysis studies using traditional web and social media.

Study	Medium	Domain	Competitive intelligence perspective	Method of analysis
Coussement and van den Poel, [14]	Email	Customer complaints	Customer	Automated
Spangler and Kreulen [54]	Email	Customer complaints	Customer	Automated
Romano et al. [47]	Product reviews	Movie box office	Customer	Automated
Duan, Gu, and Whinston [18]	Product reviews	Movie box office	Product	Automated
Abbasi and Chen [1]	Email	Enron scandal communications	Employee	Automated
Schumaker and Chen [52]	News articles	Stock market	Market	Automated
Antweiler and Frank [5]	Online Forums	Stock market	Market	Automated
Tetlock et al. [56]	News articles	Stock market	Market	Automated
Loughran and McDonald [38]	10-K filings	Stock market	Market	Automated
O 'Leary [42]	Blog postings	Stock market	Market	Automated
Li and Wu [37]	Online forums	Sports	Market	Automated
Zhang et al. [66]	News articles	Health epidemics	Environment	Automated
Wiebe et al. [46,62,63]	News articles	News/Politics/Business/Travel/English Literature	Environment	Automated
Finch and Luebbe [24]	Public listserv	Fly fishing gear	Product	Manual
Finch [23]	Newsgroup	Power tools	Product	Manual
Abbasi, Chen, and Salem [2]	Online forums	Movie reviews, and political discussions	Product	Automated
This study	Online forums	Vehicle defects	Product	Automated

Word *senses* are shown in CAPS above. The number after the pound sign (#) identifies different specific word senses. The word *categories* are given in parentheses. Word sense disambiguation [3,31,53] is important. Without disambiguation, for example, the object "TIRE" (identified as word sense "TIRE#1") may be misinterpreted as the negative verb "TIRE" (word sense "TIRE#2" = to fatigue) which is unrelated. Stemming [45] ensures that multiple occurrences of the same word in different morphological forms (e.g., "bubble", "bubbling", "bubbled") can be reconciled when analysis and reporting are performed.

In the fourth step (4.) vehicle experts tag a large sample of discussions. The vehicle experts must assess whether a defect is being discussed, how critical the defect is, and what component the defect relates to. Further annotations can also be added (e.g., specific vehicle model, year, engine, and/or build; what the implications of the defect are; why an item should not be regarded as a defect).

Once a sufficiently large training set of discussion threads has been tagged, automated text mining (5.) can be employed to classify new threads by defect existence and criticality.

Both the automated (2., 3., 5.) and human (4.) text mark-up are stored in the discussion and defect database, to allow comparison of discussions, defects, components and vehicles. Collation (6.) of threads along each of these dimensions allows analysis (7.) of the contents of the database using traditional On-Line Analytical Processing (OLAP) tools [15], such as Excel PivotCharts. Note that the analysis process may inform subsequent iterations of crawling (2.) to gather further discussion data from additional sources, or for different vehicle models, or with alternative filtering criteria, so the process is iterative.

Finally, Vehicle Quality Managers – in our study, assumed to be product managers for Honda, Toyota, or Chevrolet motor vehicles – use insights from their analysis (7.) to inform defect remediation planning (8.), execution (9.), and tracking (10.). As defects are rectified, discussion tags are updated (4.) by the vehicle expert to reflect successful resolution. This allows the Vehicle Quality Manager to focus their attention on remaining defects.

5. Methodology

In order to better understand the reporting of vehicle quality management issues in social media, we undertook a large empirical study [25] of online vehicle enthusiast discussion forums, specifically using

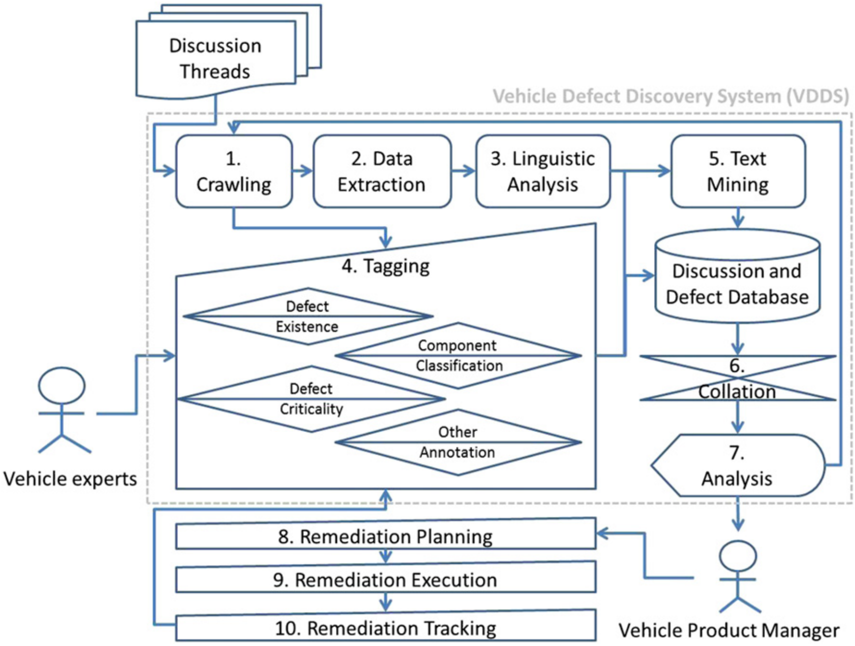


Fig. 1. Process for vehicle quality management using social media.

the case study method. The case study method of theory building is widely accepted [7,19,28,40,50,65]. We followed a research design consistent with earlier studies of consumer postings [23], and adhering to the guidelines of content analysis research [41].

5.1. Data sampling

We began with a small pilot study to test the data gathering approach. For the exploratory pilot, we employed 36 senior-level undergraduate business students to independently tag a sample of 900 discussion threads (50 threads per student, 2 students per thread) sampled from Honda's two most popular models (Civic and Accord). Discussion threads were sourced from Honda-Tech.com, the world's most popular Honda enthusiast discussion forum. We first downloaded the 2010 snapshot of the United Stated Department of Transportation, National Highway Traffic Safety Administration (NHTSA), Office of Defect Investigations, vehicle safety complaint database. We extracted the top 200 most frequent component description keywords for this dataset, which represent the most commonly defective vehicle components. We then extracted the top 900 threads from Honda-Tech.com, which contained the most mentions of these defective components.

For the pilot study, student agreement on defect existence was 82% while agreement on defect criticality (safety vs. performance vs. no-defect) was 77%. The pilot study satisfied us that performance and safety defects are commonly discussed in online vehicle enthusiast forums, but revealed that the criticality of defects, and sometimes even their existence, is difficult for a layperson to judge. We found that defect and criticality determination requires expert knowledge of vehicle components. For example, accurate defect determination requires, at minimum, that the person making the classification knows: what component a part relates to (e.g., "What is a pinion/cowl/distributor/…?"); the factory-warranted life of the component (e.g., "Should the alternator have lasted 50,000 miles?"), is the component described OEM or aftermarket (e.g., "The tread is separating on these Bridgestone tires, but were they a factory install or owner modification?"), and how serious the implications of component failure are (e.g., "Is a lighted O_2 sensor indicator a performance or safety concern?").

For the production study, we therefore employed three independent motor vehicle domain experts to conduct the thread tagging: two graduate students and one senior undergraduate student majoring in Automotive Engineering, from a nationally recognized Vehicle Systems and Safety research group. The experts were required to read each forum thread assigned to them and make decisions on the following three questions:

1. Does the thread discuss a product defect?
2. If yes, how critical is the defect?
3. What vehicle component is primarily affected by this defect?

We expanded the study to three different vehicle marques: Honda, Toyota, and Chevrolet. For the production study we obtained the permission of the forum owners and crawled all threads available at Honda-Tech.com, ToyotaNation.com, and ChevroletForum.com as of June 2010. We then found the top 1500 threads from each of these three forums, which contained the most occurrences of the top 200 defective components from NHTSA complaints. 1500 was chosen as the sample size that would produce an acceptable margin-of-error of ±2.5% at the 95% confidence level. We eliminated sub forums about news, racing and advertising because discussions in those forums generally do not relate to defects. To remove short and low quality discussions, we applied two additional criteria: (1) thread contained at least 50 words; (2) thread comprised at least 2 posts. Table 2 shows some summary information related to the three forums. We included two brands (Honda and Toyota) in our *Training Set*, to help ensure that any models built would be generalizable. To test the generalizability of our models, we used data from the third brand (Chevrolet) as the *Validation Set*.

We extracted the following information from each thread: title, text of each posting, date and time of each posting, user identifier of each posting, and number of times the thread had been viewed.

5.2. Refined constructs

Threads were tagged by defect existence, defect criticality, and component affected. Given the unexpected ambiguity encountered by taggers in the pilot study, we tightened the definition of our constructs for the production study. Strict protocols were developed to define these constructs.

5.2.1. Coding scheme for defect existence and criticality

The construct describing the severity of the defect required special elaboration. The motor industry defines five Safety Integrity Levels (SILs), or controllability categories, for classifying operability hazards according to severity [32]. SILs are judged based on both ability of the vehicle operator to avoid a severe outcome, and severity of the expected outcome. In ascending order of severity, the Safety Integrity Levels are "nuisance only (safety not affected)", "distracting (at worst, minor outcomes)", "debilitating (severe outcomes)", "difficult to control (very severe outcomes)", and "uncontrollable (extremely severe outcomes)". For the purposes of tagging discussion threads, we employed a simplification of these SILs, and defined three categories:

1. "*Performance* defects" are defects that affect customer satisfaction, but are unlikely to lead to injury or death. Examples include a malfunctioning air conditioner, broken radio, innocuous squeaking or rumbling noises, or a defective ignition switch. Performance defects correspond with the SILs "nuisance only" and "distracting" (SILs 0 to 1). An actual example of a performance defect was:

   > "*Post: can anyone help me with the codes? my car is thrwing [sic] a 1 and the rpm gauge goes crazy and the engine light turns on. thanks in advance.*
   > *Reply: a 1 is your 02 [oxygen content]. but the needle indicates a dizzy [distributor] problem.*"

2. "*Safety* defects" are defects that could lead to serious injury or death. For example, a sticky accelerator pedal, or a rupturing airbag inflator that expels metal fragments at high velocity. Safety defects correspond with the SILs "debilitating", "difficult to control", and "uncontrollable" (SILs 2 to 4). In the following example, a speedometer malfunction represents a serious safety issue as the driver may be unaware of their actual road speed:

   > "*Post: the speedometer is stuck at 0 … I don't know what to do to get it working.*
   > *Reply: check fuses and wiring. otherwise try replacing*"

3. "*Junk* (i.e., non-defects)", includes postings that do not describe manufacturing defects. For instance, postings requesting vehicle information, some for-sale listings (only those where no vehicle defects are described), routine service questions, or hobbyists asking for assistance with after-market vehicle modifications.

Table 2
Data sources.

	Forum	Number of users	Number of sub-forums	Total threads	Sample size (threads)
Training set	Honda Tech	201,975	34	1,316,881	1500
	Toyota Nation	78,203	100	216,599	1500
Validation set	Chevrolet Forum	18,498	48	26,936	1500

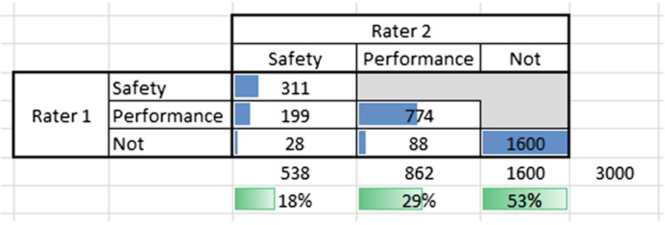
"*Post: all you guys are gonna make fun of me, but this is my first stick car, and i wanted 2 know what exactly is chirping the tires and exactly how you do it. please help*"

The training set was tagged independently by two of the domain experts. To ensure tagging consistency, each expert was required to first independently tag 500 Honda threads and 500 Toyota threads. Tagging inconsistencies were noted and the experts constructed a two-page protocol document to govern classification. For example, issues with non-OEM parts, or failure of OEM parts outside of factory warranted lifetime, were classified as "Junk" as they were deemed to be outside of the manufacturer's concern. Various components were specified that were to be marked defective if they did not last the lifetime of the vehicle (e.g., door handles, electronic wiring harnesses, radiator hardware excluding hoses). Clear cases of product abuse by the customer (e.g., high mileage vehicles that had not been adequately maintained, botched owner-conducted repairs, vehicles that had been left excessively long in storage) were discounted and tagged as 'non-defect'. The experts adjusted their discrepant tags to conform to the agreed protocol.

After all 3000 training threads (1500 Honda + 1500 Toyota) were tagged, a Kappa statistic (κ) of inter-rater reliability [12] was computed between the experts. For the "Defect vs. Non-Defect" variable, κ was 0.96, whereas for "Performance vs. Safety vs. Junk" κ was 0.90. In both cases, we can conclude that the tagging can be considered reliable. Fig. 2 summarizes the expert classification of the training set in the production study and shows that the desired improvement in classification consistency over the pilot study was achieved. Finally, the two experts were asked to resolve their remaining differences to construct a gold standard Training Set for later analysis. For the 315 cases (10% of cases) where expert disagreement was not resolved we took the more conservative of the two judgments (e.g., we classified the thread as a safety defect if *any* expert felt it was a safety issue, or a performance defect if one expert felt it was a performance defect and the other felt the item was not a defect.). The Validation Set, comprised of 1500 Chevrolet threads, was independently tagged using the same protocol by the third domain expert.

5.2.2. Classification scheme for component affected

To simplify the classification task, we used 15 major component classification labels, abridged from the 339 NHTSA unique component descriptions available. For example, the NHTSA provides 45 categories of "Service Brake" complaints and 29 categories of "Power Train (Transmission)" related complaints — we simplified these to the categories "Braking" and "Transmission" respectively. The 15 major component classification labels we used were as follows: [Acoustics], [Air Conditioning], Airbag, Braking, Electrical system, Engine, Lights, Seat Belts, Steering, Structure and Body, Suspension, Transmission, Visibility (Windows), Wheels and Tires, and Other. The categories in square brackets were added as they cover a substantial set of performance related defects that are largely ignored by the NHTSA's safety-focused classification scheme. The component classification scheme was used to generate a variety of managerial reports (e.g. safety defects by component; performance defects by component) useful to the vehicle quality management professional. However, the component classification scheme is not directly used in the remainder of the work described in this paper as this work focuses solely on predicting defect existence and criticality, rather than component affected.



		Rater 2			
		Safety	Performance	Not	
Rater 1	Safety	311			
	Performance	199	774		
	Not	28	88	1600	
		538	862	1600	3000
		18%	29%	53%	

Fig. 2. Summary of rater agreement by defect existence and criticality for the Training set.

5.3. Computed metrics

For each individual thread studied, we used automated routines to compute all metrics described in Table 3. First, we counted the number of words in each metric category (e.g. NEG, POS, HOS, …) in each thread. Next, to remove length-bias (long postings have more words), we normalized each word count metric by dividing by the total number of words in the thread and multiplying by 1000 to obtain the *incidence* (per thousand words) for each word category. We computed each metric for the full thread. For example, for Harvard Negative words (NEG), we computed the *incidence* of negative words (per thousand words) in the full thread. For the OpinionFinder system, which assesses subjectivity and polarity at the phrase-level, rather than the word-level, we used OpinionFinder 1.5 to determine the ratio of subjective expressions to all expressions (OF_SUBJ_RATIO), and the ratio of negative contextual polarity expressions to all expressions (OF_NEG_RATIO). "All expressions" includes objective expressions, subjective expressions, and expressions whose subjectivity could not be determined by OpinionFinder.

5.4. Relative term weights

To find out which terms were relatively more frequent in safety vs. performance vs. non-defects in the Training Set, we grouped threads by defect existence and criticality, and concatenated the threads into four sets of threads: safety defects, performance defects, all defects and all non-defects. We ran the Harvard General Inquirer on the full text of each group and computed the relative term weights for all term senses in each of the four groups of threads, in order to find which term senses were relatively more prevalent in defects versus non-defects, and in safety issues versus other threads. Relative term weights (*rtw*) for each word sense were computed as follows:

$$rtw_{\mathrm{DEFECT}} = \frac{df_{\mathrm{defects}} + 0.001}{df_{\mathrm{nondefects}} + 0.001}$$

$$rtw_{\mathrm{SAFETY}} = \frac{df_{\mathrm{safety\_issue}} + 0.001}{df_{\mathrm{not\_safety\_issue}} + 0.001}$$

where *df* measures the document frequency of a word sense

Table 3
Computed metrics for each thread.

Metric name	Metric description
WORDS	Count of all words
EXCLAM	Incidence of exclamations
Metrics derived from Harvard General Inquirer/Harvard Psychosocial Dictionary H4N [34,55]	
NEG/POS/HOS/VICE	Incidence of negative/positive/hostile/vice words
DIFF	Positive word incidence, less negative word incidence
RATIO	Ratio of positive to negative words
Metrics derived from OpinionFinder [46,62,63]	
OF_SUBJ_RATIO	Ratio of subjective expressions to all expressions.
OF_NEG_RATIO	Ratio of *negative* subjective expressions to all expressions.
Metrics from financial domain [38]	
FIN_POS/FIN_NEG/ FIN_LIT	Incidence of positive/negative/litigious financial domain words
FIN_DIFF	Positive word incidence, less negative word incidence
FIN_RATIO	Ratio of positive to negative words

appearing in the entire collection of a known category (defect, nondefect, safety_issue, not_safety_issue). 0.001 is a small fraction added to avoid divide-by-zero errors. *rtw* was adapted from [21].

To ensure the results would generalize across multiple brands, we conducted this analysis separately for each of Toyota and Honda. We retained only brand-independent terms with high relative prevalence in a given category (e.g., safety defects) across *both* brands.

In this study, we completed only a unigram analysis (single word terms). An *n*-gram analysis (multi-word terms) would be of interest for future studies.

6. Results and evaluation

Next we provide overall descriptive statistics that summarize our dataset (Section 6.1). We evaluate the performance of traditional sentiment analysis on our data set (Section 6.2), and describe the automotive smoke words discovered by our decision support system for vehicle defect discovery (Section 6.3). Finally, we assess the cross-brand generalizability of our vehicle defect discovery approach (Section 6.4).

6.1. Overall descriptive statistics

The 4500 threads in our data set were from 3 brands (Honda, Toyota, and Chevrolet). The threads were from 113 unique sub-forums and discussed 61 unique vehicle models and 89 different model generations. (A model generation includes all model years since the previous major design overhaul). On average, each vehicle discussion thread contained 8 postings, by 5 different users, and had been viewed 492 times, showing the forums were heavily read by enthusiasts, particularly those searching for information (and not necessarily making posts). The average thread contained 23 sentences with a total 502 words (min 50 words; max 8586 words), though only 151 (30%) of these words were typically unique. Each word contained on average 4 characters, partly due to the frequent use of abbreviations and technical lingo (e.g., "tranny" = "transmission", "dizzi" = "distributor"; "carb" = "carburetor"; "rolla" = "Toyota Corolla"; "eng mgnt" = "engine management").

Tagging of the 4500-thread training set consumed 30 person days of effort expended over a period of 11 weeks. The average time to tag a thread was 1.8 min, with a standard deviation of 2.2 min. That is consistent with earlier reading research showing average American adults reading from a computer monitor at approximately 180 words per minute [67]. Assuming a 40-hour work week and 50 weeks per year, we can project that to manually tag the full data set of 1.6 million total available Honda, Toyota, and Chevrolet threads would consume 27 person-years of effort.

6.2. Sentiment analysis

We were interested in testing the presumption that a large number of negative sentiment words predicts the existence of vehicle defects (see Section 2.4 earlier). We used the Harvard General Inquirer dictionary of positive and negative keywords [34,55], and the Financial dictionary of positive and negative keywords [38], as described in the Methodology section (Section 5.3 earlier) to determine the incidence of positive and negative *keywords* found on average in safety-critical postings, non-safety-critical postings, and non-defect postings. Similarly, we used OpinionFinder [46,62,63] to determine the incidence of negative subjective expressions (*phrases*).

Overall thread sentiment was moderately positive, with full threads containing on average 30 positive and 27 negative words per thousand words, and first postings in the thread containing on average 52 positive and 45 negative words per thousand words. Using the financial domain sentiment dictionaries, first posts contained 6 positive and 33 negative words from the domain of finance, per thousand, while the thread as a whole contained 5 positive words and 37 negative words per thousand, so there was a relative abundance of negative words from the financial domain across all threads. Each thread contained on average 28 strong modal verbs and 4 weak modal verbs, showing that threads heavily contained strong opinions and instructions. OpinionFinder has two subjectivity classifiers. The 1st subjectivity classifier found that, on average, 11% of expressions across all threads were subjective. This figure was much higher (38%) for OpinionFinder's 2nd subjectivity classifier. OpinionFinder's polarity recognizer found that, on average, 19% of all expressions across all threads were negative expressions.

We ran an Analysis of Variance (ANOVA) on these statistics. The results are shown in Table 4, which shows the means for each of the metrics we computed, across each thread type (safety vs. non-safety; defect vs. non-defect). An asterisk (*) with gray shading indicates where the difference between means is significant at the 99% confidence level. The superscript number ([1] or [2]) in square brackets alongside OpinionFinder's subjectivity ratios specifies which of OpinionFinder's two classifiers was used to find the greatest distinction between the thread types.

Interestingly, we found that negative sentiment is *not* positively correlated with defects. Notice that all threads, whether they relate to safety or not, or defects or not, have the same average incidence of Harvard negative words (NEG): roughly 27 per thousand. Looking at the ratio of positive to negative words (RATIO) we see that, on average, safety critical threads (32.6 positive words per thousand) are significantly *more positive* than other threads (29.7 positive words per thousand). Safety critical threads also contain, on average, substantially *fewer* negative (FIN NEG) words – 29.1 per thousand – than other threads (38.3 per thousand). Similarly, defect threads contain significantly *fewer* hostile (HOS) and vice (VICE) words than non-defects, which again is an unexpected result. Finally, and again counter to conventional wisdom, we found that the prevalence of subjective expressions (OF_SUBJ_RATIO) and negative subjective expressions (OF_NEG_RATIO) is *smaller* in defects than in non-defects.

We investigated further to determine why it is the case that traditional sentiment words are not predictive of vehicle safety issues or defects. We found that users were prone to be negative about performance issues (e.g., air conditioning failures) even if these did not affect their safety, and users were also prone to use negative sentiment words even if not reporting a defect with the vehicle. In the following example, which is a junk posting, Harvard negative words, indicated with (−), abound:

> "*Have you ever hit* (−) *something? I did once, but she stopped screaming* (−) *after a few minutes of smuthering [sic] with my jacket Ooops!! j/k [sic] No, but I have rear ended someone. … all of the sudden, the girl in front of me slams* (−) *on her brakes. I hit* (−) *mine too but wasn't able to stop in time. Neither car had any damage* (−). *Thank god for the 15mph bumper.*"

Table 4
Means of computed metrics for each thread type.

		Thread type			
		Safety-critical	Non-safety critical	Defect	Non-defect
METRIC	WORDS	554	547	513	580
	EXCLAM	2.5	2.5	2.5	2.5
	NEG	26.7	27.6	27.7	27.2
	POS	32.6*	29.7*	30.7	29.9
	HOS	6.3	6.2	5.9*	6.5*
	VICE	7.3	8.0	7.5*	8.1*
	DIFF	5.9*	2.1*	3.0	2.6
	RATIO	1.5*	1.4*	1.4	1.4
	OF_SUBJ_RATIO	0.10[1]	0.11[1]	0.36*[2]	0.40*[2]
	OF_NEG_RATIO	0.18	0.19	0.18*	0.20*
	FIN_POS	4.5	4.9	4.6	5.0
	FIN_NEG	29.1*	38.3*	35.7	37.5
	FIN_LIT	0.1	0.2	0.2	0.2
	FIN_DIFF	–24.6*	–33.4*	–31.1	–32.5
	FIN_RATIO	0.3	0.2	0.2	0.3

In the following example, which is a safety defect, the consumer uses only 2 negative sentiment words, indicated with (−) alongside, and 6 positive sentiment words, indicated with (+) alongside, from the Harvard dictionary:

> "*my passenger side brake light (+) will not work. Its like this, if the headlights are off the brake light (+) doesn't work, but if i turn the lights (+) on the light (+) comes on but its just the running light (+). And when i hit (−) the brakes the light (+) still doesn't light up. Anybody know what the problem (−) might be??*"

This example clearly illustrates that positive words from the General Inquirer (e.g., "light") are not necessarily positive in the automotive industry (e.g., here "light" just indicates the component affected).

In other domains, sentiment analysis has been successfully used to find product complaints [2,43,57]. In the automotive domain, however, we discovered that conventional sentiment determination may be a poor indicator of both defect existence and defect criticality.

6.3. Automotive 'smoke' words

Given our finding that traditional sentiment determination is ill-suited to vehicle defect detection and prioritization, we set to identify, from the training set, an alternative set of marker words for this purpose. We call these Automotive 'Smoke' Words. These words can be partitioned into two sets:

1) SMOKE_DEFECT: list of words that are substantially more prevalent in defects than in non-defects.
2) SMOKE_SAFETY: these words are substantially more prevalent in safety issues than in other postings.

For the interested reader, a detailed description of the Automotive Smoke words can be found in the Online Supplement that accompanies this paper.

To test the significance of our smoke words, we counted the smoke words in each thread in our training set (1500 Honda and 1500 Toyota threads), and conducted an ANOVA analysis. Table 5 reports the results. An asterisk (*) with gray shading indicates that the difference between means is significant at the 99% confidence level. As is evident from the table, defects contain a significantly higher incidence of SMOKE_DEFECT words than non-defects, and safety issues contain a significantly higher incidence of SMOKE_SAFETY words than threads that do not relate to safety issues.

We also performed a logistic fit to determine the relationship between each of our investigated metrics and the dependent variables (SAFETY problem exists, DEFECT exists). We computed a χ^2 association statistic to test the significance of the association between the computed metrics (X variables) and the SAFETY and DEFECT issues (Y variables). Table 6 summarizes the results. Metrics (X variables) with significant relationships to the dependent (Y) variables at the 99% confidence level ($\chi^2 \geq 6.64$ where $\alpha = 0.01$, $df = 1$) are shaded in gray. Associations that exceed the 99.99% confidence level ($\chi^2 \geq 10.83$ where $\alpha = 0.001$, $df = 1$) have been shaded in black. For OpinionFinder, two subjectivity classifiers were available; we used both but, for brevity, show only the results from the classifier which produced the OpinionFinder results with the strongest fit. It is interesting to observe that for the few traditional metrics which are highly significant (POS, DIFF, OF_SUBJ_RATIO, OF_NEG_RATIO, FIN_NEG, FIN_DIFF) the association is counter to the conventional wisdom: our findings are the first to reveal that safety defects have a *higher* incidence of positive words than other postings ($\chi^2 = 22.0$ for POSS, and $\chi^2 = 18.4$ for DIFF), and that defects, including safety defects, have a *lower* incidence of both subjective phrases, and negative words and phrases, than other postings (on regular DEFECTS, $\chi^2 = 22.27$ for OF_SUBJ_RATIO, $\chi^2 = 10.97$ for OF_NEG_RATIO; on SAFETY defects, $\chi^2 = 4.95$ for OF_SUBJ_RATIO, $\chi^2 = 74.0$ for FIN_NEG, and $\chi^2 = 66.0$ for FIN_DIFF). The Automotive Smoke Words show a much stronger correlation: safety issues have a significantly higher incidence of SMOKE_SAFETY words than other postings ($\chi^2 = 379.8$), and defects have a significantly higher incidence of SMOKE_DEFECT words than non-defects ($\chi^2 = 707.8$). Notice the χ^2 association statistic is much stronger, and the characteristic sigmoid (s-shaped) curve of strong logistic relationships is highly pronounced, between SMOKE words and their respective thread category (SAFETY and DEFECT), which confirms the significance of the smoke words and suggests they have very high explanatory power for SAFETY and DEFECT.

Table 5
Mean incidence (per thousand) of *smoke* words for each thread type.

		Thread type			
		Safety-critical	Not safety critical	Defect	Non-defect
METRIC	SMOKE_DEFECT	94*	70*	91*	60*
	SMOKE_SAFETY	78*	55*	60	58

The procedure for using the Automotive Smoke Word list to ultimately predict whether a posting, *p*, is related to a regular defect and/or a safety defect is as follows:

a) First, compute $words_p$ = the total number of words in posting *p*.
b) Next, the posting must be processed through Harvard General Inquirer, to identify unique word senses (e.g., "BAG#1") for each word. For lexical items not present in the General Inquirer lexicon (e.g., "NHTSA"), these should be counted separately. The output of this step should be a word sense list, showing the number of occurrences of each word sense in posting *p*.
c) Next, compute $raw\_smokes\_DEFECT_p$ = the count of how many times any word senses from SMOKE_DEFECT occur in the word sense list generated in step b) above. Also, compute $raw\_smokes\_SAFETY_p$ = the count of how many times any word senses from SMOKE_SAFETY occur in the word sense list generated in step b) above.
d) Next, normalize the raw smoke word counts from the previous step, so that unusually long (or short) postings do not bias the results. To normalize, compute the number of smoke words *per thousand words* in the posting (i.e., the prevalence of smoke words), rather than the absolute number of smoke words. The formulae are as follows:

$$prevalence\_smokes\_DEFECT_p = \frac{raw\_smokes\_DEFECT_p}{words_p} \times 100$$

$$prevalence\_smokes\_SAFETY_p = \frac{raw_smokes_SAFETY_p}{words_p} \times 100.$$

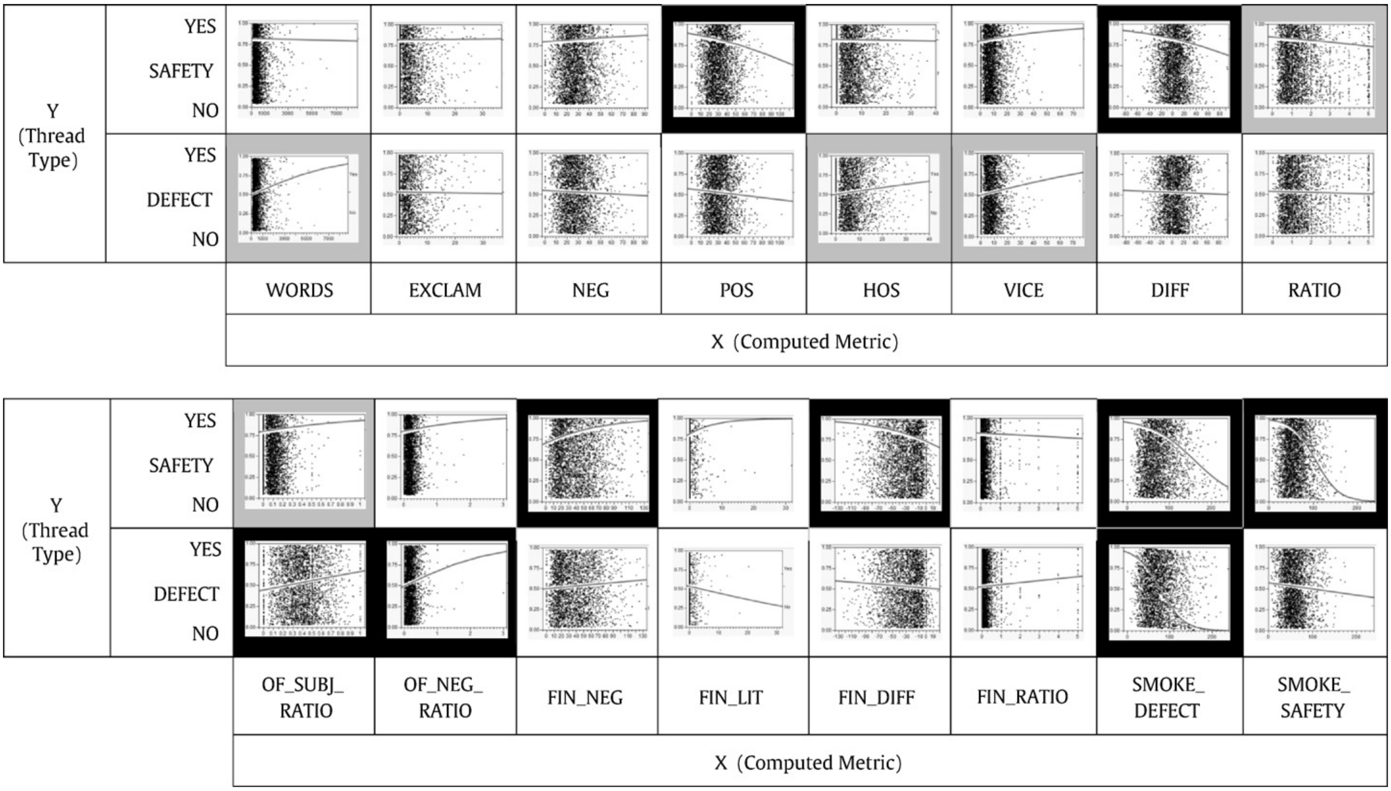
e) Finally, substitute the prevalences of the smoke words, from the previous step, into the applicable (DEFECT prediction or SAFETY defect prediction) logistic model, to obtain a prediction of the likelihood that the posting is a defect:

$$Prob\left[DEFECT_p\right] = \frac{1}{1 + e^{2.76 + \left(-0.04 \times prevalence\_smokes\_DEFECT_p\right)}}$$

$$Prob\left[SAFETY_DEFECT_p\right] = \frac{1}{1 + e^{3.99 + \left(0.04 \times prevelance\_smoke\_SAFETY_p\right)}}.$$

The parameters for these logistic models were obtained from the logistic regressions. Any values greater than 0.5 for *Prob[DEFECT$_p$]* are predicted to be a defect. Similarly, any values greater than 0.5 for *Prob[SAFETY$_p$]*are predicted to be a safety defect.

Table 6
Logistic fit of each computed metrics to each thread type (Training set: Honda and Toyota threads).



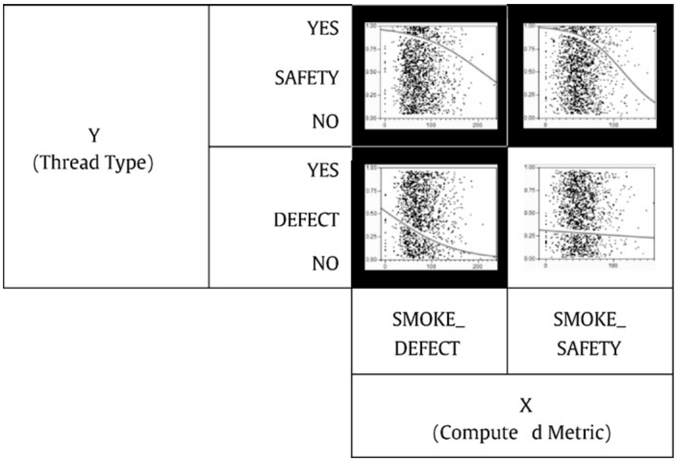
6.4. Generalizability (external validity)

To ascertain whether the smoke words we obtained for Honda and Toyota were generalizable to another, *unseen* brand we employed the Validation Set (1500 Chevrolet threads) which had been independently tagged by the third expert. Table 7 shows that the logistic fit between actual Chevrolet defects (Y variables) and the incidence our Automotive Smoke Words (X variables) was strong at the 99.99% confidence level ($\chi^2 = 111.1$ for incidence of SMOKE_SAFETY vs. actual SAFETY, and $\chi^2 = 54.6$ for incidence of SMOKE_DEFECT vs. actual DEFECT). Our results therefore have significant external validity.

7. Limitations

Due to biases inherent in web data and in document collection processes, it is important for manufacturers to be aware that the defects discovered using this approach are unlikely to be fully representative of the population of defects. Instead, this defect discovery process should be viewed as both exploratory and supplementary to other diagnostic and data gathering exercises (e.g., mechanical tests, customer surveys, customer complaints). The defect discovery process described here is intended simply to highlight defects that would have been difficult to find and analyze through manual consultation of thousands or millions of documents on the web.

Table 7
Logistic fit of smoke words to each thread type (Validation set: Chevrolet threads).



Threads were tagged through a manual process of human expert annotation. In any complex sphere, experts may overlook reported defects: for example, some automotive experts may discount reports of mysterious unintended acceleration, or reports of tire tumors, believing these are driver recklessness (excessive speed, or hitting a pothole) rather than vehicle defects. The data is prone to these training biases, and should be supplemented with other sources to better highlight emergent issues that warrant further investigation.

8. Implications for practice and research

The findings of this study have a number of implications for practitioners:

- Defects are commonly discussed in online vehicle forums, but interspersed among a large proportion of junk postings. This implies that vehicle quality management professionals would greatly benefit in terms of productivity by employing a Vehicle Defect Discovery System (VDDS) like ours to sift defects from unrelated postings.
- A traditional sentiment analysis cannot be relied upon to predict defects as, counter to the conventional wisdom, safety defects have a *higher* incidence of positive words than other threads, and a *lower* incidence of negative words and subjective expressions than other threads. This implies that, while practitioners can continue to use sentiment analysis to identify consumer complaints on the web, sentiment analysis should not be the primary mechanism used to locate vehicle defects.

- The incidence of automotive smoke words strongly predicts the existence of safety and performance defects across multiple brands. This implies that practitioners should use web crawlers, in conjunction with a word-sense disambiguation tool, and an automotive smoke word list, as described in our defect management process and VDDS, to scan multiple social media forums (discussion boards, Facebook, Twitter) for vehicle defects.
- The vehicle defect predictions made by the automated defect discovery and prioritization process introduced in this paper are strongly associated with manual defect criticality annotations by vehicle experts. The VDDS provides robust and generalizable defect discovery and classification. This implies that practitioners should employ the automated defect identification and prioritization process we described in conjunction with their manual defect tracking process.

For researchers, the implications of our findings are as follows:

- We have confirmed that generic sentiment words are indeed highly domain-specific, as contended in earlier studies for other industries [38,42]. This implies that researchers may need to define domain-specific sentiment words for other industries.
- We have defined and validated a method and system for automated defect detection and prioritization in the automotive industry using linguistic analysis and text mining. Our results show that this method can indeed uncover defects related to both safety and performance. Many steps of the method can be automated; we created a VDDS using software components for crawling of data, auto tagging, and classification of threads into defect categories. The proposed method and system could be adapted and tested in other domains related to quality control and product management.

9. Summary, conclusions, and future work

In this paper we found that auto enthusiast discussion forums contain substantial content related to motor vehicle defect existence and criticality. We found that conventional sentiment analysis, which is successful in the identification of complaints in other industries, fails to distinguish defects from non-defects and safety from performance defects. We compiled an alternative set of automotive smoke words that have higher relative prevalence in defects vs. non-defects, and in safety issues vs. other postings. These smoke words, discovered from Honda and Toyota postings, generalize well to a third brand, Chevrolet, which was used for validation. We implemented our findings in a novel Vehicle Defect Discovery System (VDDS) that provides robust and generalizable defect discovery and classification. This paper has shown that vehicle quality management can be supported by appropriate analysis of social media postings.

In future work, we intend to extend the VDDS. We plan to expand upon the current unigram (single term) analysis of postings, and determine whether rule induction methods, neural networks, or other text mining techniques can be developed to enhance the defect detection and sorting process. We also intend to explore alternative social media (Twitter, Facebook, …), additional linguistic features of the text, and a greater selection of vehicle brands. With the volume of social media postings expanding rapidly, we expect that the need for automated business intelligence tools for the exploration of this vast and valuable data set will continue to grow.

Acknowledgements

The authors are grateful to Dr. Mehdi Ahmadian, Director of the Center for Vehicle Systems and Safety, Virginia Tech, for his helpful feedback, guidance, and assistance. This research is partly supported by the Natural Science Foundation of China (Grant #70872089 and #71072129) and the National Science Foundation (Grant #DUE-0840719).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.dss.2012.04.005.

References

[1] A. Abbasi, H. Chen, CyberGate: a system and design framework for text analysis of computer-mediated communication, MIS Quarterly 32 (4) (2008) 811–837.
[2] A. Abbasi, H. Chen, H.A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Transactions on Information Systems 26 (3) (2008).
[3] E. Agirre, P. Edmonds, Word Sense Disambiguation: Algorithms and Applications, Springer, Dordrecht, 2007.
[4] V. Anand, W.H. Glick, C.C. Manz, Thriving on the knowledge of outsiders: tapping organizational social capital, The Academy of Management Executive 16 (1) (2002) 87–101.
[5] W. Antweiler, M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, Journal of Finance 59 (3) (2004) 1259–1294.
[6] S. Argamon, C. Whitelaw, P. Chase, S. Dhawle, S.R. Hota, N. Garg, S. Levitan, Stylistic text classification using functional lexical features, Journal of the American Society for Information Science and Technology 58 (6) (2007) 802–822.
[7] I. Benbasat, D.K. Goldstein, M. Mead, The case research strategy in studies of information systems, MIS Quarterly 11 (3) (1987) 369–386.
[8] J.S. Brown, P. Duguid, Organizational learning and communities-of-practice: toward a unified view of working, learning, and innovation, Organization Science 2 (1) (1991) 40–57.
[9] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann, 2002.
[10] M. Chau, H. Chen, Comparison of three vertical search spiders, IEEE Computer 36 (5) (2003) 56–62.
[11] H. Chen, H. Fan, M. Chau, D. Zeng, Testing a cancer meta spider, International Journal of Human Computer Studies 59 (2003) 755–776.
[12] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1) (1960) 37–46.
[13] D. Constant, L. Sproull, S. Kiesler, The kindness of strangers: the usefulness of electronic weak ties for technical advice, Organization Science 7 (2) (1996) 119–135.
[14] K. Coussement, D. Van den Poel, Improving customer complaint management by automatic email classification using linguistic style features as predictors, Decision Support Systems 44 (4) (2008) 870–882.
[15] A. Datta, H. Thomas, The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses, Decision Support Systems 27 (3) (1999) 289–301.
[16] K. De Valcka, G.H. Van Bruggen, B. Wierenga, Virtual communities: a marketing perspective, Decision Support Systems 47 (3) (2009) 185–203.
[17] P.M. Di Gangi, M.M. Wasko, Steal my idea! Organizational adoption of user innovations from a user innovation community: a case study of Dell IdeaStorm, Decision Support Systems 48 (1) (2009) 303–312.
[18] W. Duan, B. Gu, A.B. Whinston, Do online reviews matter? — An empirical investigation of panel data, Decision Support Systems 45 (4) (2008) 1007–1016.
[19] K.M. Eisenhardt, Building theories from case study research, Academy of Management Review 14 (4) (1989) 532–550.
[20] O.A. El Sawy, Personal information systems for strategic scanning in turbulent environments: can the CEO go online? MIS Quarterly 9 (1) (1985) 53–60.
[21] W. Fan, M.D. Gordon, P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, Decision Support Systems 40 (2) (2005) 213–233.
[22] W. Fan, L. Wallace, S. Rich, Z. Zhang, Tapping the power of text mining, Communications of the ACM 49 (9) (2006) 76–82.
[23] B.J. Finch, Internet discussions as a source for consumer product customer involvement and quality information: an exploratory study, Journal of Operations Management 17 (5) (1999) 535–556.
[24] B.J. Finch, R.L. Luebbe, Using Internet conversations to improve product quality: an exploratory study, International Journal of Quality and Reliability Management 14 (8) (1997) 849–865.
[25] B.B. Flynn, S. Sakakibara, R.G. Schroeder, K.A. Bates, E.J. Flynn, Empirical research methods in operations management, Journal of Operations Management 9 (2) (1990) 250–284.
[26] C. Fornell, B. Wernerfelt, Defensive marketing strategy by customer complaint management: a theoretical analysis, Journal of Marketing Research 24 (4) (1987) 337–346.
[27] J. Gerdes, B.B. Stringham, R.G. Brookshire, An integrative approach to assess qualitative and quantitative consumer feedback, Electronic Commerce Research 8 (4) (2008) 217–234.
[28] B. Glaser, A. Straus, The Discovery of Grounded Theory: Strategies of Qualitative Research, Wiedenfeld and Nicholson, London, 1967.
[29] R. Gopal, J.R. Marsden, J. Vanthienen, Information mining — reflections on recent advancements and the road ahead in data, text, and media mining, Decision Support Systems 51 (4) (2011) 727–731.
[30] M. Hora, H. Bapuji, A.V. Roth, Safety hazard and time to recall: the role of recall strategy, product defect type, and supply chain player in the U.S. toy industry, Journal of Operations Management 29 (7) (2011) 766–777.
[31] N. Ide, J. Veronis, Word sense disambiguation: the state of the art, Computational Linguistics 24 (1) (1998) 1–40.

[32] P.H. Jesty, K.M. Hobley, R. Evans, I. Kendall, Safety analysis of vehicle-based systems, in: F. Redmill, T. Anderson (Eds.), Lessons in System Safety, Proceedings of the 8th Safety-Critical Systems Symposium (SCSS), Springer, London, 2000.
[33] Z. Jourdan, R.K. Rainer, T.E. Marshall, Business intelligence: an analysis of the literature, Information Systems Management 25 (2) (2008) 121–131.
[34] E. Kelly, P. Stone, Computer Recognition of English Word Senses, North-Holland Linguistic Series, 1975.
[35] J. Kuruzovich, S. Viswanathan, R. Agarwal, S. Gosain, S. Weitzman, Marketspace or marketplace? Online information search and channel outcomes in auto retailing, Information Systems Research 19 (2) (2008) 182–201.
[36] E.L. Lesser, J. Storck, Communities of practice and organizational performance, IBM Systems Journal 40 (4) (2001) 831–841.
[37] N. Li, D.D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decision Support Systems 48 (2) (2010) 354–368.
[38] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, Journal of Finance 661 (1) (2011) 35–65.
[39] J.C. McCune, Snooping the Net, Management Review (1997) 58–59.
[40] D.M. McCutcheon, J.R. Meredith, Conducting case study research in operations management, Journal of Operations Management 11 (3) (1993) 239–256.
[41] K.A. Neuendorf, The Content Analysis Guidebook, Sage Publications Inc., 2001
[42] D.E. O'Leary, Blog mining-review and extensions: from each according to his opinion, Decision Support Systems 51 (4) (2011) 821–830.
[43] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (1–2) (2008) 1–135.
[44] M. Parameswaran, A.B. Whinston, Research issues in social computing, Journal of the Association for Information Systems 8 (6) (2007) 336–350.
[45] M.F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.
[46] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, Conference on Empirical Methods in Natural Language Processing (EMNLP-03), 2003, pp. 105–112.
[47] N.C. Romano, C. Donovan, H. Chen, J. Nunamaker, A methodology for analyzing web-based qualitative data, Journal of Management Information Systems 19 (4) (2003) 213–246.
[48] S.E. Sampson, Ramifications of monitoring service quality through passively solicited customer feedback, Decision Sciences 27 (4) (1996) 601–622.
[49] C. Sanes, Complaints are hidden treasure, Journal for Quality and Participation 16 (5) (1993) 78–82.
[50] D.E. Schendel, C.W. Hofer, Theory Building and Theory Testing: a Conceptual Overview, Strategic Management, Little, Brown and Company, Boston, 1979.
[51] A. Schenkel, R. Teigland, Improved organizational performance through communities of practice, Journal of Knowledge Management 12 (1) (2008) 106–118.
[52] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: the AZFin text system, ACM Transactions on Information Systems 27 (2) (2009).
[53] H. Schutze, Automatic word sense discrimination, Computational Linguistics 24 (1) (1998) 97–123.
[54] S. Spangler, J. Kreulen, Mining the Talk: Unlocking the Business Value in Unstructured Information, IBM Press, 2008.
[55] P.J. Stone, D.C. Dunphy, M.S. Smith, The General Inquirer: a Computer Approach to Content Analysis, MIT Press, Oxford, England, 1966.
[56] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: quantifying language to measure firms' fundamentals, Journal of Finance 63 (3) (2008) 1437–1467.
[57] P.D. Turney, M.L. Littman, Measuring praise and criticism: inference of semantic orientation from association, ACM Transactions on Information Systems 2 (4) (2003) 315–346.
[58] R.G. Vedder, M.T. Vanecek, C.S. Guynes, J.J. Cappel, CEO and CIO perspectives on competitive intelligence, Communications of the ACM 42 (8) (1999) 108–116.
[59] D.D. Ward, P.H. Jesty, R.S. Rivett, Decomposition scheme in automotive hazard analysis, SAE International Journal of Passenger Cars — Mechanical Systems 2 (1) (2009) 803–813.
[60] M.M. Wasko, S. Faraj, Why should I share? Examining social capital and knowledge contribution in electronic networks of practice, MIS Quarterly 29 (1) (2005) 35–57.
[61] E.C. Wenger, W.M. Snyder, Communities of practice and organizational performance, Harvard Business Review (2000) 139–145.
[62] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, Sixth International Conference on Intelligent Text Processing and Computational Linguistics, 2005, pp. 486–497.
[63] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing, Vancouver, 2005.
[64] A. Woolridge, Social media provides huge opportunities, but will bring huge problems, Economist (2011) 50.
[65] R.K. Yin, Case Study Research, Sage Publications, London, 1989.
[66] Y. Zhang, Y. Dang, H. Chen, M. Thurmond, C. Larson, Automatic online news monitoring and classification for syndromic surveillance, Decision Support Systems 47 (4) (2009) 508–517.
[67] M. Ziefle, Effects of display resolution on visual performance, Human Factors 40 (4) (1998) 555–568.

Alan S. Abrahams is an Assistant Professor in the Department of Business Information Technology, Pamplin College of Business, at Virginia Tech. He received a PhD in Computer Science from the University of Cambridge, and holds a Bachelor of Business Science degree from the University of Cape Town. Dr. Abrahams' primary research interest is in the application of decision support systems in entrepreneurship. He has published in a variety of journals including Expert Systems with Applications, Journal of Computer Information Systems, Communications of the AIS, and Group Decision and Negotiation.

Jian Jiao is a PhD candidate in Computer Science at Virginia Tech and a Software Design Engineer at Microsoft. He holds an M.S. in Computer Science from the Beijing Institute of Technology, and has previous work experience at Microsoft Research Asia and Motorola.

G. Alan Wang is an Assistant Professor in the Department of Business Information Technology, Pamplin College of Business, at Virginia Tech. He received a Ph.D. in Management Information Systems from the University of Arizona, an M.S. in Industrial Engineering from Louisiana State University, and a B.E. in Industrial Management & Engineering from Tianjin University. His research interests include heterogeneous data management, data cleansing, data mining and knowledge discovery, and decision support systems. He has published in Communications of the ACM, IEEE Transactions of Systems, Man and Cybernetics (Part A), IEEE Computer, Group Decision and Negotiation, Journal of the American Society for Information Science and Technology, and Journal of Intelligence Community Research and Development.

Dr. Weiguo (Patrick) Fan is a Full Professor of Accounting and Information Systems and Full Professor of Computer Science (courtesy) at the Virginia Polytechnic Institute and State University (Virginia Tech). He received his Ph.D. in Business Administration from the Ross School of Business, University of Michigan, Ann Arbor, in 2002, a M. Sce in Computer Science from the National University of Singapore in 1997, and a B.E. in Information and Control Engineering from the Xi'an Jiaotong University, PR China, in 1995. His research interests focus on the design and development of novel information technologies – information retrieval, data mining, text/web mining, business intelligence techniques – to support better business information management and decision making. He has published more than 100 refereed journal and conference papers. His research has appeared in journals such as Information Systems Research, Journal of Management Information Systems, IEEE Transactions on Knowledge and Data Engineering, Information Systems, Communications of the ACM, Journal of the American Society on Information Science and Technology, Information Processing and Management, Decision Support Systems, ACM Transactions on Internet Technology, Pattern Recognition, IEEE Intelligent Systems, Pattern Recognition Letters, International Journal of e-Collaboration, and International Journal of Electronic Business.